Springer Texts in Education

Jimmie Leppink

Statistical Methods for Experimental Research in Education and Psychology

Springer

Springer Texts in Education

Springer Texts in Education delivers high-quality instructional content for graduates and advanced graduates in all areas of Education and Educational Research. The textbook series is comprised of self-contained books with a broad and comprehensive coverage that are suitable for class as well as for individual self-study. All texts are authored by established experts in their fields and offer a solid methodological background, accompanied by pedagogical materials to serve students such as practical examples, exercises, case studies etc. Textbooks published in the Springer Texts in Education series are addressed to graduate and advanced graduate students, but also to researchers as important resources for their education, knowledge and teaching. Please contact Natalie Rieborn at textbooks.education@springer.com for queries or to submit your book proposal.

More information about this series at http://www.springer.com/series/13812

Jimmie Leppink

Statistical Methods for Experimental Research in Education and Psychology

Springer

Jimmie Leppink
Hull York Medical School
University of York
York, UK

ISSN 2366-7672 ISSN 2366-7680 (electronic)
Springer Texts in Education
ISBN 978-3-030-21240-7 ISBN 978-3-030-21241-4 (eBook)
https://doi.org/10.1007/978-3-030-21241-4

© Springer Nature Switzerland AG 2019
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

For my mentor, and to the world

Writing a book is a truly humbling exercise.
No matter how many sources you studied and seek to include in your work,
at the end of the journey, you feel you have barely scratched the surface,
you realise how little you know,
and how much there is and always will be out there to learn.

To those whom I offended with a harsh judgement from my side,
be it for a seemingly inappropriate choice in an empirical study,
for a supposed lack of knowledge, or else,
I apologise.

It is okay to not know things,
and what we know, is effectively a spark of the universe;
when we join forces, connect those small lights, and learn from each other,
we, together, can reach for the skies.

I dedicate this book to a great person who has been a support and inspiration,
with lessons not only about research but

equally about life,
the late Dr. Arno Muijtjens (11 April 1950–2 April 2019):
because for me, Arno, you have been a mentor and teacher in many ways,
and you have been, and will be, an example to follow.

Preface

We are living in exciting times. The movement called Open Science is visible everywhere, all the way from the use of materials, the design of new research, data collection, data analysis, as well as the reporting of findings and sharing of data. Preregistration and registered reports are being adopted by increasing numbers of journals across fields to complement or replace the traditional peer-review system. Zero-cost Open-Source software packages allow people to save and share analyses with anyone in the world without having to face paywalls. Data sharing offers tremendous opportunities to bring the discourse about empirical phenomena to a next level and is slowly but steadily becoming more common. Where a common question used to be which methods are better or worse than which other methods, the key question these days appears to be how combinations of methods can help us make better decisions for research and practice. This book attempts to shed some light on this apparent key question.

York, UK Jimmie Leppink

Acknowledgements

My interest in Methodology and Statistics was first stimulated by my teachers in Methodology and Statistics in the Bachelor of Science in Psychology at Maastricht University, the Netherlands, Dr. Nick Broers and Prof. Dr. Gerard van Breukelen. During my Ph.D., I was fortunate to learn from many people at Maastricht University's Department of Methodology and Statistics, including my aforementioned teachers as well as Emeritus Professor Dr. Martijn Berger and Dr. Frans Tan. During my Master of Science in Quantitative Analysis in the Social Sciences (Methodology and Statistics) at the Catholic University of Leuven, Belgium, I was fortunate to have great teachers, including Emeritus Professor Dr. Jacques Tacq, Prof. Dr. Geert Molenberghs, Prof. Dr. Kelvyn Jones, and the late Emeritus Professor Dr. Allan L. McCutcheon. For several years (from the day I opened it on the 7th of February 2012 until the day Dr. Patrick Blessinger took over, the 5th of November of 2018: Thank you for that! I am very happy that the group continued to grow and thrive after I resigned), I managed a LinkedIn discussion group entitled "Research, Methodology, and Statistics in the Social Sciences" which hosted over 141,000 people from across the world in Autumn of 2018, and over the years learned from many great scholars including Prof. Dr. Thom Baguley, Emeritus Professor Dr. John D. Eveland, and Dr. Shane Tutwiler. In my work as Postdoc and Assistant Professor, I learned from many people including Dr. Fred Stevens, Dr. Jeroen Donkers, Dr. Carlos Collares, Dr. Paul Ginns, Dr. Saad Chahine, Dr. Kulamakan (Mahan) Kulasegaram, Dr. Kalman Winston, Prof. Dr. Patricia O'Sullivan, Prof. Dr. Cees van der Vleuten, Prof. Dr. Jeroen van Merriënboer, Prof. Dr. Erik Driessen, and—a true mentor for me—the late Dr. Arno Muijtjens. A big Thank You to goes to Maastricht University for an incredible 16.5 years (September 2002–January 2019) of adventure in different roles: Student, Tutor, Trainer, Research Assistant, Ph.D. Candidate, Postdoc (5 years during which I had plenty of opportunities to explore new things), and Assistant Professor. The product of the education, training, and opportunities to explore the world you have given me will always be visible in my work. I thank the Catholic University of Leuven for such a great experience with the Master of Science and the flexibility in the programme which allowed me to complete it next to the last bits of my Ph.D. and start of my Postdoc (October 2011–July 2012). Many thanks to Hull York Medical School and the University of York for my most recent step in my career (January 2019); with great colleagues like Prof. Dr. Martin Veysey, Dr. Andrew Brown, and Dr. Liz Glidewell, it is an honour and a pleasure to work here, and I hope this is just the start of a long journey together. Finally, I am fortunate that my wife, Dr. Patricia Pérez-Fuster, has been and always is a great support, and that we have already managed to publish several articles on Methodology and Statistics together. All the people mentioned here and many more, including the reviewers of the proposal with which this book started, and the reviewers of the full manuscript Dr. Paul Ginns and Dr. Shane Tutwiler, have had their impact on the content of this book; for all this, I am immensely grateful.

York, UK Jimmie Leppink

Contents

Part I Common Questions

Part II Types of Outcome Variables

Part III Types of Comparisons

Part IV Multilevel Designs

About the Author

Jimmie Leppink (28 April 1983) obtained degrees in Psychology (M.Sc., September 2005–July 2006, Cum Laude), Law (LLM, September 2007–July 2008), and Statistics Education (Ph.D., September 2008–March 2012) from Maastricht University, the Netherlands, and obtained a degree in Statistics (M.Sc., October 2011–July 2012, Magna Cum Laude) from the Catholic University of Leuven, Belgium. He defended his Ph.D. Thesis in Statistics Education in June of 2012, and was a Postdoc in Education (April 2012–March 2017) and Assistant Professor of Methodology and Statistics (April 2017–January 2019) at Maastricht University's School of Health Professions Education. Since January 2019, he has been working as a Senior Lecturer in Medical Education at Hull York Medical School, which is a joint medical school of the University of Hull and the University of York. His research, teaching, and consulting activities revolve around applications of quantitative methods in Education, Psychology, and a broader Social Science context as well as the use of learning analytics for the design of learning environments, instruction, and assessment in Medical Education and the broader Higher Education.

Common Symbols and Abbreviations

α	Statistical significance level in NHST, and Cronbach's or Krippendorff's alpha in psychometric analysis
β	Standardised regression coefficient
η^2, partial η^2	Effect size estimates in (M)AN(C)OVA
κ	Cohen's kappa
μ	Population mean
ρ	Spearman's rho
σ	Population standard deviation
τ	Kendall's tau
φ	Coefficient phi for two-way contingency tables
χ^2	Chi-square
ω^2	Effect size estimate in (M)AN(C)OVA, generally slightly less biased than η^2 though the difference between ω^2 and η^2 decreases with increasing sample size
-2LL, -2RLL	The minus two log-likelihood aka deviance of a model using FIML (-2LL; fixed effects) or using REML (-2RLL; random effects): LL stands for log likelihood
A, B, C	In single-treatment-factor designs used to denote conditions (e.g., A and B being two treatments, and C the control), in two- or three-factor designs used to denote the different factors; sometimes also used to name a hypothetical experiment (e.g., Experiment A) or different assessors/raters
A-B, B-A	Used to denote orders of conditions in an experiment that includes a within-subjects factor
A-by-B	Term used for two-way contingency tables as well as for an interaction effect of A and B
AD1	First-order ante-dependence residual covariance structure
AIC	Akaike's information criterion
AICc	Corrected AIC, a variant of AIC that has a slightly lower tendency towards more complex models than AIC, though the difference between AIC and AICc reduces with increasing sample size

ANCOVA	Analysis of covariance, one of the methods used in this book
ANCOVRES	Analysis of covariate residuals, one of the methods discussed in this book
ANOVA	Analysis of variance, one of the methods used in this book
AR1	First-order autoregressive residual covariance structure
B, b	Non-standardised regression coefficient (i.e., using the original scales of predictor and outcome variable, not *SD*s as scales for these variables)
BF	Bayes factor
BF_{01}, BF_{10}	BF of $\boldsymbol{H_0}$ vs. $\boldsymbol{H_1}$ (i.e., 01) and of $\boldsymbol{H_1}$ vs. $\boldsymbol{H_0}$ (i.e., 10), respectively
BIC	Schwarz' Bayesian information criterion
BLUE	Best linear unbiased estimate
c, *r*	In combination sometimes used to denote columns and rows
CI	Confidence interval; in this book, I hold a plea for both the '1—α' and the '1—2α' CI. Provided assumptions are met, a C% CI should include the population parameter of interest in C% of all possible samples of the same size drawn from that population
CRI	Credible interval aka posterior interval, a Bayesian alternative to the Frequentist confidence interval
CS	Compound symmetry
D	Used to denote a difficulty-levels independent variable
d	Cohen's *d*, a measure of effect size which expresses the differences between two *M*s in *SD*s instead of in actual scale (e.g., points or minutes) units
df	Degrees of freedom
DR	Difference in rating
DRF	Deviance reduction factor, also known as R^2_{MCF} in fixed-effects categorical outcome variable models, and for any comparison of fixed-effects solutions in fixed-effects and mixed-effects models in randomised controlled experiments involving categorical or quantitative outcome variables, a straightforward indicator of the reduction in -2LL or deviance by a more complex model relative to a simpler (i.e., special case) variant of that more complex model (e.g., Model 1 vs. Model 0, or a main-effects plus interaction model vs. a main-effects model)
EMM	Estimated marginal mean(s)
F	Probability distribution and test statistic; for comparisons that involve two conditions or categories (i.e., df_{groups} = 1 or df_{effect} = 1), $F = t^2$. As sample size goes to infinity, the

	difference between F for $df_{\text{groups}} = 1$ or $df_{\text{effect}} = 1$ and x_1^2 converges to zero
f	Effect size estimate that is similar to standardised β and can be used for required sample size calculations for (M)AN(C)OVA
FIML, ML	Full information maximum likelihood, used for the estimation of fixed effects and provides a valid approach to dealing with missing data without imputation under MAR and under MCAR
FOST	Four one-sided testing, part of PASTE, and unites TOST and ROPE
g	Hedges' *g*, an effect size estimate that is similar to Cohen's *d*
GLB	Greatest Lower Bound, along with McDonald's omega one of the proposed alternatives to Cronbach's alpha
GPower	A statistical package used in this book, version 3.1.2
GR	Group
$\boldsymbol{H_0}$	Null hypothesis
$\boldsymbol{H_{0.1}}$, $\boldsymbol{H_{0.2}}$, $\boldsymbol{H_{0.3}}$, $\boldsymbol{H_{0.4}}$	The coherent set of four H_0s tested in FOST, the first two of which are also used in TOST
$\boldsymbol{H_1}$	Alternative hypothesis
HF	Huynh–Feldt residual covariance structure
ICC	Intraclass correlation
ITT	Intent to treat: the average effect of offering treatment
Jamovi	In other sources sometimes denoted with small J as *jamovi*, a statistical software package used in this book, version 0.9.5.16
JASP	A statistical software package used in this book, version 0.9.2.0
JZS	JeffreysZellner–Siow
K, *k*	Commonly used to denote the number of items, raters (assessors), stations, or repeated measurements per participant; sometimes also used to denote the number of statistical tests of comparisons (to be) made in a given context (the latter is also denoted as C_p)
KR-20	One of the Kuder–Richardson coefficients, a special case of Cronbach's alpha for dichotomous items or ratings
LB	Lower bound
LR	Likelihood ratio
M	(arithmetic) mean in a sample or condition
M_d	Mean difference
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random

M_N	A factor by which to multiply N based on ICC = *0 to account for* ICC > *0*
Model 0	Null model ($\boldsymbol{H_0}$ or simplest of models under comparisons)
Model 1	Alternative model ($\boldsymbol{H_1}$), a more complex version of the null model, and when there are several alternatives to Model 0, together with Model 0 among several competing models in model comparison/selection (e.g., also Models 2–4 when we deal with possible main and interaction effects of two factors)
Mplus	A statistical software package used in this book, version 8
N	Total sample size (i.e., all conditions together)
n	Sample size per condition
NHST	Null hypothesis significance testing
N_{O}	Number of possible orders
O	Observation
OR	Odds ratio
p	Probability; in NHST (i.e., the p-value), the probability of the test statistic value observed or further away from H_0, if H_0 is true (i.e., a conditional probability, the condition being truth of H_0)
PASTE	Pragmatic approach to statistical testing and estimation, the core approach in this book, which is about combining different methods of testing and estimation to make informed decisions
$P(\boldsymbol{H_0})$, $P(\boldsymbol{H_1})$	The probability of $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$, respectively, *prior to* seeing the data. These concepts fit within a Likelihoodist, Bayesian, and to some extent also within an information-theoretic (i.e., information criteria) approach to statistical testing but *not* within the Frequentist approach
$P(\boldsymbol{H_0}\mid\mathrm{O})$, $P(\boldsymbol{H_1}\mid\mathrm{O})$	The probability of $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$, respectively, *after* seeing the data. These concepts fit within a Likelihoodist, Bayesian, and to some extent also within an information-theoretic (i.e., information criteria) approach to statistical testing but *not* within the Frequentist approach
$P(\mathrm{O}\mid\boldsymbol{H_0})$, $P(\mathrm{O}\mid\boldsymbol{H_1})$	The probability of findings observed under $\boldsymbol{H_0}$ and under $\boldsymbol{H_1}$, respectively. These concepts fit within a Likelihoodist, Bayesian, and to some extent also within an information-theoretic (i.e., information criteria) approach to statistical testing but *not* within the Frequentist approach; note that the Frequentist p-value is *not* $P(\mathrm{O}\mid\boldsymbol{H_0})$ but $P(\mathrm{O}$ *or further away from* $\boldsymbol{H_0} \mid \boldsymbol{H_0})$
QDA	Question-design-analysis (bridge, heuristic)
R	An environment in which, amongst others, statistical packages can be developed and used; also used in this book, version 3.5.0

r	Pearson's correlation coefficient
R1-R12	Used to denote different (competing) candidate models for the residual covariance structure in mixed-effects models
R^2, R^2 adjusted	Proportion of variance explained in fixed-effects models for quantitative outcome variables, with a penalty for model complexity in the case of the adjusted variant
R^2_{C}, R^2_{M}, R^2_{R}	Useful in multilevel analysis, at least when using RI (i.e., CS) models or not too complex extensions of that model (e.g., one RI term, one RS term, and their covariance): R^2_{C} combines the fixed and random effects, R^2_{R} focusses on the random effects (i.e., *ICC*), and R^2_{M} is about the fixed effects and can, therefore, be interpreted in a similar fashion as R^2 in fixed effects models; with more complex residual covariance structures, computing these or other R^2-statistics becomes more complicated
R^2_{CS}	*R*-squared statistic attributed to Cox and Snell, potentially useful for categorical outcome variable models though its upper bound is (well) below 1
R^2_{MCF}	*R*-squared statistic attributed to McFadden, my recommended default *R*-squared statistic for categorical outcome variable models
R^2_{N}	*R*-squared statistic attributed to Nagelkerke, potentially useful for categorical outcome variable models
R^2_{T}	*R*-squared statistic attributed to Tjur, potentially useful for categorical outcome variable models though its upper bound is (well) below 1
REML	Restricted maximum likelihood, used for the estimation of random effects
RI	Random intercept
RI-RS	Model consisting of at least one RI and at least one RS term
ROPE	Region of practical equivalence, a Bayesian concept similar to the region of relative equivalence in TOST, and part of the TOST-ROPE uniting FOST
r_{RES}	Residual correlation
RS	Random slope
RStudio	A statistical package used in this book, version 1.1.456
SABIC	Sample-size adjusted BIC
S	Sometimes used to denote a score outcome variable
SD	Standard deviation in a sample or condition
SE	Standard error
SocNetV	Stands for Social Network Visualizer, and is a statistical package used in this book, version 2.4

SPSS	Stands for Statistical Package for the Social Sciences, and is a statistical package used in this book, version 25
SS	Sum or squares
Stata	A statistical package used in this book, version 15.1
SUTVA	Stable unit treatment value assumption
T	Time point/occasion
t	Probability distribution and test statistic, slightly wider than the standard Normal distribution z though the difference between t and z converges to zero as sample size goes to infinity; squaring t, we obtain F for $df_{\text{groups}} = 1$ or $df_{\text{effect}} = 1$
TOST	Two one-sided tests, an approach to (relative) equivalence testing
TOT	Treatment of the treated: the average effect of receiving treatment
Type I error	Concluding there is an effect where there is none
Type II error	Concluding there is no effect where there is one
Type M error	(Substantial) misestimation of an effect of interest in magnitude
Type S error	Seeing a negative effect that is actually positive or vice versa
U	Mann–Whitney's nonparametric test statistic
UB	Upper bound
UN	Unstructured residual covariance model, a model in which V_{RES} is allowed to be different for each item or occasion and in which r_{RES} is allowed to be different for each pair of items or occasions
V	Cramér's V
VAS	Visual analogue scale, a continuous scale on which a respondent or assessor can pick a point that is supposed to provide a measurement of interest
V_{FIXED}	Fixed-effects variance
V_{RANDOM}	Random-effects variance
V_{RES}	Residual variance
V_{RI}	RI variance
V_{RS}	RS variance
VS-MPR	Vovk–Sellke maximum p-ratio
X	Commonly used to denote an independent or treatment variable
Y	Commonly used to denote an outcome variable
Z	Probability distribution (i.e., standard Normal distribution) and test statistic; squaring z yields x_1^2

Part I
Common Questions

1 The Question-Design-Analysis Bridge

Abstract

This is the first of four chapters of Part I of this book. The focus of Part I lies on common questions in experimental research. In this first chapter, the question-design-analysis (QDA) heuristic is introduced: there is a bridge connecting research questions and hypotheses, experimental design and sampling procedures, and common statistical methods in that context. This heuristic runs throughout the parts and chapters of this book and is used to present statistical analysis plans for a wide variety of situations and which include alternative routes to deal with assumption departures. This chapter briefly introduces the different statistical methods that are covered in this book through the QDA heuristic. Both likely candidates and alternative methods are presented for different questions and designs. Further, common flaws that move away from the original research question or from a particular hypothesis and/or fail to appropriately account for one experimental design feature or another are discussed. Examples of flaws discussed in this chapter are two-sided testing where one-sided testing would be expected or vice versa, treating two-way data as one-way data, treating a mediating variable as a confounding variable, and treating a within-subjects effect as between-subjects. This chapter provides the foundation for the subsequent chapters in this first part of the book: different approaches to statistical testing and estimation (Chap. 2), important principles of measurement, validity, and reliability (Chap. 3), and methods to deal with missing data (Chap. 4).

Introduction

During his Presidential Address at the First Indian Statistical Congress in 1938, Sir Ronald Aylmer Fisher (1890–1962) reflected (1938, p. 17): "*A competent overhauling of the process of collection, or of the experimental design, may often*

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_1

increase the yield ten or 12-fold, for the same cost in time and labour. To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. To utilise this kind of experience he must be induced to use his imagination, and to foresee in advance the difficulties and uncertainties with which, if they are not foreseen, his investigations will be beset." Fisher is one of the greatest statisticians of all time (Efron, 1998), and his books *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935) have served as a foundation for most if not all literature on experimental research since. Even literature in which no direct reference is made to Fisher (1925) or Fisher (1935) still cites work that is based on Fisher (1925, 1935). In the 95 words from his Presidential Address just quoted, Fisher taught us that anticipation of possible challenges and careful planning and designing constitute a *conditio sine qua non* for experimental research. Moreover, the goal of experiments is not just to answer questions; an experiment ought to be "*designed to form a secure basis of knowledge*" (Fisher, 1935, p. 8). Experimental research is not about using theory to solve local problems (e.g., which teaching approach should be used in Programme A at University X in country C); it is about *advancing* theory, and in order to be able to advance theory, we need *carefully designed* experiments.

In the words of Winer (1970, p. 1), those best qualified to design an experiment are those who are "*(1) most familiar with the nature of the experimental material, (2) most familiar with the possible alternative methods for designing an experiment, (3) most capable of evaluating the potential advantages and disadvantages of the alternatives. Where an individual possesses all these qualifications, the roles of experimenter and designer are one.*" In contemporary experimental research practice, this situation is relatively rare. Usually, experimental research is carried out in teams, not in the last place because different team members bring different knowledge, experience, and skills. Content experts are typically the ones who are most familiar with the nature of the experimental material but are not necessarily most familiar with different ways of designing an experiment let alone pros and cons of those alternatives. Simultaneously, methodologists and statisticians may be very familiar with the latter but not with the former. Unfortunately, eight decades after Fisher's Presidential Address (1938), methodologists and statisticians are still in too many cases contacted for help once the experiment has been carried out, in the hope that they can help researchers to 'correct' for some of the mistakes made at one or several stages in the process. Even if such corrections are possible, which is often not the case, *methodological control* (i.e., control by *design*) is virtually always better than *statistical control* (i.e., control by *analysis*); as a rule of thumb, the more of the latter is needed (if at all possible), the weaker the conclusions that can be drawn from the experiment. Therefore, in Fisher's words: researchers should seek the support of a methodologist or statistician *before*, not only after, the experiment. Whether you are designing a single experiment or you are in the process of writing a grant proposal that presents a coherent series of experiments, always consider having at least one well-trained methodologist or statistician on board from the start.

Focus on the Design and Analysis of Larger-Sample Experiments

Succinctly put, an experiment is a study that is aimed at establishing causal relations between activities, practices or procedures on the one hand and variables that *respond* to these activities, practices or procedures on the other hand. The activities, practices or procedures to be compared in an experiment together constitute one or more *independent variables*, whereas the variables that respond to these activities, practices or procedures are referred to as *dependent variables* and are also known as *response variables* and *outcome variables*. For example, two teaching methods—a newly developed one and a conventional one—may constitute an independent variable for researchers who are interested in the question under what conditions the newly developed method (i.e., *treatment condition*) can be expected to result in better learning outcomes (i.e., *outcome variable*) than the conventional method (i.e., *control condition*). However, in a *true experiment*, participants are *sampled* from a target population of interest *at random* and are *allocated* to the conditions *at random* as well. In other words, the two conditions do not result from already existing groups, such as Class A in School J happening to be taught the conventional way and Class B in School K being taught in the newly developed format. The latter would be an example of a *quasi-experiment*. The fact that this book focusses on *experimental research* and does not cover *quasi-experimental research* should by no means be interpreted as an implication that quasi-experimental research is not useful or is necessarily inferior to experimental research. Different ways of doing research may yield temporary answers to somewhat different questions and may be bound to a partially different set of assumptions. Quasi-experimental research, like experimental research, has its use provided that it is done correctly and the interpretations and implications are restricted to where they apply.

Every author has to make choices, some of which relate to being all-inclusive versus focussing on a particular content. I have chosen for the latter. That said, although quasi-experimental research is not covered in separate chapters, comparisons between experiments and quasi-experiments as well as dangers of applying statistical techniques to a quasi-experiment as if it was a true experiment are discussed in several chapters of this book. Great general reads on quasi-experimental research include Cook and Campbell (1979), Cook and Wong (2008), Huitema (2011), and Stuart and Rubin (2010).

Another choice has been to focus on larger-sample experiments and to not include specific chapters on single-case experiments. As for quasi-experiments, this choice is not intended to serve as a statement that single-case experiments have no use or are necessarily inferior to larger-sample experiments. The rationale behind not covering single-case experiments in detail is simply that single-case research is a world by itself. Moreover, several books (e.g., Barlow & Hersen, 2009; Ledford & Gast, 2018), book sections (e.g., Huitema, 2011), and overview articles (e.g., Forbes, Ross, & Chesser, 2011; Michiels, Heyvaert, Meulders, & Onghena, 2017;

Michiels & Onghena, 2018; Normand, 2016; Onghena, Maes, & Heyvaert, 2018; Phan & Ngu, 2017; Smith, 2012) have already been written on this topic, and the intention of this book is not to provide an overview of thousands of statistical methods for all kinds of research out there. Note that the term ‘larger-sample’ has been chosen intentionally over ‘large-sample’, to denote a preference to move beyond *we-have-to-publish-fast* small samples but leave open to the interpretation of the reader what is ‘large’ and what is ‘small’. In many educational and also quite some psychological research settings, a two-group experiment with $n = 64$ per group may already be considered ‘large’, whereas some sociologists who do experiments in the context of say different presentation formats of surveys may call samples of a size (N) of a few 100 s or participants ‘fairly small’.

Finally, another topic that is only mentioned briefly in some places in this book is that of meta-analysis. Although authors of a book on experimental research could well choose to include meta-analysis, this book focusses on the design and analysis of larger-sample experiments, and refers readers to great reads like Konstantopoulos (2010), Kruschke and Liddell (2017), Lipsey and Wilson (2001), Patall and Cooper (2008), and Viechtbauer (2010) for an in-depth study of the world of meta-analysis.

The Bridge

Bloom (2008) provides an excellent, concise overview of five elements of a successful experiment: (1) *research questions*; (2) *experimental design*; (3) *measurement methods*; (4) *implementation strategy*; and (5) *statistical analysis*. These elements are not isolated; there is a bridge that connects them. Besides, in my view, experimental design, measurement methods, and implementation strategy are hard to see as separate elements in practice. Moreover, experiments are usually not just informed by research questions; one or more testable hypotheses, general or more specific, are normally available as well. Therefore, in this book, the term *design* includes the elements of experimental design (e.g., Bloom, 2008; Howell, 2017), *sampling* (e.g., Kish, 1965), *measurement* methods (e.g., Crocker & Algina, 1986), and implementation strategy, and the term *question* in the QDA heuristic is used for the coherent set of research questions and hypotheses. Finally, the term *analysis* in this heuristic also includes reporting, in manuscripts or presentations, such that an audience can understand what choices have been made in a given experiment and why.

From Question to Design

To start, the type of research questions and testable hypotheses determine our likely options for the experimental design. If we are interested in whether a new online health intervention aimed at reducing alcohol consumption in a given target population is more effective in doing so than a conventional online health intervention, a two-group experiment may do. However, if researchers have reasons to expect that the effect of the new online health intervention is different in different sub-populations (e.g., men and women) within that target population, a *two-way factorial design* (e.g., Field, 2018; Howell, 2017), with intervention (new or *experimental treatment* condition vs. conventional or *control* condition) and gender (men vs. women) as factors, is more appropriate. In the aforementioned two-group experiment, *simple random sampling* of participants and random allocation of the randomly sampled participants to conditions could do. However, in the two-way factorial case it would be better apply a form of *stratified random sampling* (e.g., Kish, 1965): to apply simple random sampling and random allocation to the men, and to the women, in such a way that the proportions of men and women in both conditions are equal. Stratified random sampling, in the latter case, could yield a considerable increase in precision, especially when sample sizes are not that large (e.g., $N < 100$). Stratification or blocking is then applied, and random sampling and random allocation are then applied for each stratum or block (here, the strata or blocks would be men and women).

This brings us to two main reasons why *random sampling* constitutes the most rigorous way to study causal relations between independent variables like the type of health intervention in the example and outcome variables like a measure of alcohol consumption or of a reduction of alcohol consumption. Both random sampling of participants and random allocation of these participants to experimental conditions are ways to eliminate *bias*, because only the laws of probability determine which participants are included in the experiment and which participants are assigned to which condition in that experiment. In this case, all pre-existing differences between groups or conditions are due to chance. If we were to do the same experiment with the same sample size an infinite number of times and applied random sampling and random allocation in every single experiment, on average—across all experiments—the differences between conditions at the start of the experiment, prior to any treatment, could be expected to be zero. Of course, between experiments, there would be differences between conditions at start, sometimes in favour of one condition sometimes in favour of the other condition, and sometimes larger sometimes smaller. However, these differences between experiments would decrease if we took larger samples. Where the laws of probability apply, all possible sources of uncertainty about estimates of phenomena of interest are randomised, confidence intervals (CIs) and statistical significance tests can account for that uncertainty, and the *internal validity* of the experiment can be guaranteed. In the words of Bloom (2008, p. 128): "*Absent treatment, the expected values of all past, present, and future characteristics are the same for a randomized treatment group and control group. Hence, the short-term and long-term future experiences of the control group provide valid estimates of what these experiences would have been for the treatment group had it not been offered the treatment.*"

In small samples, large differences between the sample (size N) and the population may be quite likely, and large differences between conditions (size n of N) prior to any treatment are likely as well. Apart from small samples, non-random sampling and non-random allocation have also been identified as sources that decrease the likelihood of findings from one experiment being replicated in a next experiment (Hedges, 2018; Ioannidis, 2005a, b; Tipton, Hallberg, Hedges, & Chan, 2017). This is not to say that it is absolutely impossible to generalise sample findings to a population of interest in cases where random sampling is practically not feasible. However, researchers should be aware that randomly allocating a relatively 'large' sample to conditions in an experiment does not compensate for applying some kind of convenience sampling or some kind of purposeful sampling, and that failing to apply random sampling may well come at the cost of a reduced generalisability of findings to a population of interest. In the words of Bloom (2008, p. 116): "*One cannot, however, account for all uncertainty about generalizing an impact estimate beyond a given sample (its external validity) without both randomly sampling subjects from a known population and randomly assigning them to experimental groups*." Furthermore, random sampling and random allocation should not be used as an excuse for having small samples. Even if random sampling and random allocation are applied, in the case of small samples findings of interest can vary wildly from one experiment to the next. Consequently, even if the phenomenon of interest is the same in a population sampled from in both initial and replication experiment and the latter is an exact replication of the former, with small samples it is quite likely to not replicate findings from the initial experiment in that direct replication.

Increasingly, researchers carry out experiments where not individuals but existing clusters of individuals are sampled and allocated to conditions, for example health centres or schools some of which participate in a treatment condition some of which participate in a control condition. As for the sampling and allocation of individuals, random sampling plus random allocation constitutes the best practice, for it can be expected to provide unbiased estimates of intervention effects. However, similarities and interactions between individuals within clusters induces a within-cluster between-participant correlation (aka *intraclass correlation*, *ICC*; e.g., Leppink, 2015a; Snijders & Bosker, 2012; Tan, 2010) that reduces the effective sample size from *N* to somewhere in between that sample size *N* and the *k* number of clusters (i.e., the higher the *ICC*, the more the effective sample size goes down to *k*), and statistical power and precision of treatment effects with it (Bloom, 2008; Leppink, 2015a; Snijders & Bosker, 2012). Bloom (2008, p. 126) provides a useful formula for calculating the so-called design effect or the number of times the standard error (*SE*) for cluster randomisation compared to the *SE* for individual randomisation, as a consequence of the *ICC* in cluster-randomised studies:

$$\text{design effect} = \sqrt{(1 + [(n - 1) * ICC])}.$$

In this formula, *n* is a constant number of individuals per class or cluster. In its squared form, hence without the square root, the formula provides an indication of how many times as many participants per cluster we would need to achieve the same precision as if *ICC* was zero. Where clusters of participants are concerned, *ICC*-values in the range of 0.01–0.20 are very common, and the expected value depends on the nature of the clusters, the type of experimental manipulation, and the type of outcome variable in question. If in a given context *ICC* = 0.10 is a reasonable assumption and the current number of participants per cluster is 20, the design effect is the square root of 2.9 (ca. 1.703). To achieve the same precision for cluster randomisation with *ICC* = 0.10 instead of *ICC* = 0, we would need 2.9 times as many sample members. Note that if *ICC* = 0, the design effect equals 1 regardless of *n*, meaning no need to plan a larger sample size, but *ICC* = 0 in the case of cluster randomisation would be an unrealistic assumption.

Even if in an experiment treatment can vary within clusters, the *ICC* is something to be accounted for by aiming for a larger sample size. The same holds for cases where so-called *multistage sampling* (e.g., Kish, 1965) is applied. In the case of health centres or schools, for example, two-stage sampling is applied when not all individuals of health centres or schools sampled are invited to participate but only a sample of each centre or school is invited to participate. In such cases, the first stage of sampling is that of which centres or schools are contacted, and the second stage of sampling is found in sampling individuals within the centres or schools contacted. Whether we deal with workplaces, schools, households or other natural clusters, individuals coming from the same clusters tend to be similar in ways that are likely to result in an *ICC* to be accounted for.

In some cases, statistical power and precision can be gained by using predictive power of *covariates* (Bloom, 2008; Van Breukelen, 2006; Van Breukelen & Van Dijk, 2007), provided that these covariates are anticipated and *planned prior to data collection* and make sense from a theoretical perspective (Gruijters, 2016), and are *not affected by* the treatment (Leppink, 2015b, 2017a). As explained earlier, where the laws of probability apply, group differences unrelated to treatment are due to nothing but chance, hence participants' values on a covariate measured *before* the start of the treatment do not influence which condition a given participant will be part of and vice versa. The only case in which a covariate may reasonably be *affected by* experimental treatment is where the covariate is measured *after* the start of that treatment. However, even in that case, covariates cannot and should not be expected to help researchers to 'control' for group differences unrelated to the treatment; instead, such a covariate should be treated as a *mediator* instead of as a confounder, and this has implications for how we design our experiments (e.g., allow study time or compliance to vary and be measured instead of keeping these variables constant across conditions by design; Leppink, 2015b, 2017a). Also note that covariates (mediators or not) may or may not moderate treatment effects of interest (Field, 2018; Huitema, 2011; Leppink, 2018a, b) and to ensure sufficient statistical power and precision for moderation of a meaningful magnitude we need large enough samples.

Finally, in some cases, researchers have an interest in changes over time, and more specifically, differences between conditions in a change over time, in a given outcome variable of interest. In these cases, it is important to carefully plan measurements and account for within-participant between-measurement correlation (i.e., the *ICC* but now at the level of individual participants instead of at the level of clusters of participants): *N* participants being measured *k* times each cannot and should not be seen as an effective sample size of *N* times *k* (Leppink, 2015a; Snijders & Bosker, 2011; Tan, 2010). In the context of repeated measurements from the same participants, *ICC*-values in the range of 0.30–0.60 are not uncommon. Whether we have to deal with $ICC > 0$ due to cluster randomisation, $ICC > 0$ due to repeated measurements, or both, we must account for that when planning our sample size (e.g., Hox, Moerbeek, & Van der Schoot, 2017).

From Design to Analysis

One of the main conclusions from the previous section is that randomly drawing sufficiently large samples from populations of interest that are then allocated randomly to the different treatment conditions is to be considered the gold-standard method. This is not to say that small samples are useless or that a failure to randomise at either stage of sampling or allocation means we can throw the data of our experiment in the bin, but the further we move away from the gold standard the more our interpretations and generalisations of findings may be questioned. Applying statistical methods that are suitable for experiments that meet the gold standard to studies that are quite far away from that gold standard is tricky business. For example, dealing with pre-existing groups in a quasi-experiment, we will likely need to undertake additional steps at the data-analytic stage that are not needed in the case of a true experiment (Cook & Wong, 2008; Hedges, 2018; Stuart & Rubin, 2010). Omitting these additional steps, as if we were dealing with an experiment instead of with a quasi-experiment, comes at the risk of inappropriate conclusions and recommendations for future research and practice in a field.

Even if we deal with a true experiment, the features of the experimental design have to be accounted for in the analysis. For example, data from a two-way factorial design ought to be analysed as two-way not one-way data, because a one-way analysis does not provide a formal test of the interaction between the two factors (i.e., moderation aka effect modification) and comes at the cost of a substantially reduced statistical power and precision for group differences (Leppink, O'Sullivan, & Winston, 2017).

In the context of the confounder-mediator distinction, if we treat a mediator as a confounder, we will erase at least part of a treatment effect of interest as if it did not exist (Leppink, 2017a) and that will likely result in inadequate recommendations for future research or practice. The same holds for omitting the analysis or reporting of variables that moderate treatment effects of interest; not analysing or not reporting such moderations stimulates people to draw the conclusion that a treatment effect of interest is fairly stable across the range of a third variable (i.e., the moderator) while in fact it may differ substantially for different areas of that third variable (e.g., Field, 2018; Hayes, 2018; Huitema, 2011; Leppink, 2018b).

Failing to account for cluster randomisation or repeated measurements constitutes another recipe for incorrect conclusions with regard to treatment effects of interest. Even in the case of complete data (i.e., 100% response, 0% non-response or missing data otherwise), ignoring cluster structures or repeated-measurements structures tends to result in too narrow CIs and too small *p*-values in statistical significance tests for between-subjects effects, and ignoring repeated-measurements structures additionally tends to result in too wide CIs and too large *p*-values for within-subjects effects (Leppink, 2015a; Tan, 2010).

Note that the features to be accounted for in the analysis discussed thus far largely relate to the design, which ought to logically follow from the research questions and hypotheses at hand. With regard to hypotheses, we researchers in Education and Psychology have the habit of engaging in two-sided hypothesis testing where one-sided testing would be allowed. For example, based on theory and previous research we may expect that providing novices in programming with worked examples of how to programme results in faster learning and hence better post-test performance after an experiment than providing novices with no worked examples. In that, we may test the null hypothesis of no difference against a one-sided alternative that the worked examples (i.e., experimental treatment) condition will (on average) perform better than the no worked examples (i.e., control) condition. As discussed in Chaps. 2 and 10, the advantage of a one-sided test over a two-sided test is then an increased statistical power given a sample size *N* and a somewhat smaller sample size *N* needed for a given desired statistical power. Of course, hypotheses are to be formulated prior to data collection and should not be revised after seeing the data; formulating one-sided hypotheses after seeing the data is not allowed and can be expected to result in more false alarms (i.e., incorrect rejections of null hypotheses).

Compliance and Non-compliance

An important assumption in causal inference is that of *stable unit treatment value assumption* (SUTVA). Under SUTVA, a treatment applied to one participant does not affect the outcome of another participant (Cox, 1958; Rubin, 1978), and there is only a single version of each condition; if either of these two is violated, causal effects are hard to define, and an experiment cannot be expected to yield unbiased estimates of effects of interest.

In most experiments in Education and Psychology, different conditions are well defined. Moreover, in most of the cases, the participants undergo and complete the experiment in the condition they were assigned to. However, there are cases where some participants unintendedly participate in a condition they were not assigned to. This form of non-compliance disturbs the estimation of the treatment effect of interest and hence needs to be accounted for in the analysis. How to do so depends on the nature of non-compliance (e.g., participants allocated to the control condition receiving treatment, participants allocated to the/a treatment condition participating in the control condition, or both) and frequency thereof. Bloom (2008) provides a very useful overview of estimating causal effects in the case of this kind of non-compliance, focusing on two key questions: the average effect of offering treatment (commonly referred to as ‘intent to treat’ or ITT), and the average effect of receiving treatment (also called ‘treatment of the treated’ or TOT). Although ITT can in many cases be estimated quite easily, there is no valid way of estimating TOT, “*because there is no way to know which control group members are counterparts to treatment group members who receive treatment*” (Bloom, 2008, p. 120).

Remainder of This Book

The goal of this book is neither to cover all possible statistical methods out there nor to focus on a particular software package. There are many excellent statistical textbooks on the market that present both basic and advanced concepts at an introductory level and/or provide a very detailed overview of options in a particular statistical software package. Some recent examples are Andy Field's 5th edition (as well as previous editions) of *Discovering Statistics Using IBM SPSS Statistics* (2018), Andy Field and colleagues' *Discovering Statistics Using R* (2012), Thom Baguley's *Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences* (2012), Bradley Huitema's *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies* (2011, 2nd edition; I have not read the 1st edition, hence only the referral to the 2nd edition), David Howell's 8th edition (as well as previous editions) of *Statistical Methods for Psychology* (2017), and Jari Metsämuuronen's impressive three-volume work entitled *Essentials of Research Methods in Human Sciences* (2017). Besides, although perhaps some of the examples were more relevant in former than they are in current times, many of the old sources definitely remain more than worth the read (e.g., Burns & Dobson, 1981; Greenwood, 1989; Kish, 1965; Tacq, 1997), and the same goes for both older and more recent books on research methods and methodologies that may not focus on experimental research but provide a very nice view of what methods and methodologies are out there for us (e.g., Creswell, 2012, as well as previous editions; Coe, Waring, Hedges, & Arthur, 2012). This is not yet another book in this kind of genre. This book focusses on experimental research in two disciplines that have a lot of common ground in terms of theory, experimental designs used, and methods for the analysis of experimental research data: *Education* and *Psychology*. Although the methods covered in this book are also frequently used in many other disciplines, including Sociology and Medicine, the examples in this book come from contemporary research topics in Education and Psychology. The data used are not from actual research but are simulated such that they represent examples of different types of data—that fairly well or clearly do not meet certain assumptions—and allow to discuss different types of approaches to data analysis for different types of variables.

Various statistical packages, commercial and zero-cost Open Source ones, are used. Commercial packages used in this book are *Mplus* version 8 (Muthén & Muthén, 2017), *Stata* 15.1 (StataCorp 2017), and *SPSS* version 25 (IBM Corporation, 2017). Zero-cost Open Source packages used throughout this book are several packages in the *R* programme version 3.5.0 (R Core Team, 2018) and *RStudio* version 1.1.456 (RStudio Team, 2018), *GPower* version 3.1.2 (Buchner, Erdfelder, Faul, & Lang, 2009), *JASP* version 0.9.2.0 (Love, Selker, Marsman, et al., 2018), *Jamovi* version 0.9.5.16 (Jamovi project, 2019), and *SocNetV* version 2.4 (Kalamaras, 2018). In the remainder of this book, I just use the *italic underlined* terms to refer to the versions and references here. Although this book uses both commercial and non-commercial packages, at the time of writing, the non-commercial (i.e., zero-cost Open Source) packages just mentioned are developing at a very fast pace, and already provide researchers with options that are either not available in common commercial packages or in these commercial packages require quite some work and experience.

Each statistical method is discussed in a concrete context of a research question along with a directed (one-sided or one-tailed) or undirected (two-sided or two-tailed) hypothesis and an experimental setup in line with that. Therefore, the titles of the chapters in this book do not include any names of statistical methods such as *analysis of variance* (ANOVA; e.g., Field, 2018; Fisher, 1921, 1925; Howell, 2010, 2017) or *analysis of covariance* (ANCOVA; e.g., Huitema, 2011; Leppink, 2018a, b). In a total of seventeen chapters (of which this one is the first) divided over five parts, this book covers a wide range of topics, research questions, and hypotheses that call for experimental designs and statistical methods, fairly basic or more advanced.

Part I: Common Questions (Chaps. 1–4)

In this first chapter, the QDA bridge is presented, and this heuristic runs throughout the chapters of this first part as well as throughout the chapters of the subsequent parts of this book. The next chapters in this first part of the book cover different approaches to statistical testing and estimation (Chap. 2), important principles of measurement, validity, and reliability (Chap. 3), and ways of dealing with missing data (Chap. 4). The general principles presented in this first part return in each of the subsequent parts of this book.

In Chap. 2 (*Statistical Testing and Estimation*), different approaches to statistical testing and estimation are discussed. Although some of these approaches have been used extensively, other approaches have been introduced but remain underused. Using concrete examples from educational and psychological research, four approaches to statistical testing are compared: traditional *null hypothesis significance testing* (NHST; Bakker, Van Dijk, & Wicherts, 2012; Cohen, 1990; Cohen, 1994; Kline, 2004), *two one-sided tests* (TOST) *equivalence testing* (Goertzen & Cribbie, 2010; Hauck & Anderson, 1984; Lakens, 2017), *information criteria* (Anderson, 2008; Burnham & Anderson, 2002), and *Bayesian hypothesis testing* (Etz & Vandekerckhove, 2018; Rouder, Speckman, Sun, Morey, & Iverson, 2012; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wagenmakers et al., 2018). While the traditional NHST approach has its use, the other three approaches enable researchers to do things that cannot be done with the traditional approach. For instance, in research articles across journals, statistically non-significant *p*-values are interpreted as 'confirming' the null hypothesis. While no statistical approach can provide absolute evidence for (i.e., prove the truth of) either the null or the alternative hypothesis, contrary to the traditional approach, the alternatives can help researchers establish evidence in favour of 'no difference' or a difference so small that it may not be practically relevant. For statistical estimation and linking to statistical testing, CIs and Bayesian posterior intervals aka credible intervals (CRIs) are discussed. While 95% CIs are commonly used for interval estimation and associated with traditional NHST at the 5% statistical significance level, 90% CIs also have their uses, including in the context of TOST equivalence testing where two one-sided tests are carried out at a 5% statistical significance level each. Similarities and differences between approaches for statistical testing and estimation are discussed. Next, a general approach to statistical testing and estimation—which combines elements of each of the aforementioned approaches—is proposed and this approach is applied throughout the chapters of this book.

Many variables that are important in educational and psychological research come from psychometric instruments. Performance or other behaviour may be assessed by different raters. Constructs such as motivation, mental effort, and emotion may be studied through self-report questionnaires. Researchers and educators may use tests in an attempt to measure knowledge or skills. Chapter 3 (*Measurement and Quality Criteria*) introduces important principles of *measurement*, *validity*, and *reliability* (Bovaird & Embretson, 2008; Hoyle, 2008; Iramaneerat, Smith, & Smith, 2010; Kellow & Willson, 2010; Mueller & Hancock, 2010; Osborne, Costello, & Kellow, 2010; Stemler & Tsai, 2010; West & Thoemmes, 2008; Wolfe & Dobria, 2010). Even though researchers commonly report means (*M*s) and standard deviations (*SD*s) for outcome variables of interest, the use of these and other statistics used for group comparisons relies on the assumption that the instruments are valid and reliable. Chapter 3 does not cover all possible aspects of reliability and validity (that would require a book in itself), but focusses on statistical aspects of instrument reliability and validity. Across fields, group comparisons based on data acquired with psychometric instruments have been justified (or sometimes discarded) by researchers through the use of Cronbach's alpha (Cronbach, 1951). Although Cronbach's alpha has been widely interpreted as an indicator of reliability, validity or 'scale quality', Chapter 3 reviews literature that advises against Cronbach's alpha and in favour of alternatives such as McDonald's omega (e.g., Crutzen & Peters, 2017; McDonald, 1978, 1999; Peters, 2014; Revelle & Zinbarg, 2009; Sijtsma, 2009). Chapter 3 discusses some of these alternatives to Cronbach's alpha in a broader approach to psychometrics along with *factor analysis* (Field, 2018; Hoyle, 2000; Osborne et al., 2010; Rummel, 1970; Thompson, 2004), *item response theory* (e.g., Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991), *latent class analysis* and *latent profile analysis* (Ding, 2018; Hagenaars & McCutcheon, 2002; McCutcheon, 1987), *network analysis* (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012; Van Borkulo et al., 2014), *mixed-effects modelling* aka *multilevel modelling* (Hox et al., 2017; Molenberghs & Verbeke, 2005; Singer & Willett, 2003; Snijders & Bosker, 2012; Tan, 2010; Verbeke & Molenberghs, 2000), and *Bland-Altman analysis* (Altman & Bland, 1983; Bland & Altman, 1986, 2003).

For the sake of simplicity of the introduction, Chaps. 2 and 3 do not yet deal with missing data situations. How to deal with missing data depends on the expected nature of the missingness (Cole, 2010; Enders, 2010; Little & Rubin, 2002; Molenberghs & Kenward, 2007; Van Buuren, 2012). More than four decades ago, Rubin (1976) proposed a framework of three types of missing data: *missing complete at random* (MCAR), *missing at random* (MAR), and *missing not at* *random* (MNAR). Nevertheless, missing data are often treated inappropriately. Two common inappropriate approaches to missing data are mean imputation and listwise deletion. In the case of mean imputation, a missing data point is replaced by either the mean of the cases that do have a value on that variable or the mean for the case at hand of the items that do have a value. This tends to result in underestimated *SD*s. In the case of listwise deletion aka 'complete case' analysis, all cases with missing data on at least one variable are excluded from the analysis. For instance, in a paper one may read that some participants were deleted from the analysis due to incomplete data. Listwise deletion comes with an unnecessary loss of information. In another deletion approach, called pairwise deletion, incomplete cases are deleted only for variables where they have missing. Hence, the sample size may vary from one (pair or set of) variables to the next and that may create some challenges for the computation and comparison of *SE*s and for the comparisons of competing models. Moreover, especially when missingness is not MCAR, both mean imputation and the deletion approaches just mentioned are likely to result in biased estimates for differences and relations of interest. Chapter 4 (*Dealing with Missing Data*) discusses several alternatives to the inappropriate approaches just mentioned, with advantages and disadvantages: *last observation carried forward* (Peto et al., 1977), *matching* aka *hot-deck imputation* (Little & Rubin, 2002; Roth, 1994), *regression imputation* (Allison, 2002), *full information maximum likelihood* (FIML; Collins, Schafer, & Kam, 2001), and *multiple imputation* (MI; Rubin, 1987; Schafer, 1997; Van Buuren, 2012). Although no chapter can identify 'the best' solution to any possible missing data situation, Chapter 4 provides some general guidelines and rules of thumb—relating to the type of missingness (MCAR, MAR, MNAR), whether a variable that has missing data has been measured more than once (i.e., repeated measurements), the proportion of missing data on a variable (e.g., 5, 20, 80%), and whether the variable that has missing is an outcome variable or not—that can be used in many more or less common situations.

Part II: Types of Outcome Variables (Chaps. 5–8)

After the general questions covered in Part I, Part II discusses methods and models for four different types of outcome variables, using examples from contemporary research in Education and Psychology: two-category aka *dichotomous* (Chap. 5), *multicategory nominal* (Chap. 6), *ordinal* (Chap. 7), and *quantitative* (Chap. 8). Part II provides an overview of similarities and differences in statistics and interpretation between different types of outcome variables and this helps to discuss frequent misconceptions and inappropriate practices. For instance, in *regression analysis* (e.g., Agresti, 2002; Anderson & Rutkowski, 2010; Baguley, 2012; Field, 2018; Howell, 2017; Huitema, 2011; McCullagh, 1980; Metsämuuronen, 2017; Nussbaum, Elsadat, & Khago, 2010), researchers often resort to dummy coding (i.e., 0/1 coding) not knowing that alternatives are available that, in the light of the research questions and hypotheses at hand, may be more appropriate. Also, in experiments where choice behaviour is an important outcome variable, researchers not rarely compare groups in terms of frequencies on separate categories of an outcome variable not knowing that there are ways to include all categories in a single model. Paradoxically, ordinal variables are sometimes treated as interval/ratio level outcome variables (often in linear models, e.g., Tacq & Nassiri, 2011) and in other cases as multicategory nominal outcome variables. These and other issues are discussed in Part II, and at the end of this part, the reader should understand that, although there are differences in statistics and interpretation between types of outcome variables, the approach to statistical testing and estimation outlined in Part I of this book works for each of the different types of outcome variables.

Chapter 5 (*Dichotomous Outcome Variables*) covers methods for dichotomous outcome variables (e.g., *binary logistic regression*; Agresti, 2002; Field, 2018). Although many experiments involve comparisons between conditions on quantitative variables, pass/fail decisions, recover/fail to recover distinctions, and event occurrence/event absence constitute examples of potentially interesting dichotomous outcome variables. Useful plots, descriptive statistics, and tests are discussed not only for dichotomous variables measured once in time but in the context of event-history analysis of event occurrence as well. Although event-history analysis is commonly associated with *survival analysis* (Hosmer, Lemeshow, & May, 2008; Kleinbaum, 1996; Miller, 1997) in hospitals, (simulated) traffic research for example may focus on the occurrence or absence of accidents in different groups of participants studied.

Chapter 6 (*Multicategory Nominal Outcome Variables*) discusses methods for multicategory nominal outcome variables (e.g., *multinomial logistic regression*; Anderson & Rutkowski, 2010). Although this type of outcome variable is less common than quantitative or dichotomous outcome variables, it may constitute a main outcome variable in for instance experiments that focus on choice behaviour or association as a function of treatment. For example, in research on emotion, different content or different formats of presenting content may trigger different types of emotions and different words with it. These emotions or words may at best be considered nominal rather than ordinal categories. Useful plots, descriptive statistics, and tests are discussed, and many of the concepts discussed in Chap. 5 in the context of dichotomous outcome variables return as well (with extensions to multicategory nominal outcome variables).

A commonly undervalued and mistreated type of outcome variable is the ordinal one. Outcome variables with three or four categories are not rarely treated as interval or ratio level outcome variables or in some cases—unwillingly—as multicategory nominal outcome variables, usually with incorrect outcomes as a consequence. Whether we deal with a performance outcome variable such as a categorisation of diagnostic performance as ‘poor’, ‘satisfactory’, and ‘excellent’ or an effort rating like ‘very little effort’, ‘little effort’, ‘considerable effort’, ‘a lot of effort’ in response to treatment, this type of outcome variables is to be treated as ordinal. Chapter 7 (*Ordinal Outcome Variables*) discusses methods for ordinal outcome variables (e.g., *ordinal logistic regression*; Agresti, 2002, 2010; McCullagh, 1980). In terms of plots, descriptive statistics, and tests, this chapter largely builds forth on the foundations laid in Chap. 5 for dichotomous outcome variables and in Chap. 6 for multicategory nominal outcome variables. Extensions focus on what distinguishes multicategory ordinal outcome variables from dichotomous or multicategory nominal outcome variables.

After Chap. 7, it will be clear to the reader why the distinction between types of outcome variables matters and why for instance ordinal outcome variables need not and should not be treated as interval or ratio outcome variables. A brief recap of the main takeaways of Chaps. 5–7 is provided at the beginning of Chap. 8 (*Quantitative Outcome Variables*) for the reader who opens the book and skips the other chapters in Part II. Next, common statistics of correlation, effect size, and model performance are discussed. Linear models (e.g., Field, 2018; Howell, 2017; Huitema, 2011), non-linear models (e.g., Brown, 2001; Tan, 2010), and models for counts (e.g., *Poisson regression*; Agresti, 2002; Nussbaum et al., 2010) are discussed in Chap. 8. Note that the term 'quantitative outcome variables' is used with the sole intention to make the distinction between outcome variables that are of interval or ratio level of measurement (Chap. 8) and outcome variables that are of nominal or ordinal level of measurement; it is not intended to imply that the latter are only 'qualitative' outcome variables. Also note that in this book, I am not advocating for any kind of qualitative-quantitative divide that appears to be upheld in some fields. Instead, my plea throughout this book is that such divides are unrealistic and unconstructive, because they feed misconceptions like there never being any hypotheses in qualitative research, qualitative research always being exploratory and quantitative research always being confirmatory, qualitative research being non-linear and quantitative research being linear, qualitative research being subjective and quantitative research being objective, that qualitative researchers are constructivists while quantitative researchers are (post-)positivists, that qualitative research is done 'by hand' while quantitative research is done with software, and more of this kind. Any kind of qualitative-quantitative divide fails to appreciate that some core assumptions form part of all research, that research can be qualitative and quantitative at the same time, and that (Van der Zee & Reich, 2018, p. 3): "*For both qualitative and quantitative research, interpretation of results depends on understanding what stances researchers adopted before an investigation, what constraints researchers placed around their analytic plan, and what analytic decisions were responsive to new findings. Transparency in that analytic process is critical for determining how seriously practitioners or policymakers should consider a result*." Especially nowadays with developments in analytics, text mining, and Big Data (Toon, Timmerman, & Worboys, 2016), any kind of possibly meaningful qualitative-quantitative divide has been become a thing of the past. Some philosophers may still defend the 'incompatibility' thesis that states that qualitative and quantitative research cannot go together, just like some other philosophers may still consider combining Frequentist and Bayesian methods a form of 'heresy'. My view on this matter is more pragmatic: just like there are many practical examples of how qualitative and quantitative research can be combined into *mixed methods research* (e.g., Creswell, 2012; Johnson & Onwuegbuzie, 2004; Levitt et al., 2018), Frequentist, Bayesian, and other statistical methods can be combined as well, and most statistical packages these days include these different types of methods (e.g., Jamovi, JASP, SPSS, Stata, and increasing numbers of packages in R).

Part III: Types of Comparisons (Chaps. 9–12)

To facilitate explanation and understanding of the different types of concepts in Part II, comparisons in Part II are limited to a single treatment factor with two groups. In Part III, the different types of outcome variables discussed in Part II are revisited but this time with examples of single treatment factors that comprise three or more groups and absence of specific hypotheses (Chap. 9) or presence thereof (Chap. 10), with examples of two and three factors simultaneously (Chap. 11), and with examples that involve covariates (Chap. 12). This part of the book serves several purposes, including the following. Firstly, there are quite a few different approaches to multiple testing and which one, if any, is needed depends on the set of hypotheses available prior to data collection. Chapters 9 and 10 therefore discuss common approaches to multiple testing and useful approaches to planned comparisons, respectively. Next, Chapters 11 and 12 cover the concepts of main effects, interaction effects, and simple effects. Although when dealing with two or three factors researchers often do study interaction effects, a check on factor-covariate interaction (an important assumption underlying for instance ANCOVA) is often omitted. Chapter 12 therefore compares different methods for factor-covariate combinations.

When a treatment factor consists of more than two groups, applying a correction for multiple testing to keep the rate of false alarms (seeing differences in a sample while in the population sampled from there are none) limited is desirable. This is especially the case when prior to data collection there are no clear expectations or hypotheses that call for one or a few very specific tests. Chapter 9 (*Common Approaches to Multiple Testing*) discusses several types of *correction for multiple testing*—Tukey's (Tukey, 1949), Scheffé's (Bohrer, 1967; Scheffé, 1959), Bonferroni's (Bonferroni, 1936), Dunn's (Dunn, 1979), Dunnett's (Dunnett, 1955), Games-Howell's (Games & Howell, 1976; Games, Keselman, & Clinch, 1979), Holm's (Holm, 1979), and Bayesian prior probability fixing (Westfall, Johnson, & Utts, 1997)—with their advantages and disadvantages. Next, a new approach to (overall and) multiple testing is proposed.

When a treatment factor consists of more than two groups but—based on theory, previous research or common sense—we do have specific hypotheses with regard to which groups differ and eventually in what direction they differ, we may not need to compare all groups with each other but may gain statistical power to detect treatment effects of interest by performing tests in line with the hypotheses we have. For instance, when the independent variable is a dosage of medication and participants in the different groups receive 0, 5, 10, and 15 mg, respectively, and we expect a linear relation between dosage and say performance in a driving simulator, one test may do. In another context, when in an experiment with three conditions we expect that two treatment conditions A and B lead to better driving performance than a control condition and additionally condition B will do better than condition A, two one-sided tests—one for the difference between control and treatment A, and another for the difference between treatments A and B—can do. These and other *planned comparisons*, and how they follow from a specific set of hypotheses, are discussed in Chap. 10 (*Directed Hypotheses and Planned Comparisons*). Finally, Chap. 10 also discusses *preregistration* and *registered reports* as ways to register prespecified hypotheses and justify planned comparisons.

When two or more factors are involved, different types of effects can be distinguished: main effects, interaction effects, and simple effects. Chapter 11 (*Two-Way and Three-Way Factorial Designs*) discusses important guidelines for the testing, estimation, interpretation, and reporting on each of these two types of effects. The distinction between dummy coding and other types of coding discussed in earlier chapters is revisited in Chap. 11 as well to discuss some common misconceptions with regard to the meaning of model coefficients. For instance, when performing a logistic regression on a dichotomous outcome variable with two dummy coded treatment factors A and B (i.e., each 0/1 coded) and their interaction (i.e., a product of 0s and 1s, hence a 0/1 variable as well), researchers at times interpret the regression coefficients of A and B in a model that includes the A-by-B interaction as 'main effects', while in fact they are simple effects.

For the sake of simplicity of the introduction of new topics, Chaps. 9–11 do not cover examples that include covariates. However, in many studies, one or more covariates are added to the analysis. Although adding covariates sometimes makes sense, the reasons for which they are added are not always appropriate (e.g., Gruijters, 2016) and some important assumptions such as with regard to factor-covariate interaction frequently remain unchecked, with inappropriate outcomes and interpretations as a consequence. Chapter 12 (*Factor-Covariate Combinations*) compares different methods for dealing with covariates, including ANCOVA, *analysis of covariate residuals* (ANCOVRES; Kane, 2013; Leppink, 2018b), *path analysis* (e.g., Baron & Kenny, 1986; Hayes, 2018; Preacher & Hayes, 2004, 2008), and *moderated regression* (e.g., Champoux & Peters, 1987; Darrow & Kahl, 1982; Hayes, 2018; Leppink, 2018b).

Part IV: Multilevel Designs (Chaps. 13–16)

In Part IV of the book, different types of hierarchical data structures are discussed. Although there is more awareness among researchers of multilevel designs and multilevel analysis these days compared to say 10–15 years ago, journals across fields continue to regularly publish articles in which a multilevel structure has been ignored or otherwise treated inappropriately. This is problematic, because—as explained earlier in this chapter—inadequately accounting for such structures can substantially distort testing and estimation outcomes. Therefore, in this part of the book, four common types of multilevel situations in experimental research are discussed: participants who interact with each other because they are part of the same social networks (e.g., centres) or because a particular type of interaction is part of the instruction of the experiment (Chap. 13), participants being assessed by two or more raters (Chap. 14), treatment groups being measured repeatedly over time (Chap. 15), and experiments where the type of treatment varies not only between but within participants as well (Chap. 16). Appropriate methods and commonly encountered inappropriate methods are compared in terms of outcomes and interpretation, to help increase awareness of the usefulness of multilevel models when dealing with multilevel data.

In increasing numbers of experiments, participants interact with each other during the experiment, for instance because they are part of the same centres or organisations or because a specific type of interaction is part of the instruction of the experiment. This type of interaction is the focus of Chap. 13 (*Interaction Between Participants*). Different types of situations are discussed: dyads (e.g., couples), small-size groups (e.g., project teams), and larger groups or social networks. In each of these types of situations, individuals are treated as actors nested within higher-level actors or units (e.g., pairs, teams, or centres). Chapter 13 provides examples of how failing to account for this structure can result in substantial distortions of our perspective on treatment effects of interest.

In not so few experiments where performance or another behavioural outcome variable is measured, scores of learners or other individuals result from an assessment by two or more independent raters. The scores of different raters are often averaged into a single score per participant. While this does not always result in incorrect conclusions with regard to treatment effects of interest, a more accurate analytic approach is found in treating the raters as stations that have to be passed by participants. Chapter 14 (*Two or More Raters*) presents worked examples of how to run that type of analysis and acquire estimates of treatment effects and inter-rater reliability simultaneously.

Almost a century ago, Fisher (1925) launched the term *split-plot* design to define agricultural experiments in which split plots of land received different treatments and were measured across time. Later on, this term was also adopted in Psychology, Education, and other disciplines to refer to similar kinds of experiments with human participants: different groups are given a different treatment and are measured at two or more occasions over time. In some cases, there are two measurements one of which takes place before the treatment (i.e., pre-test) and the other takes place after treatment (i.e., post-test). In other cases, there are two measurements one of which takes place immediately after treatment (i.e., post-test) and the second one after some time (i.e., follow up). In yet other cases, there are three or more measurements, some of which before some of which after treatment. Methods for each of these types of situations are discussed in Chap. 15 (*Group-by-Time Interactions*). Analogously to the rater situations discussed in Chap. 14, occasions can be treated as stations to be passed by each of the participants.

In some cases, different groups may receive different treatments at different occasions, the order of treatment varies across groups, and there is a measurement of an outcome variable of interest at each occasion. In the simplest setup, there are two treatments, A and B, which are taken in a different order by each of two groups: A-B in one group (with a measurement after A followed by a measurement after B), B-A in the other group (with a measurement after B followed by a measurement after A). In other cases, there are more treatments and more orders for a larger number of groups with it, or there are only a few—and perhaps only two—treatments which can vary in each of a larger number of trials. At each trial, there is a measurement of an outcome variable of interest. Consider students who are asked to read ten articles, each article is read in each of two possible formats determined in a random order, and after each article students are asked to rate on a visual analogue scale (VAS: 0–100) how much effort it took to read the article. These are all examples of situations where treatment varies both between and within participants and a measurement of an outcome variable of interest takes place in each trial (e.g., after each condition or for each article). As in Chaps. 14 and 15, the occasions or trials can still be viewed as stations to be passed by each of the participants, but there is something that varies from station to station that has to be accounted for in our models. Chapter 16 (*Models for Treatment Order Effects*) provides worked examples for how to do that.

Part V: General Recommendations (Chap. 17)

The final chapter of this book, Chap. 17 (*A General Pragmatic Approach to Statistical Testing and Estimation*), provides a synthesis of Parts I–IV in the form of a set of general recommendations on core questions discussed in this book: design and sample size, ways to increase statistical power, dealing with missing data, psychometrics of measurement instruments, testing and estimating treatment effects, and dealing with covariates. To readers who wonder what about multilevel questions, these are covered as part of the aforementioned core questions. For instance, the distinction between single-level designs and two- or multilevel designs has implications for both design and sample size and has implications for dealing with eventual missing data. In the context of psychometrics of measurement instruments, the final chapter provides a general mixed-effects aka multilevel modelling approach that does not require the use of latent variables and can be applied to all kinds of outcome variables covered in this book.

2 Statistical Testing and Estimation

Abstract

In Chap. 1, the QDA bridge is presented as an approach to experimental research. In this chapter, a pragmatic approach to statistical testing and estimation (PASTE) is presented. In line with the QDA heuristic introduced in Chap. 1, all statistical testing and estimation is driven by research questions and hypotheses and appropriately accounts for the features of the experimental design and the data acquired in the experiment. To get a first grip on the data, to check which assumptions may be reasonable, and to decide on how to proceed with the numbers, we ought to first carefully inspect our data with graphs and, based on what we see in these graphs, with appropriate simple descriptive statistics. These graphs and descriptive statistics, together with the research questions and hypotheses at hand and the features of the experimental design to be accounted for, help us to decide on appropriate ways to test our hypotheses and to estimate effects of interest. Point estimates are preferably accompanied by CIs and/or CRIs. Statistical testing is preferably done using multiple criteria, though which criteria to use partly depends on the nature of the hypotheses to be tested. In this chapter, different approaches to statistical significance testing as well as different approaches to hypothesis testing using information criteria and Bayesian hypothesis testing are discussed and compared, and concepts of sequential analysis and machine learning are discussed as well. After the introduction of each of the aforementioned concepts and methods, this chapter concludes with a coherent set of general guidelines that may serve as a general pragmatic approach to statistical testing and estimation. This approach, called PASTE, is also used in all subsequent chapters of this book. In a nutshell, PASTE is not about preferring one approach over another (e.g., Bayesian over statistical significance) but about using combinations of methods to make evidence-based decisions with regard to findings at hand and to formulate appropriate recommendations for future research and practice.

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research*
in Education and Psychology, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_2

Introduction

The QDA bridge, the heuristic introduced in Chap. 1, states that there is a bridge between research questions and hypotheses (i.e., *questions*), experimental design and methods used to collect data (i.e., *design*), and statistical analysis (i.e., *analysis*). That is, both *questions* and *design* inform *analysis*. Moreover, once we have collected our data, we will also have to see which of the likely candidates for statistical analysis (i.e., in the light of questions and design) are appropriate given the nature of the data we have collected. For instance, the presence of several very extreme cases, extreme skewness and other potentially severe departures from normality constitute possible reasons for us to revise initial choices for statistical methods that do not need normally distributed data but may well provide somewhat inadequate results when used with particular data unless we take additional steps, such as applying a transformation (e.g., taking the square root or log of a no-zero-values time variable that is clearly skewed to the right; Field, 2018), before we use these methods (e.g., Osborne, 2010a; and see Chaps. 3, 5, and 8 of this book for more on this and other ways to deal with departures from assumptions). Calculating *M*s, *SD*s, correlations or other statistics without any kind of graphical inspection of our data first may be a recipe for disaster.

Some readers may wonder if 'disaster' is not somewhat of a too heavy qualification for possible outcomes of the exercise of calculating numbers without any kind of visual check with appropriate graphs. As a matter of fact, from time to time, medical doctors and educational researchers complain to me that statisticians (including myself) can be very strict (or, in their words, harsh) in their reviews when it comes to the design and analysis of experiments or other empirical studies. In their view, educational and psychological research are not 'rocket science', we 'just use' statistics as a tool to draw some conclusions, and we cannot all be statisticians. I fully agree with the latter: we cannot all be statisticians, just like we cannot all be medical doctors. However, if I as a statistician—with no medical background—provided some medical advice to people that is not based on solid evidence but is inadequate, medical doctors might get very upset with me, and right so, especially if my advice put the health of people who took my advice at risk or worsened health issues they had been suffering from already. Likewise, when people use statistical methods without following some basic rules, this will likely upset statisticians, because the recommendations for future research and practice following from this work undermine scientific practice, may harm society or at least put specific groups of people at risk of being exposed to potentially harmful treatments, and may result in an unnecessary loss of tax payers' confidence in scientific practice (in the field at hand or more broadly). Therefore, just like statisticians should refrain from providing inadequate medical advice, medical doctors and other researchers should not violate the rules of good scientific practice by inventing and using their own rules. Good methodological and statistical practice is a moral and ethical obligation to our fields, to our colleagues, to next generations of researchers, and to society. This is not to say that every problem has one appropriate statistical solution. On the contrary, most problems in educational and psychological research have several possible solutions, and the latest when that becomes evident is when you ask different teams of researchers to independently analyse the same data (e.g., Leppink & Pérez-Fuster, 2019; Silberzahn et al., 2018) or the same team of researchers illustrate how different ways of analysing the same data may shed light on different aspects of an effect of interest (e.g., Twisk et al., 2018; Twisk, Hoogendijk, Zwijsen, & De Boer, 2016).

Very good sources on graphical checks include Field (2018), Metsämuuronen (2017), Osborne (2010), and Osborne and Overbay (2010), and graphical checks can be done with virtually any statistical software package. Depending on the nature of an outcome variable of interest, pie charts (for nominal variables) or bar charts (for nominal and ordinal variables) with frequencies, or box plots and histograms (for quantitative outcome variables) per condition or group can help us acquire a good first understanding of the distribution of a response variable of interest in each of the conditions. Depending on what this distribution looks like in the different conditions, some methods will constitute better candidates for analysis than other methods. That said, it is recommendable to not restrict ourselves to one statistical method when writing a study proposal but instead to mention a few candidate methods some of which may become more or less likely depending on what the data to be collected looks like. In Chaps. 5, 6, 7, 8 of this book, this exercise is done for the different outcome variables we may encounter in experimental research. For the sake of simplicity, we will introduce a *pragmatic approach to statistical testing and estimation* (PASTE) in this chapter in a context of two-group experiments with quantitative outcome variables that are distributed such that no transformations or decisions with regard to 'extreme' cases have to be made and a good old *Student's t-test* (e.g., Field, 2018; Fisher Box, 1987) constitutes an appropriate way of testing hypotheses.

Example Experiment

Researchers in a Health Professions Education department are interested in the effect of a new type of simulation training on the development of communication skills among undergraduate medical students. The researchers decide to do an experiment in which this new simulation training constitutes the experimental treatment condition and the conventional form of simulation serves as control condition. In both conditions, participating undergraduate medical students individually undergo training with the same type of simulated patients. The only way in which the two conditions differ is that specific instructions are provided during the training in the treatment condition but not in the control condition. At the end of the training, participants from both conditions individually complete the same post-test with a simulated patient. For simplicity of the example, this post-test yields a quantitative integer score that can range from 0 to 10, and higher scores indicate better post-test performance. Suppose, the researchers have no pronounced expectations for a potential effect of the additional instructions in the treatment condition to be positive (i.e., on average, the treatment condition results in better performance than the control condition) or negative (i.e., on average worse performance in the treatment condition compared to the control condition), for example because the evidence with regard to providing this kind of instruction in the literature thus far is mixed. Therefore, the researchers want to test the null hypothesis of 'no difference' between conditions in average post-test performance ($\boldsymbol{H_0}$: no treatment effect, $\mu_{\text{treat}} = \mu_{\text{control}}$), and their alternative hypothesis is that there is a difference, positive or negative ($\boldsymbol{H_1}$: treatment effect, $\mu_{\text{treat}} \neq \mu_{\text{control}}$).

Required Sample Size Calculation

The researchers decide to randomly sample $N = 128$ undergraduate medical students and randomly allocate them such that both conditions host $n = 64$ participants. They do so, because they know that these numbers yield a statistical power of 0.80 for a difference between conditions of 0.50 *SD*, using a two-sided (i.e., $\boldsymbol{H_1}$: $\mu_{\text{treat}} \neq \mu_{\text{control}}$) Student *t*-test (introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland, who used "Student" as his writing name; see for instance: Fisher Box, 1987) at statistical significance level $\alpha = 0.05$. Required sample size calculations like this one can be done easily in zero-cost software like *GPower* and *Jamovi*. A statistical power of 0.80 means that, assuming an effect of a specified magnitude (here: 0.50 *SD*), on average eight out of every ten experiments with the given sample size (here: $n = 64$ per condition) and statistical significance level (here: $\alpha = 0.05$) would yield a statistically significant result (here: $p < 0.05$). A power of (at least) 0.80 is not only recommendable in the light of having a decent chance of detecting an effect of interest in a new experiment; it is simply necessary to do meaningful replication research. Assuming a power of 0.80, the chance that two independent experiments in a row—let us call them an initial experiment and a replication experiment—yield a statistically significant outcome is 0.80^2 (i.e., 0.80 * 0.80) or 0.64. In other words, even with two experiments that each have a statistical power of 0.80, there is only 64% chance that both experiments will yield a statistically significant outcome. If we add a third independent experiment with the same power of 0.80, say a second replication of the initial experiment, the chance that all three experiments with a power of 0.80 yield a statistically significant outcome equals 0.80^3 (i.e., 0.80 * 0.80 * 0.80) = 0.512, hence hardly more than 50%.

Assumptions

The researchers in the example experiment are aware of all this and would like to strive for a higher statistical power, but do not have more resources to do so. As planned, they succeed in having 128 participants, 64 per condition, complete the experiment. As recommended, before they proceed with statistical testing and

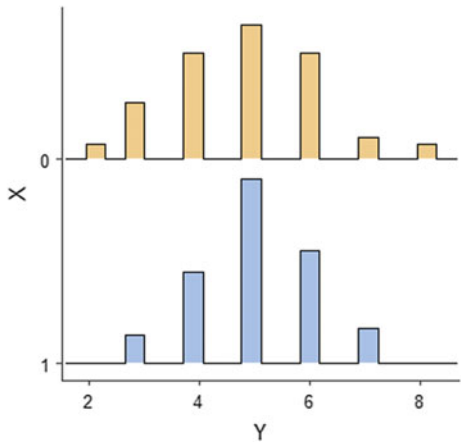


Fig. 2.1 Histogram of the distribution of post-test performance (Y) in the control (X = 0) and treatment (X = 1) condition (*Jamovi*)

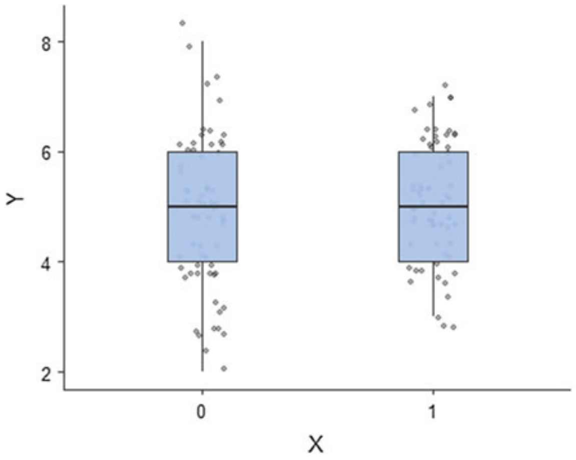


Fig. 2.2 Boxplot of the distribution of post-test performance (Y) per condition (X = 0: control; X = 1: treatment) (*Jamovi*)

estimation, they first graphically inspect their data. Figures 2.1 and 2.2 present the histograms and box plots, respectively, of the distributions of post-test performance in the control (X = 0) and treatment (X = 1) condition (*Jamovi*).

Based on these graphical checks, the researchers proceed with descriptive statistics. In the control condition, post-test performance was M = 4.844 (SD = 1.312; skewness = 0.080, kurtosis = −0.078). In the treatment condition, post-test performance was M = 5.078 (SD = 1.013; skewness = −0.066; kurtosis = −0.294). Post-test score ranged from 2 to 8 in the control group and from 3 to 7 in the treatment condition, and the median post-test score was 5 in both conditions.

The histograms and skewness and kurtosis values indicate that the distributions of post-test performance in both conditions do not deviate that much from normality. In linear regression models, the residuals around the best linear unbiased estimate (BLUE; the regression line) are assumed to follow a Normal distribution in the population of interest. Small departures from normality are often not a problem, especially in somewhat larger samples. Moreover, following the *central limit theorem* (e.g., Field, 2018), even if there are departures from the aforementioned normality assumption, with increasing sample size the *sampling distribution* of *M*s and mean differences (*M_d*s) more and more closely approximates a Normal distribution. When dealing with somewhat smaller samples, software like the *userfriendlyscience R* package (Peters, 2017) can help researchers to estimate the sampling distribution given the data at hand. In the example experiment, the population sampled from could well follow a Normal distribution *and* otherwise sample sizes are large enough to safely assume an approximately Normal sampling distribution.

Another assumption to be checked relates to the *SD*s (or, in squared form: the variances) of the two conditions in the population sampled from: they may be (approximately) equal, or (clearly) unequal. Many researchers test both the normally distributed population and equal *SD*s assumptions through statistical significance tests (e.g., Fasano & Francheschini, 1987; Justel, Peña, & Zamar, 1997; Levene, 1960; Shapiro & Wilk, 1965; Smirnov, 1948; Stephens, 1974). However, this approach is not without problems. Firstly, absence of evidence is not evidence of absence; a statistically non-significant *p*-value generally cannot and should not be interpreted as evidence in favour of 'no deviation' from normality, of 'no difference' between *SD*s, et cetera. Secondly, as with any statistical significance test, small samples usually leave researchers poorly equipped to detect potentially meaningful differences, while large samples may result in statistically significant differences that from a practical point of view may not have much meaning if any. Thirdly, any time, researchers can report *both* a *t*-test assuming equal *SD*s (i.e., the classical Student's *t*-test; Fisher Box, 1987) *and* a *t*-test assuming unequal *SD*s (Welch, 1947). In the classical Student's *t*-test, the number of *degrees of freedom* (*df*) equals $N - 2$, hence here $df = 126$. In Welch's *t*-test, the departure from equal *SD*s is accounted for by lowering *df*. The more the departure from equal *SD*s observed in the experiment, the more these two *t*-tests can be expected to differ; if the difference between *SD*s observed in an experiment is fairly small, the two *t*-tests will usually yield very similar outcomes. In the example experiment, the largest *SD* is $1.312/1.013 \approx 1.295$ times the smallest *SD*. Some may call this difference substantial, others may call it fairly small. For the data at hand, assuming equal *SD*s we find $t_{126} = 1.131$, $p = 0.260$, and assuming unequal *SD*s we find $t_{118.4} = 1.131$, $p = 0.260$. Assuming equal *SD*s, the 95% CI for the M_d (of 0.234) extends from −0.176 to 0.644, and under unequal *SD*s the 95% CI for that M_d extends from −0.176 to 0.645. In other words, under equal and unequal *SD*s, we obtain almost identical testing and estimation outcomes. The M_d of 0.234 (in favour of the treatment condition) corresponds with a difference of 0.200 *SD*s, Cohen's $d = 0.200$ (Cohen, 1988). Although this kind of effect size estimates is to be always evaluated in the context of a study at hand, most researchers would agree that *d*-values of this kind generally reflect relatively 'small' effects. The 95% CI for Cohen's *d* (of 0.200) extends from −0.148 to 0.547. The treatment explains about

1% of the variance in post-test performance (i.e., $R^2 = 0.010$) or perhaps even less (adjusted $R^2 = 0.002$, which corresponds with 0.2% of variance explained).

In a nutshell, the researchers conclude that they do not have sufficient evidence to reject $\boldsymbol{H_0}$ of 'no difference'. The *p*-value obtained from a two-sided test equals 0.260, which is larger than the pre-specified statistical significance level. Consequently, the 95% intervals for the M_d and for Cohen's *d* include '0', the difference specified under $\boldsymbol{H_0}$. Cohen's *d* and the model fit statistics of R^2 and adjusted R^2 indicate that the effect observed in the sample is small. However, these findings should *not* be interpreted as evidence of 'no difference'. No statistical method should be expected to provide absolute evidence for any hypothesis, and that would rarely if ever be the goal of statistics anyway, but if we want to establish evidence in favour of $\boldsymbol{H_0}$ relative to $\boldsymbol{H_1}$, we need one or more alternatives to this 'no difference' NHST approach.

Likelihoods and Ratios

Already more than five decades ago, philosopher Ian Hacking (1965) suggested an intuitive way of thinking about comparative support for $\boldsymbol{H_0}$ relative to $\boldsymbol{H_1}$ and vice versa: dataset A provides evidence for $\boldsymbol{H_1}$ more than for $\boldsymbol{H_0}$ if A is more probable under $\boldsymbol{H_1}$ than under $\boldsymbol{H_0}$. In that case, the ratio of likelihoods aka *likelihood ratio* (LR) of $\boldsymbol{H_1}$ over $\boldsymbol{H_0}$ exceeds 1, and since the LR of $\boldsymbol{H_0}$ over $\boldsymbol{H_1}$ is the inverse of the aforementioned this one is smaller than 1. In line with this approach, Royall (1997) distinguished three questions: (1) what to *believe* now that we have seen A, (2) how to *act* now that we have seen A, and (3) how to interpret A as *evidence* regarding $\boldsymbol{H_0}$ versus $\boldsymbol{H_1}$? In the view of Royall, the *belief* question is captured by *Bayesian* posteriors, the *act* question is in line with *Frequentist* approach of which NHST is part, and the approach of the LR is captured by scholars of the *Likelihoodist* school of thought. Moreover, according to Royall, all information or evidence in a sample is contained in the likelihood function and can be captured in LRs. In the words of Royall (2004, p. 129), "*evidence has a different mathematical form than uncertainty. It is likelihood ratios, not probabilities, that represent and measure statistical evidence* […]. *It is the likelihood function, and not any probability distribution, that shows what the data say.*" Note that the LR [$P(\mathrm{O}|\boldsymbol{H_1})/P(\mathrm{O}|\boldsymbol{H_0})$] is what in Bayes' theorem (Bayes, 1763; Laplace, 1812) constitutes the shift from prior odds [$P(\boldsymbol{H_1})/P(\boldsymbol{H_0})$] to posterior odds [$P(\boldsymbol{H_1}|\mathrm{O})/P(\boldsymbol{H_0}|\mathrm{O})$]:

$$P(\boldsymbol{H_1}|\mathrm{O})/P(\boldsymbol{H_0}|\mathrm{O}) = [P(\mathrm{O}|\boldsymbol{H_1})/P(\mathrm{O}|\boldsymbol{H_0})] * [P(\boldsymbol{H_1})/P(\boldsymbol{H_0})].$$

Thus, the LR can be viewed as a shift from prior odds to posterior odds that provides a measure of relative support for $\boldsymbol{H_1}$ versus $\boldsymbol{H_0}$ or (when numerator and denominator are switched in prior odds, LR, and posterior odds) vice versa. Now, when we deal with continuous variables—such as distributions of *M*s or M_ds—and composite hypotheses (i.e., 'a difference' or 'values in a given range'), we need a generalised LR based on Wilks's theorem (Wilks, 1938). According to this theorem, the log (generalised) LR (i.e., the difference in *deviance* or −2LL of two competing models, one of which is an extension or reduction of the other) converges to a χ^2-distribution as the sample size goes to infinity, if $\boldsymbol{H_0}$ is true. $\boldsymbol{H_0}$ is rejected if the difference in log likelihood between the model called $\boldsymbol{H_0}$ and the model called H_1 exceeds the critical χ^2-value at $\alpha = 0.05$ for *df* = the difference in number of parameters between the two models.

Applying this generalised LR test to our example experiment, we are dealing with a χ^2-distribution with *df* = 1, hence χ_1^2. Just like in the case of two groups ($df_{\text{groups}} = 1$) the *F*-distribution is a squared *t*-distribution, the χ_1^2-distribution is a squared *z*-distribution (i.e., *z* is the standard Normal distribution). As sample size goes to infinity, the *t*-distribution more and more approaches the *z*-distribution and the *F*-distribution more and more approaches the χ^2-distribution. The critical χ_1^2-value at $\alpha = 0.05$ is 3.84. The deviance (−2LL) for model '$\boldsymbol{H_0}$' or 'Model 0' (aka the 'null model') equals 403.138, while that for model '$\boldsymbol{H_1}$' or 'Model 1' is 401.844. Note that the −2LL of the model that includes a parameter that is not included in the simpler model—here: the treatment effect is part of Model 1 but not of Model 0—is always lower (i.e., never higher) than that of the simpler model. However, the generalised LR test (in the literature often simply called 'LR test', even though the generalised LR is a generalisation of the 'simple' LR from the above equation) indicates whether this reduction in −2LL is statistically significant. The difference in deviance or −2LL is used as the observed χ_1^2-value and is found as follows:

$$\text{observed}\,\chi_1^2\text{-value} = [-2\text{LL Model 1}] - [-2\text{LL Model 0}].$$

In our example experiment, we find: observed χ_1^2-value = 403.138 − 401.844 = 1.294; $\chi_1^2 = 1.294$ corresponds with $p = 0.255$. The difference in *p*-value of 0.260 in the *t*-test and 0.255 in this generalised LR test arises because although the t_{126}-distribution more closely approaches a *z*-distribution (i.e., given *df* = 1, the square root of χ^2 equals *z*) than *t*-distributions for smaller *df* it is still a bit wider than the *z*-distribution.

Note that with the generalised LR test, we are in the end back to a statistical significance test. Another ratio approach that relates to *p*-values is found in the Vovk-Sellke maximum *p*-ratio (VS-MPR; Sellke, Bayarri, & Berger, 2001). Based on a two-sided *p*-value, as in our example experiment, the maximum possible odds in favour of $\boldsymbol{H_1}$ over $\boldsymbol{H_0}$ can be calculated as follows provided that $p < 0.37$:

$$\text{VS-MPR} = 1/[-\text{e} * p * \log(p)].$$

For the example experiment, that yields VS-MPR = 1.050, indicating that this *p*-value is at most 1.050 times more likely to occur under $\boldsymbol{H_1}$ than under $\boldsymbol{H_0}$. For $p = 0.05$, VS-MPR would be around 2.46.

There Is More to Statistical Significance Testing

Whether we use p-values from t-tests, LR tests, or the VS-MPR, evidence in favour of $\boldsymbol{H_1}$ over $\boldsymbol{H_0}$ is more easily established in larger samples than in smaller samples. That is, given effect size and statistical significance level, larger samples have a higher statistical power than smaller samples. Underpowered experiments (i.e., given a particular test and effect size) have constituted a source of concern for a long time (Montgomery, Peters, & Little, 2003; Whisman & McClelland, 2005), and should be a concern not only for single experiments but for meaningful replication research as well. Experiments with a power of 0.50 for effects of a realistic magnitude (e.g., $0.2 < d < 0.5$) are not uncommon in educational and psychological research, and following the aforementioned formula of '*statistical power per experiment ^ the number of independent experiments*', we learn that there is only 25% chance that two independent experiments with a power of 0.50 both yield a statistically significant result and only 12.5% that three independent experiments with a power of 0.50 yield a statistically significant result. These days, readers can find many good resources for power analysis and required sample size calculations for different types of designs, including: *GPower*; Dong, Kelcey, and Spybrook (2017), Dong, Kelcey, Spybrook, Maynard (2016), Dong and Maynard (2013), *Jamovi*; Kelcey, Dong, Spybrook, and Cox (2017), Kelcey, Dong, Spybrook, and Shen (2017), Spybrook, Kelcey, and Dong (2016).

Two-Sided or One-Sided Testing?

Had the researchers prior to the example experiment had the alternative hypothesis that the treatment has a *positive* effect ($\boldsymbol{H_1}$: $\mu_{\text{treat}} > \mu_{\text{control}}$), a one-sided test could have been defended, and $t_{126} = 1.131$ would have corresponded with $p = 0.130$ and a 95% CI ranging from −0.092 to ∞. If on the contrary, the researchers prior to the example experiment had the alternative hypothesis that the treatment has a *negative* effect ($\boldsymbol{H_1}$: $\mu_{\text{treat}} < \mu_{\text{control}}$), a one-sided test in the other direction would have been defendable, and $t_{126} = 1.131$ would have corresponded with $p = 0.870$ and a 95% CI ranging from −∞ to 0.491. Doing a one-sided test because prior to the experiment we expect a difference in a particular direction can help us gain statistical power (higher effectiveness) or allows us to achieve the same statistical power with a somewhat smaller sample size (higher efficiency). For example, with a two-sided test at $\alpha = 0.05$ and $d = 0.50$ in the population of interest, we need $n = 64$ participants per condition ($N = 128$), whereas with a one-sided test we need only $n = 51$ participants per condition ($N = 102$), to achieve a statistical power of 0.80 (e.g., *GPower*, *Jamovi*). Registering your hypotheses prior to data collection through for instance research proposals approved for funding as well as *registered reports* (Center for Open Science, 2018) constitutes the best practice for defending one-sided testing. In the traditional publication system, peer review takes place only *after* the data have been collected, and there is no way other than perhaps through a check of the original funded research proposal to check if researchers had reasons to perform a one-sided test prior to seeing the data indeed. In registered reports, peer review of the introduction and method section is done *prior* to data collection, and a provisionally accepted manuscript will be published regardless of the findings if authors comply with the protocol agreed after initial peer review, or can explain minor deviations where they occur (e.g., unexpected attrition or logistic challenges). Registered reports constitute a powerful approach to defending one-sided testing where that is reasonable.

Sequential Testing

Another reason to consider registered reports lies in the possible use of *sequential testing* (Armitage, McPherson, & Rowe, 1969; Dodge & Romig, 1929; Lakens, 2014; Pocock, 1977; see also Chap. 10). In the practice of research involving human participants, an "*important question is how we are going to be able to increase sample sizes without greatly reducing the number of experiments one can perform*" (Lakens, 2014, p. 702). Contrary to a fixed sample size planned a priori, sequential analyses are about continuing data collection until an interim analysis reveals a statistically significant difference. There are three main arguments against such sequential analyses. Firstly, Bayesians would argue that sequential analysis constitutes a valid practice to data collection and testing in a Bayesian but not in a Frequentist approach (e.g., Wagenmakers, 2007). Secondly, doing sequential analyses means more statistical significance tests. Although these tests are not statistically independent (e.g., a test after $n = 15$ also involves the data from the previous test which was done after say $n = 13$), there will be an increase in Type I error probability and this needs to be accounted for by adjusting the statistical significance level downward (e.g., Lakens, 2014; Simmons, Nelson, & Simonsohn, 2011). Thirdly, as smaller samples come with wider CIs and more fluctuation across studies, follow-up studies will be needed to provide more accurate effect size estimates. Despite these arguments against sequential testing, there is evidence that doing sequential analyses with an appropriately adjusted statistical significance level (i.e., in response to the second argument against) can help researchers to reduce the sample size of studies by 30% or more (Lakens, 2014). Moreover, there is software to help researchers determine how to adjust the statistical significance level such that the overall statistical significance level remains nicely below 0.05. Sherman's (2014) package in *R* called *phack* (from *p*-hacking, e.g., Bakker, Van Dijk, & Wicherts, 2012) constitutes a nice example of that.

If you already have a in mind a limited number of interim tests before data collection, you may not even need simulations such as in *phack* but can find appropriate adjustments of the statistical significance level in literature such as Fleming, Harrington, and O'Brien (1984), Lai, Shih, and Zhu (2006), Lakens (2014) and Pocock (1977). There are different approaches to how to correct the statistical significance level alpha, the easiest one of which perhaps comes from Pocock (1977). Suppose that we want to do a two-group experiment and would normally strive for a total of $N = 128$ participants or $n = 64$ per condition for reasons explained in the example experiment. Suppose, we would do an interim analysis halfway, so at $N = 64$ hence $n = 32$ per condition. That would constitute two statistical significance tests: one halfway and one at the end. Using $\alpha = 0.0294$ for each of the two tests would keep the overall Type I error probability (α) at 0.05. If we were to split the final total N not in two but three equal parts (e.g., first interim analysis after $N = 50$, second interim analysis after $N = 100$, and final analysis after $N = 150$), using $\alpha = 0.0221$ for each of the three tests would keep the overall α at 5%. For four equal parts that would come down to $\alpha = 0.0182$ for each test, and for five equal parts that would mean $\alpha = 0.0158$ per test. In other approaches (e.g., Fleming, Harrington, & O'Brien et al., 1984; Lai, Shih, & Zhu, 2006), the alpha of interim analysis would be much smaller than 0.05 and the alpha of the final analysis could remain at 0.05 or slightly below 0.05, to provide a penalty for larger variability in smaller samples.

Some readers may wonder if Pocock's approach provides a sufficiently strong correction of alpha indeed and whether a Bonferroni correction to multiple testing is not more appropriate here. The answer to this is that the Bonferroni correction constitutes a slightly conservative way of Type I error probability inflation correction in the case of statistically independent tests. For example, consider a three-group experiment with no a priori expectations with regard to group differences. Researchers first perform an omnibus one-way ANOVA and follow up on a statistically significant outcome of that omnibus test with three comparisons: condition 1 versus condition 2, condition 1 versus condition 3, and condition 2 versus condition 3. These $k = 3$ tests can be considered statistically independent, and the probability of at least one Type I error can be calculated as follows:

$$\alpha_{\text{total}} = 1 - \left[(1 - \alpha_{\text{per-comparison}})^{k}\right].$$

For $k = 2$ (i.e., two independent tests), $\alpha_{\text{total}} = 0.0975$; for $k = 3$, $\alpha_{\text{total}} = 0.142625$. Bonferroni slightly conservatively corrects for that by dividing $\alpha_{\text{per-comparison}}$ by k and using that as corrected alpha for each test. Hence, for $k = 2$, Bonferroni would state a corrected alpha of 0.025, and for $k = 3$, that would be about 0.0167. Pocock's recommended corrected alpha values of 0.0294 for two tests and 0.0221 for three tests are slightly higher than the Bonferroni approach and slightly higher than the corrected alpha values one would achieve if using the above formula: for $k = 2$, $\alpha_{\text{per-comparison}} \approx 0.0253$, and for $k = 3$, $\alpha_{\text{per-comparison}} \approx 0.0169$. This is because the tests in sequential testing are not statistically independent. After all, every subsequent test also includes the data used in previous tests (plus new data). While the above formula is correct for independent tests, it exaggerates the inflation of Type I error probability when there is dependency between tests. The same applies to updating meta-analyses with new studies.

Equivalence Testing

Note that thus far, the story has been one of 'no difference' NHST. One might argue that in our target populations of interest, treatment effects and other differences are rarely if ever *exactly* zero, and consequently, with sample sizes going to infinity we could reject any 'no difference' null hypothesis. This is what Meehl (1990) also calls the *crud factor*: "*in social science everything correlates with everything to some extent, due to complex and obscure causal influences*" (p. 125). However, many treatment effects or differences alike may be so small that a community may agree that from a practical point of view they are too small to really matter. From this perspective, we need an approach that allows us to test if a treatment of interest is in that 'too small to matter' range, and the traditional 'no difference' NHST approach does not allow us to do that. An approach that *does* enable us to address that question is TOST equivalence testing. In TOST equivalence testing, we first have to agree—in the scientific community—on the range of 'too small to matter' treatment effects. In the context of the example experiment, researchers in the field may agree that *d*-values in the range of −0.3 to 0.3 are too small to matter. This constitutes the range of *relative equivalence*, relative because we are dealing with a range of values not with an absolute equivalence or '0' difference. This range of −0.3 to 0.3 is translated into two null hypotheses:

$$H_{0.1}\colon d < -0.3 \text{ (i.e., more negative treatment effect);}$$
$$H_{0.2}\colon d > 0.3 \text{ (i.e., more positive treatment effect).}$$

The alternative hypothesis, $\boldsymbol{H_1}$, captures the range of 'too small to matter', hence $-0.3 \leq d \leq 0.3$. If and only if *both* $\boldsymbol{H_{0.1}}$ *and* $\boldsymbol{H_{0.2}}$ can be rejected, we declare sufficient evidence to assume relative equivalence. Lakens (2017, 2018) developed an *R* package called *TOSTER*, which enables researchers to do TOST equivalence testing for a range of situations, including the type of data collected in the example experiment. This package can be run in *R* and *RStudio* and is also incorporated in *Jamovi*. Doing this for example experiment, we find—for both assuming and not assuming equal *SD*s—for $\boldsymbol{H_{0.1}}$: $d < -0.3$, $p = 0.003$ and for $\boldsymbol{H_{0.2}}$: $d > 0.3$, $p = 0.286$. In other words, we can reject $\boldsymbol{H_{0.1}}$ but fail to reject $\boldsymbol{H_{0.2}}$.

Earlier in this chapter, we saw that 95% CIs include '0' if a *p*-value obtained from a 'no difference' null hypothesis significance test is not statistically significant at the 5% level. The 90% CI (i.e., general: 1–2α) is relevant in the context of TOST equivalence testing, because two tests are carried out at 5% each. If both $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$ yield statistically significant *p*-values, the 90% CI for *d* includes neither values in the range specified under $\boldsymbol{H_{0.1}}$ nor values in the range specified under $\boldsymbol{H_{0.2}}$. In the example experiment, the 90% CI of *d* extends from −0.092 to 0.491. This explains why $\boldsymbol{H_{0.1}}$, also called the lower bound null hypothesis, can be rejected, but $\boldsymbol{H_{0.2}}$, which is also called the upper bound null hypothesis, cannot be rejected. In other words, where researchers might at first erroneously interpret $p = 0.260$ obtained from the 'no difference' null hypothesis significance test as evidence in favour of ‘no difference’, TOST indicates that we have insufficient evidence to assume relative equivalence.

What TOSTER also enables researchers to do is required sample size calculations given the equivalence bounds, statistical significance level used in the two tests, and desired statistical power. For instance, using $\alpha = 0.05$ for the two tests with equivalence bounds $d = -0.3$ and $d = 0.3$, for a statistical power of 0.80 we would need 190.3077 participants per condition ($n = 191$), hence 382 participants in total ($N = 382$). If we are okay with a power of 0.70, the numbers would be 159.7622 ($n = 160$) and 320 (N), respectively. In other words, establishing significant evidence in favour of relative equivalence is more difficult than it may appear at first. That said, in a meta-analysis across a series of experiments, we may establish such evidence even if individual experiments do not.

Information Criteria

A different approach to hypothesis testing altogether is found in *information criteria*. Although quite a few criteria have been developed, some of which may be more useful than others in a given context, put very briefly they can help researchers to decide which model of a set of competing models ought to be preferred in the light of the data, not necessarily because that model is ‘true’ or represents ‘truth’ but because it is best in terms of predictive accuracy. Adding meaningful variables (e.g., large treatment effects) to a model will likely increase predictive accuracy, whereas adding not so meaningful variables (e.g., small treatment effects) will unlikely increase predictive accuracy. Different criteria differ in the extent to which they penalise for adding not so meaningful variables.

Akaike’s Information Criterion

A first commonly encountered information criterion is Akaike’s information criterion (AIC; Akaike, 1973, 1992). Given a set of competing models, the preferred model is the one with the lowest AIC value. In the example experiment, we find AIC = 407.138 for Model 0 ($\boldsymbol{H_0}$: no difference) and 407.844 for Model 1 ($\boldsymbol{H_1}$: difference). In other words, Model 0 aka ‘no treatment effect’ is to be preferred. A concept that is closely related to the aforementioned (generalised) LR test is that of relative likelihood:

$$\text{relative likelihood} = \exp[(\text{AIC}_{\text{min}} - \text{AIC}_{\text{alt}})/2].$$

In our example experiment, we find about 0.703. Although AIC constitutes a useful criterion, it has somewhat of a preference towards somewhat more complex models. This tendency does not constitute much of an issue in large samples but may result in researchers preferring too complex models especially when sample sizes are small (Claeskens & Hjort, 2008; Giraud, 2015; McQuarrie & Tsai, 1998). Several adaptations of AIC have been developed to correct for this tendency, including AICc (Burnham & Anderson, 2002; Cavanaugh, 1997; Hurvich & Tsai, 1989). With increasing sample size, the tendency of AIC towards overly complex reduces, and as the sample size goes to infinity, the difference between AIC and AICc goes to zero (Burnham & Anderson, 2004).

Schwarz' Bayesian Information Criterion

Another commonly encountered information criterion is Schwarz's Bayesian information criterion (BIC; Schwarz, 1978). AIC and BIC use the same information from the likelihood function, but BIC provides a more severe penalty for adding perhaps not so meaningful variables and consequently tends towards simpler and sometimes somewhat too simple models (Weakliem, 1999). Like with AIC, the model with the lowest BIC is generally the one to be preferred. In the example experiment, we find BIC = 412.842 for Model 0 and 416.400 for Model 1. Differences in BIC values can be interpreted as follows: 0–2: anecdotal, mention and that is it; 2–6: positive; 6–10: strong; and above 10: very strong (Kass & Wasserman, 1995). In other words, the difference of almost four points constitutes positive evidence in favour of 'no treatment effect'. Analogous to the relative likelihood, a ratio for differences between BIC values can be computed, and this ratio can be interpreted as an approximate Bayes factor (BF) (Kass & Raftery, 1995):

$$\text{approximate BF} = \exp[(\text{BIC}_{\text{min}} - \text{BIC}_{\text{alt}})/2].$$

For the example experiment, the outcome is approximately 0.169. More generally, AIC, AICc, and BIC constitute three alternatives to BFs that do not require researchers to formulate prior distributions and each provide different penalties for overfitting (i.e., a tendency towards overly complex models), with AIC providing the weakest penalty and BIC providing the strongest penalty.

Sample-Size Adjusted Bayesian Information Criterion

A variation on BIC which in terms of penalty for overfitting lies somewhere in between AIC and BIC is found in the sample-size adjusted BIC (SABIC; Enders & Tofighi, 2008; Tofighi & Enders, 2007). AIC and BIC are by default provided in many software packages, and *Mplus* also by default provides the SABIC along with AIC and BIC. In our example experiment, we find SABIC = 406.517 for Model 0 and 406.913 for Model 1. In line with the other information criteria discussed in this section, the model with the lowest SABIC value is to be preferred, hence here Model 0.

Statistical Significance and Information Criteria

NHST and information criteria are based on a different logic. Regardless of the null hypothesis tested—point (e.g., 'no difference') or composite (i.e., values in a given range, common in one-sided testing)—the p-value is the probability of the observed value of the test statistic (e.g., t or χ^2) or further away from $\boldsymbol{H_0}$, if $\boldsymbol{H_0}$ is true. As such, very small p-values indicate that, in the light of $\boldsymbol{H_0}$, observing the findings from the sample or further away from the $\boldsymbol{H_0}$ is unlikely. Even though different information criteria depart from somewhat different questions and assumptions, they help researchers to choose between competing models. That said, Forster (2000) indicates that in terms of complex-simple preference two-sided 'no difference' NHST, using $\alpha = 0.05$, appears to be situated somewhere in between AIC and BIC. That is, the p-value may in some cases indicate insufficient evidence to reject a 'no difference' $\boldsymbol{H_0}$ (e.g., $p = 0.07$) while AIC indicates a slight preference for Model 1 ($\boldsymbol{H_1}$), while in some other cases, the p-value may lead researchers to reject a 'no difference' $\boldsymbol{H_0}$ (e.g., $p = 0.04$) while BIC indicates a slight preference for Model 0 ($\boldsymbol{H_0}$).

Bayesian Estimation and Hypothesis Testing

What the approaches discussed thus far have in common is that there is no need to think in terms of *prior distributions* or probability distributions of parameters of interest before seeing the data. Bayesian methods are all about updating prior distributions with regard to our treatment effects or other differences of interest with incoming data into *posterior distributions* or probability distributions about the same parameters of interest after seeing the data (e.g., Wagenmakers et al., 2018). This process of updating is not one that stops after a single study; the probability distribution *posterior* to Experiment 1 can serve as a *prior distribution* for Experiment 2. In other words, Bayesian methods enable researchers to incorporate information from previous studies into their models for next studies. Although not so few people have claimed that Bayesian methods are still 'new' meaning they were introduced only recently, they were introduced in psychological research in the 1960s (Edwards, Lindman, & Savage, 1963).

Credible Intervals as a Bayesian Alternative to Confidence Intervals

From the *posterior* distribution, 95% *posterior intervals* aka *CRIs* aka *highest-density (credible) intervals* can be computed. Contrary to CIs, the width of a CRI depends not only on the data but on the prior distribution as well. After all, Bayesian inference is about updating a prior distribution to a posterior distribution with new data. We saw before that the 95% CI for Cohen's d in the example experiment extends from −0.148 to 0.547. Using the default prior distribution for the difference between two *M*s (*JASP*; Rouder, Speckman, Sun, Morey, & Iversonr, 2012), a Cauchy distribution with *M* (M_d) 0 and scale 0.707 (i.e., 0.5 times the square root of 2), the 95% CRI for Cohen's *d* extends from −0.143 to 0.510. Using a *wide* prior, a Cauchy distribution with *M* (M_d) 0 and scale 1, instead, we find a 95% CRI from −0.155 to 0.540. Using an *ultrawide* prior, a Cauchy distribution with *M* (M_d) 0 and scale 1.4142 (i.e., the square root of 2), instead, we find a 95% CRI from −0.151 to 0.544. Finally, taking the *widest* possible prior in *JASP*, a Cauchy distribution with *M* (M_d) 0 and scale 2, we find a 95% CRI from −0.143 to 0.535. The wider the prior, the less information we incorporate in that prior and the more the 95% CRI resembles the 95% CI. Moreover, apart from the prior, the 95% CRI becomes more similar to the 95% CI with more data coming in. Although when using realistic prior distributions, 95% CIs and 95% CRIs may rarely be identical, for *M*s and M_ds in larger-sample experiments they can be expected to be very similar.

The 95% CRI can also be used to argue for the presence or absence of a meaningful effect. Somewhat similar to the aforementioned equivalence testing approach based on Frequentist statistics, Bayesians have what is called the *region of practical equivalence* (ROPE; Kruschke, 2014; Kruschke & Liddell, 2017). The idea behind ROPE is that if the 95% CRI does not exceed either of the boundaries of ROPE—for example $d = -0.3$ and $d = 0.3$ being the boundaries—we have sufficient evidence to declare relative or practical equivalence. Simultaneously, if a 95% CRI includes none of the values from the ROPE, we can safely reject (hypotheses in) that region, and CIs—90, 95% or other depending on what statistical significance level we consider appropriate in a given context—can be used in the same way. Although when using a Cauchy prior, the 95% CRI is—in comparison to the 95% CI—somewhat shrunk towards '0', the 95% CRI is usually still somewhat wider than the 90% CI even if we use default priors. In the example experiment, the 90% CI for *d* extends from −0.092 to 0.491, which is somewhat narrower than the 95% CRI using a default prior (−0.143 to 0.510). Hence, one could argue that the Bayesian ROPE procedure is slightly more conservative than the Frequentist equivalence testing procedure. For a closer read on Bayesian estimation and power analysis for the ROPE procedure, see Kruschke (2013, 2018).

Although ROPE is not based on *p*-values, and the 95% CRI is usually a bit wider than the 90% CI, ROPE and TOST could provide two approaches to *four one-sided testing* (FOST). We reject the range of values under ROPE whenever the 95% CRI has no overlap with ROPE, and only when the 95% CRI completely falls within ROPE, we declare sufficient evidence for relative equivalence. With TOST, a similar logic can be applied, but now in terms of four one-sided statistical significance tests. For relative equivalence:

$$H_{0.1}\colon d < -0.3 \text{ (i.e., more negative treatment effect);}$$
$$H_{0.2}\colon d > 0.3 \text{ (i.e., more positive treatment effect).}$$

Next, to *reject* relative equivalence:

$H_{0.3}$: $d \geq -0.3$ (i.e., either a less negative or a positive treatment effect);

$H_{0.4}$: $d \leq 0.3$ (i.e., either a less positive or a negative treatment effect).

We can use the same '1–2α' (i.e., usually 90%) interval for the set of $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$ as for the set of $\boldsymbol{H_{0.3}}$ and $\boldsymbol{H_{0.4}}$. FOST has three possible outcomes: (a) sufficient evidence for *relative equivalence* (when both $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$ can be rejected); (b) sufficient evidence to *reject relative equivalence* (when either $\boldsymbol{H_{0.3}}$ or $\boldsymbol{H_{0.4}}$ can be rejected); and (c) *inconclusive* (when at least one of $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$ cannot be rejected and none of $\boldsymbol{H_{0.3}}$ and $\boldsymbol{H_{0.4}}$ can be rejected). Note that 'sufficient evidence' *should not* be interpreted as 'absolute' evidence; we may still be wrong.

Figure 2.3 graphically presents the rationale behind FOST and the possible scenarios

A, B1, B2, C1, C2, and C3 represent six 90% CIs. In the case of A, we reject both $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$ and therefore conclude sufficient evidence for relative equivalence. Note that the rejection of both $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$ implies that none of $\boldsymbol{H_{0.3}}$ and $\boldsymbol{H_{0.4}}$ can be rejected. Again, this is not absolute evidence. Moreover, relative equivalence does not necessarily imply that 'no difference' lies in the 90% CI. For example, in a large meta-analysis, involving many studies, we may find a 90% CI for d extending from 0.05 to 0.25; with this interval, we reject both $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$ and therefore conclude sufficient evidence for relative equivalence, but the interval does not include '0'.

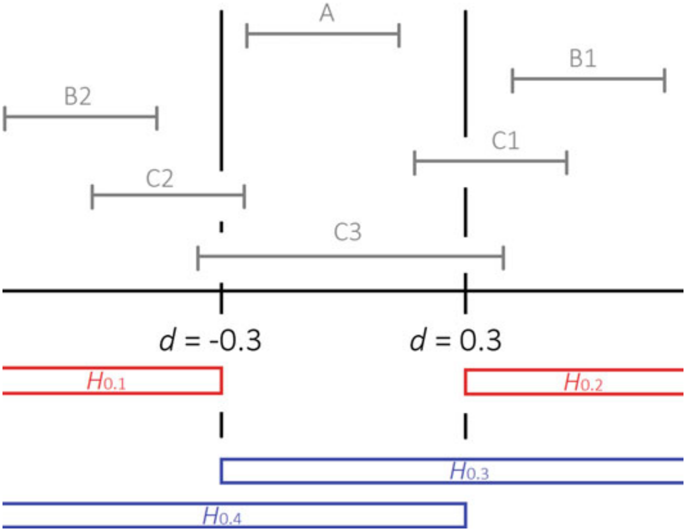


Fig. 2.3 Four one-sided testing (FOST): sufficient evidence for relative equivalence (A), sufficient evidence against relative equivalence (B1, B2), or inconclusive (C1, C2, C3)

In the cases of B1 (reject $\boldsymbol{H_{0.4}}$) and B2 (reject $\boldsymbol{H_{0.3}}$), we have sufficient evidence against relative equivalence. Again, this is not to be interpreted as absolute evidence, and this is unlikely to occur in a single experiment unless *d* in the population is very large or, when *d* is more moderate (e.g., 0.5 in B1 or −0.5 in B2) the sample size is large enough to obtain a 90% CI that does not does include any of the *d*-values in the [−0.3; 0.3] range, such that we can reject either $\boldsymbol{H_{0.3}}$ or $\boldsymbol{H_{0.4}}$. Note that rejecting $\boldsymbol{H_{0.3}}$ implies rejecting $\boldsymbol{H_{0.2}}$ as well but that we fail to reject $\boldsymbol{H_{0.1}}$, and that rejecting $\boldsymbol{H_{0.4}}$ implies rejecting $\boldsymbol{H_{0.1}}$ as well but that we fail to reject $\boldsymbol{H_{0.2}}$.

In single experiments, C1, C2, and C3 are more likely to occur than any of A or B1 or B2. In the case of C1, we fail to reject $\boldsymbol{H_{0.2}}$ and fail to reject both $\boldsymbol{H_{0.3}}$ and $\boldsymbol{H_{0.4}}$. In the case of C2, we fail to reject $\boldsymbol{H_{0.1}}$ and fail to reject both $\boldsymbol{H_{0.3}}$ and $\boldsymbol{H_{0.4}}$. Finally, in the case of C3, we fail to reject any of $\boldsymbol{H_{0.1}}$, $\boldsymbol{H_{0.2}}$, $\boldsymbol{H_{0.3}}$, and $\boldsymbol{H_{0.4}}$.

In sum, FOST works as follows: (a) if we can reject both $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$, we can conclude sufficient evidence in favour of relative equivalence; (b) if we can reject either $\boldsymbol{H_{0.3}}$ or $\boldsymbol{H_{0.4}}$, we can conclude sufficient evidence against relative equivalence; and (c) in all other cases, we remain inconclusive, meaning we have neither sufficient evidence in favour nor sufficient evidence against relative equivalence. With increasing sample size of experiments as well as with replication experiments and meta-analyses, the likelihood of scenario (c) can be expected to decrease and—depending on the magnitude of a treatment effect of interest—the likelihood of either scenario (a) or scenario (b) can be expected to increase. Nevertheless, regardless the size of a sample or meta-analysis, none of these scenarios is to be evaluated in terms of absolute evidence.

A Bayesian Approach to Model Comparison

A Bayesian alternative to *p*-values is found in BFs (e.g., Rouder et al., 2012; Wagenmakers, 2007; Wagenmakers et al., 2018). Succinctly put, the BF quantifies a shift from prior odds $[P(\boldsymbol{H_1})/P(\boldsymbol{H_0})]$ to posterior odds $[P(\boldsymbol{H_1}|\mathrm{O})/P(\boldsymbol{H_0}|\mathrm{O})]$, hence from *before* to *after* seeing data O from Experiment A. However, contrary to the aforementioned concept of LR, the BF itself *depends* on the prior distribution and this remains a main source of critique to BFs (e.g., Gelman & Carlin, 2017; Kruschke, 2011; Liu & Aitken, 2008). This critique is especially important where effects of interest are rather small or 'medium' (e.g., Cohen's $d = 0.5$) at best, which is the case in much of educational and psychological research. Larger-sample experiments on large effects will usually result in BFs that indicate strong or very strong evidence in favour of $\boldsymbol{H_1}$ over $\boldsymbol{H_0}$ whether we use a default prior, a wide prior, or an ultrawide prior. However, in the example experiment, where the observed effect is small, things look a bit different. Under the aforementioned default prior (*JASP*; Rouder et al., 2012), we find a BF for $\boldsymbol{H_1}$ over $\boldsymbol{H_0}$ (i.e., BF_{10}) of 0.337, which corresponds with $\mathrm{BF}_{01} = 2.964$ (i.e., one BF is the inverse of the other). Under the aforementioned wide prior, we find $\mathrm{BF}_{01} = 3.977$, and under the aforementioned ultrawide prior, we find $\mathrm{BF}_{01} = 5.458$. The wider the prior we use, the more the BF approach will tend towards the simple model (i.e., $\boldsymbol{H_0}$ or Model 0) and the more difficult to it is to establish evidence in favour of $\boldsymbol{H_1}$ over $\boldsymbol{H_0}$. BFs in the range of 1 to about 3 indicate 'anecdotal' evidence, BFs in the range of about 3 to 10 indicate 'moderate' evidence, and BFs above 10 indicate stronger evidence in favour of one hypothesis relative to the other hypothesis under comparison (e.g., Jeffreys, 1961; *JASP*). Some prefer a slightly more conservative approach (e.g., Kass & Raftery, 1995): BFs of 1–3 indicate negligible evidence, BFs of 3–20 indicate positive evidence, BFs of 20–150 indicate strong evidence, and BFs over 150 indicate very strong evidence. Either way, in the example experiment, under the default prior, we would speak of 'anecdotal' or 'negligible' evidence in favour of $\boldsymbol{H_0}$, while under a wide or ultrawide prior we would speak of 'moderate' or 'positive' evidence in favour of $\boldsymbol{H_0}$.

This exercise illustrates that although in an experiment of a size like the example experiment or larger the influence of a different prior on the 95% CRI may be rather small, the difference in BF may be considerable. For instance, the difference between a 'default' (Rouder et al., 2012) and an ultrawide prior in BF in the example factor is about a factor 1.84. This underlines the importance of doing a BF robustness check or sensitivity analysis to examine to what extent different priors may yield different BFs and conclusions with it. Software packages like *JASP* provide useful graphs for such an analysis.

Comparison of Estimation Intervals and Testing Criteria

Regardless of which statistics we use for statistical testing and estimation, our statistics will vary from one experiment to the next. Table 2.1 presents Cohen's d, p-value and 95% CI in each of twenty experiments of $N = 128$ ($n = 64$ per condition) each, on the same treatment effect, $d = 0.50$ in the population of interest, using a two-sided test at $\alpha = 0.05$.

The 95% CIs indicate that there is still a margin of error of about 0.35 around the Cohen's d point estimate in an experiment, and the latter varies from 0.211 in Experiment 9 to 0.873 in Experiment 2. Experiments 4, 7, 9, and 18 do not yield a statistically significant difference at $\alpha = 0.05$ (two-sided testing). This is in line with what one would expect given a statistical power of 0.80. Although the idea of a statistical power of 0.80 is not that for every fixed set of twenty studies exactly sixteen should yield a statistically significant outcome, across an infinite number of repetitions of this exercise we would on average expect sixteen out of twenty studies to yield a statistically significant outcome.

Table 2.2 presents Bayesian posterior point and interval estimates of Cohen's along with BF_{10}, based on a default prior (*JASP*; Rouder et al., 2012), for the same series of twenty experiments.

As explained previously, with a Cauchy prior, 95% CRIs tend to be shrunk towards 0 when compared to 95% CIs. For that same reason, the highest density point estimates of Cohen's d in Table 2.2 are slightly closer to 0 than the point estimates presented in Table 2.1. Note that the BF varies wildly, all the way from

Table 2.1 Cohen's d, p-value, and 95% CI (lower and upper bound, LB and UB) in each of twenty experiments of $N = 128$ ($n = 64$ per condition) (*JASP*)

Experiment	Cohen's d	p-value	95% LB	95% UB
1	0.387	0.030	0.037	0.736
2	0.873	<0.001	0.509	1.235
3	0.625	<0.001	0.268	0.978
4	0.341	0.056	−0.008	0.690
5	0.509	0.005	0.156	0.861
6	0.679	<0.001	0.321	1.034
7	0.283	0.112	−0.066	0.631
8	0.585	0.001	0.230	0.938
9	0.211	0.234	−0.137	0.558
10	0.663	<0.001	0.306	1.018
11	0.604	<0.001	0.248	0.957
12	0.502	0.005	0.149	0.853
13	0.635	<0.001	0.279	0.989
14	0.750	<0.001	0.390	1.107
15	0.738	<0.001	0.378	1.095
16	0.448	0.013	0.096	0.798
17	0.399	0.026	0.048	0.748
18	0.340	0.057	−0.010	0.688
19	0.575	0.001	0.220	0.928
20	0.585	0.001	0.230	0.937

0.361 (anecdotal evidence in favour of $\boldsymbol{H_0}$) in Experiment 9 to 6254.592 (very, very strong evidence in favour of $\boldsymbol{H_1}$) in Experiment 2. Note also that the BF is lowest in Experiments 4, 7, 9, and 18, where the Frequentist approach yielded no statistically significant outcome.

Table 2.3 presents AIC and BIC values for each of Model 0 ($\boldsymbol{H_0}$) and Model 1 ($\boldsymbol{H_1}$) in each of the twenty experiments. In sixteen out of twenty experiments (80%)—Experiments 2, 3, 5, 6, 8–17, 19, and 20—all criteria point in the same direction. In the other four experiments (20%), there is some disagreement.

In Experiment 1, the Frequentist approach yields $p = 0.030$, BF_{10} indicates anecdotal evidence in favour of $\boldsymbol{H_1}$, AIC also prefers $\boldsymbol{H_1}$, but BIC indicates a slight preference for $\boldsymbol{H_0}$. In other words, while three criteria indicate some preference towards the more complex model, BIC is not convinced. In Experiment 4, $p = 0.056$, $\mathrm{BF}_{10} = 1.012$, AIC is in favour of $\boldsymbol{H_1}$, but BIC favours $\boldsymbol{H_0}$. In Experiment 7, $p = 0.112$, $\mathrm{BF}_{10} = 0.602$, and BIC is in favour of $\boldsymbol{H_0}$ as well, but AIC slightly favours $\boldsymbol{H_1}$. Finally, in Experiment 18, $p = 0.057$, $\mathrm{BF}_{10} = 0.999$, AIC is in favour of $\boldsymbol{H_1}$, while BIC favours $\boldsymbol{H_0}$.

In Experiment 1, $\mathrm{BF}_{10} = 1.630$. Using the aforementioned relative likelihood formula based on the difference in AIC to quantify the evidence in favour of $\boldsymbol{H_1}$ over $\boldsymbol{H_0}$, we find 4.019. When we use the difference in BIC, conform the

Table 2.2 Bayesian point and interval estimates of Cohen's *d* and BF_{10} in each of the twenty experiments of $N = 128$ ($n = 64$ per condition) (JASP)

Experiment	Cohen's *d*	95% LB	95% UB	BF_{10}
1	0.354	0.014	0.699	1.630
2	0.830	0.470	1.192	6254.592
3	0.583	0.243	0.932	46.121
4	0.314	−0.019	0.659	1.012
5	0.471	0.137	0.823	7.630
6	0.638	0.283	0.995	120.887
7	0.255	−0.070	0.595	0.602
8	0.545	0.198	0.895	23.863
9	0.190	−0.140	0.527	0.361
10	0.619	0.275	0.979	90.809
11	0.563	0.222	0.918	32.381
12	0.467	0.118	0.813	6.895
13	0.593	0.249	0.947	55.120
14	0.702	0.362	1.067	465.806
15	0.699	0.347	1.051	366.264
16	0.414	0.075	0.767	3.330
17	0.366	0.030	0.705	1.857
18	0.311	−0.017	0.649	0.999
19	0.530	0.177	0.885	20.492
20	0.539	0.201	0.893	23.792

approximate BF approach, we find 1.036. In Experiment 2, $BF_{10} = 6254.592$. Calculating the ratio based on AIC, we find 30760.910, and when we calculate the ratio based on BIC, we find 7390.865. In Experiment 4, we find a ratio based on AIC of 2.375 and a ratio based on BIC of 0.571. In Experiment 7, the ratios are 1.336 for AIC and 0.321 for BIC. In Experiment 9, the ratios are 0.757 for AIC and 0.222 for BIC. Finally, in Experiment 18, the ratios are 2.342 for AIC and 0.563 for BIC.

JASP allows researchers to calculate the BF_{10} and ratios based on AIC and BIC reported here in the 'linear regression' menu, by specifying the 'default' prior (Jeffrey-Zellner-Siow, JZS; Rouder et al., 2012), the 'AIC' prior, and the 'BIC' prior, respectively. Although AIC and BIC do not require researchers to think about prior distributions per se, one way to view AIC (and AICc) and BIC is as a form of Bayesian analysis using different priors (Burnham & Anderson, 2002, 2004). Going back to the example experiment we started with, BF_{10} using the default (JZS) prior is 0.337. Under AIC and BIC prior (JASP), we find 0.702 and 0.169, respectively.

In sum, experiments like the ones of the size discussed in this chapter, when observed effects are large (e.g., Experiment 2), all criteria (*p*, BF, AIC, BIC) will probably indicate a preference for a model that includes that effect (here: Model 1, $\boldsymbol{H_1}$) over a model that excludes that effect (here: Model 0, $\boldsymbol{H_0}$). Likewise, when

Table 2.3 AIC and BIC values for the two models in each of the twenty experiments (*Mplus*)

Experiment	AIC H_0	AIC H_1	BIC H_0	BIC H_1
1	958.537	955.755	964.241	964.311
2	967.753	947.085	973.457	955.641
3	942.091	931.998	947.795	940.554
4	960.635	958.905	966.340	967.461
5	948.451	942.279	954.155	950.835
6	944.576	932.394	950.280	940.950
7	950.740	950.160	956.444	958.716
8	964.887	956.227	970.591	964.783
9	950.325	950.882	956.029	959.438
10	966.457	954.895	972.161	963.451
11	947.260	937.936	952.964	946.492
12	996.024	990.074	1001.728	998.630
13	987.805	971.325	993.509	985.881
14	958.099	943.005	963.803	951.561
15	959.443	944.867	965.147	953.423
16	955.921	951.567	961.625	960.123
17	955.650	952.580	961.354	961.136
18	962.889	961.187	968.593	969.743
19	986.257	977.929	991.961	986.485
20	965.115	956.462	970.819	965.018

observed effects are small (e.g., Experiment 9), all criteria will probably prefer a model that excludes that effect (here: Model 0, $\boldsymbol{H_0}$) over a model that includes that effect (here: Model 1, $\boldsymbol{H_1}$). In all these cases, it will probably not be difficult to explain one's preference of one model over another, except perhaps when for some reason or another (e.g., multiple testing) a more stringent testing is warranted (e.g., $\alpha = 0.01$, only BFs or differences in AIC or BIC of a particular magnitude or larger). In cases where there is disagreement between criteria, researchers will have to motivate choices of model preference but in such cases the evidence in favour of one model over another is not so strong anyway. For instance, in Experiment 1, researchers who base their decisions on the p-value (0.030) or the difference in AIC may be confident of having evidence for a treatment effect, while BF$_{10}$ indicates that this evidence may be weak at best, and the difference in BIC indicates that a model that assumes no treatment effect may do.

Some have argued that we should strive for more stringent criteria for statistical testing. For example, a very recent proposal is to change the 'default' 0.05 statistical significance level to 0.005 (Benjamin et al., 2018, p. 6): "*This simple step would immediately improve the reproducibility of scientific research in many fields. Results that would currently be called significant but do not meet the new threshold should instead be called suggestive.*" As a justification for their proposal, Benjamin et al. (2018, p. 7) argue that a two-sided p-value of 0.005 grossly "*corresponds to Bayes factors between approximately* 14 *and* 26 *in favour of* $\boldsymbol{H_1}$." However, since the BF depends on the prior distribution while the *p*-value does not, it is difficult to establish such relations. Take for instance Experiment 12: we find $p = 0.005$ and $BF_{10} = 6.895$. Although a BF of this size is well within the 3–10 range that is indicative of 'moderate' evidence, it is not yet indicate of 'strong' evidence (what BFs of 14 and 26 indicate). Besides, to achieve a statistical power of 0.80 for $d = 0.50$, we would need $N = 218$ participants or $n = 109$ per condition for a two-sided test at $\alpha = 0.005$, and we would need $N = 192$ or $n = 96$ per condition for a one-sided test at $\alpha = 0.005$.

Another approach is found in justifying your statistical significance level (Lakens et al., 2018). In this approach, researchers transparently report and justify all choices made when they design a study, and this includes the statistical significance level. In response to Benjamin et al. (2018), the proponents of this justification approach argue that there is insufficient evidence that the current standard of $\alpha = 0.05$ is a leading cause of non-reproducibility indeed, that a call for a new standard of $\alpha = 0.005$ will not necessarily result in widespread implementation of that new standard, and that such a lower threshold can come with negative consequences that are not discussed by Benjamin et al. (2018). Potential negative consequences identified by Lakens et al. (2018) are fewer replication studies (we will already need more of our potentially scarce resources for initial studies), a reduced generalisability and breadth (there is some trade-off between sample size and number of experiments we may be able to run), and a renewed exaggeration of the focus on single *p*-values (after all, statisticians and other well-trained scientists are aware that a reliance on mere *p*-values is not a good thing, but many researchers unfortunately use *p*-values as their only criterion). Whether we use $\alpha = 0.05$ or $\alpha = 0.005$, whether we rely on AIC, BIC, BF_{10} or other criteria, strong evidence for the presence or absence of a treatment effect is rarely if ever established in a single experiment; replication research, systematic review, and meta-analysis are key concepts for experimental researchers.

A Different Way of Learning from Data

In many settings, educational and psychological researchers do not have the financial or logistic resources to run experiments with numbers that would guarantee a statistical power of 0.80 for tests at $\alpha = 0.005$. If we took the numbers from the example study ($N = 128$, $n = 64$), the statistical power to detect $d = 0.50$ would decrease from 0.80 to 0.49. This is a considerable loss of statistical power. That said, especially in times of online data collection, there are cases where experiments with samples in the 100s are possible. In such cases, a potentially powerful approach to statistical testing and estimation is found in *cross-validation* (Geisser, 1975; Kurtz, 1948; Mosier, 1951; Osborne, 2010b; Stone, 1974; Yu, 2010). This approach is also part of *machine learning*, the basis of *artificial intelligence* or systems' abilities to learn from data in order to carry out a particular task (e.g., Samuel, 1959; Tiffin & Paton, 2018). Depending on whether that learning takes place with full information about an outcome variable (e.g., linear and logistic regression), no information about an outcome variable (e.g., principal component analysis; e.g., Field, 2018) or partial information about that outcome variable, that learning is called *supervised*, *unsupervised* or *semi-supervised* learning.

In large-sample experiments (i.e., several 100s of participants per condition or more), cross-validation can help to reduce the risk of *overfitting* (Osborne, 2010b): choosing for more complex models than needed. In its simplest form, it works as follows. First, we randomly divide our sample into a *training* sample (65–80% of *N*, with larger *N* allowing for a percentage closer to 80) and a *testing* (aka *evaluation*) sample (the remaining 20–35% of *N*, with larger *N* allowing for a percentage closer to 20). Next the training set is used to determine which of a set of competing models should be preferred. For this, we can use the criteria discussed thus far in this chapter: R^2 and eventually adjusted R^2 as indicators of the proportion of variance in the outcome explained by a model as well as different testing criteria (some of the following: *p*, AIC, AICc, BIC, SABIC, BF). In a simple two-group experiment, this comes down to a linear regression model with slope (i.e., Model 1: treatment effect), where the slope estimates the magnitude of the treatment effect, or without slope (i.e., Model 0: no treatment effect). When experiments include more than a single factor (e.g., a second factor or at least one covariate), several competing models can be compared. The model that is identified as the best model based on the training sample is then saved and tested on the data from the *testing* sample. This does not only work for M_ds but for other types of models as well (e.g., exploratory factor analysis on the training data followed by confirmatory factor analysis on the testing data; Mulaik, 1987; Yu, 2010; or building multiple regression models with training data and testing them with testing data; Chernick, 1999; Osborne, 2010b; Yu, 2010).

Note that the approach outlined in this section is recommended for (very) large-sample experiments only. In cases such as the example experiment and series of replication experiments in this chapter ($N = 128$), which provide a much more common—and for many settings still somewhat ‘ideal’—scenario with regard to sample size (i.e., $N < 100$ is still common for many two-group experiments), splitting the dataset into two parts would bring us back to wildly varying estimation and testing outcomes and is therefore not recommended. However, where samples are large, one can think of many practical examples.

A Pragmatic Approach to Statistical Testing and Estimation (PASTE)

Based on the concepts, criteria, and issues discussed in this chapter, we now conclude with a pragmatic approach to statistical testing and estimation (PASTE) that is used for statistical testing and estimation in the remainder of this book. The core of PASTE is to use different criteria instead of a single criterion for statistical testing and estimation in a coherent, consistent way (Leppink, 2018a, b; some of the criteria outlined in this book were first formulated in the two articles referred to, though by far not in the same level of detail as in this book, some criteria have been added, and some recommendations on criteria introduced earlier have been slightly modified in this book).

Visual and Numerical Checks of Assumptions

Every statistic is based on assumptions. To acquire a solid understanding of how reasonable some core assumptions are, we must graphically inspect our data before we calculate any numbers. For example, assumptions about distributions of residuals in a population of interest and the severity of departures from these assumptions are much more easily checked through histograms than through statistical significance tests. Whether it concerns normality or equal *SD*s (equal variances), relying on statistical significance tests is problematic because the outcome of such tests does not yet tell us what kind of deviations we are dealing with and how severe these deviations are, and interpreting a statistically non-significant *p*-value as evidence in favour of a null hypothesis—whether that null hypothesis concerns normality, equal *SD*s, absence of treatment effect, or something else—is a logical fallacy. When in doubt about one assumption or another, a safe way may be to report findings under different assumptions (e.g., assuming normality vs. not assuming normality, or assuming *SD*s to be equal or not) if doing so results in (meaningfully) different outcomes at all (e.g., in the example experiment in this chapter, the 95% CIs under equal *SD*s versus under unequal *SD*s are almost identical).

Model Comparison and Four One-Sided Testing (FOST)

Omitting the previous step to proceed with statistical testing and model comparison straight away is a potential recipe for disaster. Whether assumptions are met or not, statistical software programmes will provide you with R^2-values, adjusted R^2-values, *p*-values, AIC, BIC, BFs, and statistics alike. However, every statistic provides meaningful outcomes under certain assumptions and, although some assumptions are more critical than others, substantial deviations from assumptions may well invalidate numbers based on those assumptions.

Once we have checked the necessary assumptions and proceed with statistical testing and model comparison, it is recommended to not base our conclusions on a single criterion. Whether we prefer *p*-values, BFs, or other, every criterion has its pros and cons. AIC may sometimes hint at a more complex model, where other criteria provide good reasons not to prefer that more complex model. Likewise, in some cases, BIC may prefer a simple model where researchers may have solid reasons to rely more on AIC and *p*-values (the latter perhaps even from one-sided tests). BFs provide an interesting alternative to *p*-values, AIC, and BIC, but are quite sensitive to prior distribution choices. Given a reasonable sample size, as in the example experiment and replication series in this chapter, different criteria will likely align when effects are large (i.e., a preference towards a model including that effect) or small (i.e., a preference towards a model not including that effect), and more explanation with regard to choices made is needed where different criteria indicate different preferences.

Both p-values and BFs can be easily used when researchers have clear reasons to expect a difference in one direction prior to an experiment. In the example experiment, $BF_{10} = 0.337$ comes from the default two-sided tests. If the alternative was one-sided in favour of the treatment condition (i.e., a positive treatment effect), a one-sided test in that direction would yield $BF_{10} = 0.579$, and if the alternative was one-sided against the treatment condition (i.e., a negative treatment), a one-sided test in that direction would yield $BF_{10} = 0.095$.

In cases where we do not have reasons to engage in one-sided testing, p-values and BFs can be used along with AIC, BIC, and/or AICc and SABIC. Given that in situations where there is disagreement between criteria, AICc, SABIC, and two-sided BF and p-value are usually situated somewhere in between AIC (strongest tendency towards more complex) and BIC (strongest tendency towards simpler), in the case of two-sided testing—which remains the default in most of educational and psychological research—we may as well report just AIC and BIC, and eventually some of the other criteria along.

What is good to provide along with statistical testing criteria regardless of whether we test one-sided or two-sided is indicators of the proportion of explained variance. Doing so helps researchers to appreciate if a difference in (adjusted) R^2 between a model with or without treatment effect, or (in later chapters) between a model with or without a particular covariate, is substantial enough to justify a choice for a more complex model or is perhaps rather small and from a practical perspective not really interesting. For example, in Experiment 1 of the replication series, the treatment explains 2.9% (adjusted R^2) to 3.7% (R^2) of the variance in post-test performance. Some may argue that although this is not a lot, the treatment explains a proportion of the post-test variance that is substantial enough to matter in a practical setting and use that argument along with AIC and p-value to justify their preference for Model 1 (treatment effect) over Model 0. Others may argue that 3.7% of the variance is, in the context at hand, not substantial enough and point at the preference towards Model 0 indicated by BIC and/or that $BF_{10} = 1.630$ indicates weak evidence for a treatment effect at best. By providing a range of criteria—(adjusted) R^2, AIC, BIC, and eventually other—along with the necessary descriptive statistics, readers can have a meaningful discourse about findings reported in an article and can draw their own conclusions.

Whenever the interest lies in establishing evidence in favour of (relative) equivalence, p-values from 'no difference' null hypothesis significance tests should not be used. In such cases, TOST equivalence testing and the Bayesian ROPE approach constitute two powerful approaches that—certainly in single experiments or small series of experiments—may well indicate that although p-values from 'no difference' null hypothesis significance tests are not statistically significant we do not have sufficient evidence to declare relative equivalence either. Although ROPE is not based on statistical significance, ROPE and TOST can be used in a similar fashion in terms of fully, partially or not at all overlapping with an a priori declared region of relative or practical equivalence. ROPE is by default based on a 95% CRI, while TOST is by default based on a '1–2α' interval, which in the case of $\alpha = 0.05$ means a 90% CI but can be increased whenever a lower α is considered more appropriate (e.g., multiple testing situations). FOST is a logical extension of TOST which, like the ROPE procedure, has three possible outcomes: sufficient evidence in favour of relative equivalence, sufficient evidence against relative equivalence, or inconclusive. TOST, ROPE, and FOST are not meant to stimulate a renewed black-or-white cut-off-style thinking like '$p = 0.051$' being terrifying and '$p = 0.049$' being worth a bottle of champagne or like $BF_{10} = 2$ 'proving $\boldsymbol{H_1}$' and $BF_{01} = 2$ 'proving $\boldsymbol{H_0}$'. A 90% interval from $d = -0.301$ to $d = 0.304$ is not much different from an interval from $d = -0.294$ to $d = 0.289$, and the same goes for a CRI. A good recent example of a combined use of ROPE and BFs indicating evidence for relative or practical equivalence comes from Etherton et al. (2018), who performed a Bayesian analysis of a series of multimethod ego-depletion studies that all in all involved data from $N = 840$ participants. Main finding from their meta-analysis: a point estimate of Hedges' g (i.e., similar to Cohen's d, see: Hedges, 1981) with a 95% CRI of [−0.05; 0.24] and BFs, under each of a variety of prior distributions, above 25 in favour of $\boldsymbol{H_0}$ (i.e., $BF_{01} > 25$).

Point and Interval Estimation

TOST and ROPE and the TOST-ROPE uniting FOST model remind us of the need to compute not just point estimates but interval estimates (CIs in the case of TOST, CRIs in the case of ROPE) as well. We are rarely really only interested in whether or not 'there is an effect' of something on something; in virtually all educational and psychological contexts, we would rather like to know what kind of *magnitude* of effects we are dealing with. Where we agree that testing at $\alpha = 0.05$ is defendable, we should probably report *both* the two-sided 95% *and* the two-sided 90% CI of treatment effects of interest. The 95% CI includes all possible null hypotheses that would not be rejected at $\alpha = 0.05$ in the case of two-sided testing. The 90% CI serves two purposes. Firstly, it indicates the lower bound in the case of a one-sided test if a negative treatment effect is expected, and it indicates what is the upper bound if a one-sided test for a positive treatment effect is considered. Secondly, the 90% CI can in its entirety be used for TOST equivalence testing. For researchers who prefer a Bayesian approach, the 95% CRI constitutes a Bayesian alternative to the 95% CI that can also be used for ROPE. That said, perhaps Bayesians should consider reporting both CIs and CRIs. Although in the case of minimal information in the prior distribution (e.g., an ultrawide prior for the M_d in a two-group experiment) and sufficiently large N (e.g., the experiment taken as an example in this chapter) the 95% CI and the 95% CRI can be expected to yield very similar results, the difference between these two intervals provides an objective indicator of the degree of subjective influence of the specified prior distribution.

A Final Note on the Units

Note that the CIs and CRIs in this chapter have (largely) revolved around Cohen's *d*, not around the number of points of difference in post-test score. Which metric to use partly depends on the research question and partly depends on what facilitates comparison of effects of interest across experiments. TOST and ROPE can be used with Cohen's *d* as well as with other units, including the number of points of difference in post-test score. However, the latter is heavily scale-dependent and may be difficult to compare across experiments unless all experiments use exactly the same instrument. Intervals based on scales (units) such as Cohen's *d* allow researchers to speak of practically meaningful versus perhaps not so meaningful treatment effects and other differences independent of the scales used in a random experiment. That said, which scales or units to use also depends on the nature of the data. For M_ds, Cohen's *d* in many cases provides a sound standardised scale, at least as long as the data structure is not such that multilevel analysis is needed. Besides, the type of outcome variable is also important to consider; while Cohen's *d* may often constitute a useful metric in the context of M_ds, it is generally inappropriate for categorical outcome variables. Chaps. 5, 6, 7, 8 provide appropriate measures of effect size for different types of outcome variables.

3 Measurement and Quality Criteria

Abstract

Chapters 1 and 2 provide the first two parts of a general framework for statistical analysis in experimental research. In a nutshell, statistical analysis ought to be question/hypothesis-driven, should account for the core features of the experimental design and data acquired with that design, and involves combinations of statistical testing and estimation criteria rather than a single criterion. For the sake of simplicity, the examples discussed in the first chapters do not touch issues of measurement and the reliability and validity of that measurement. However, the vast majority of outcome variables as well as not so few predictors of outcome variables come from psychometric measurements such as self-reported ratings, assessment by experienced or not so experienced raters or performance measured otherwise (e.g., through multiple-choice tests). In this chapter, different approaches to measurement are discussed and compared in terms of their relative pros and cons. Although many of these pros and cons are not limited to experimental research, all issues discussed in this chapter are important to consider in experimental research. This includes a revision of some common practices that may distort our perspectives on the reliability of our instruments, such as the default use of Cronbach's alpha regardless of the nature of the instrument and data acquired with that instrument.

Introduction

Psychometric measurement has been a topic of discussion at least since Thorndike (1904), and over the past century many great scholars have contributed to theories about psychometric measurement, including Birnbaum (1968), Crocker and Algina (1986), Cronbach (1951, 1975, 1976), Cronbach, Gleser, Nanda, and Rajaratnam (1972), Hambleton (1978, 1980, 1983), Lord and Novick (1968), Rasch (1960),

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research*
in Education and Psychology, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_3

Stevens (1946), Thurstone and Chave (1929), Torgerson (1958), Weitzenhoffer (1951), and Yerkes (1921), to name a few. These days, there are many different approaches to measurement, some of which are known by a wider audience than others and some of which are more commonly used in (published) research than others. This chapter is by no means an attempt to cover all measurement approaches and methods that can be found in the literature. The goal of this chapter is to provide a concise overview of methods to estimate the reliability of our measurement instruments and/or to study some statistical aspects of the validity of our measurement instruments. This chapter focusses on methods that are commonly encountered in experimental educational and psychological research or that perhaps remain underused but may be useful for experimental research in one way or another. First, some measures of consensus are presented. Next, several measures of consistency are discussed. After this overview of consensus and consistency measures, several latent variable methods are compared. Finally, some alternative approaches are discussed that at first resemble latent variable methods presented in this chapter but are different not in the last place because they are not based on latent variables.

Two Examples of Measurement Practice Leading Nowhere

Progress in science is to a large extent about *learning from error*, including through discourse on unconstructive practices. Although critique on such practices is sometimes booed away as bullying or otherwise undesirable behaviour, critique on methods is not the same as criticising the scholars using the methods to be criticised. Such a constructive dialogue *is* possible and ought to be *stimulated*, because this dialogue can be held without discrediting *any* of the hard labour done by scholars in a particular field and without disrespecting the scholars themselves.

Example 1: Cognitive Load

In the context of *learning from error*, I would like to share an example of a measurement practice I believed in for years and that I contributed to with several widely cited publications (e.g., Leppink, Paas, Van der Vleuten, Van Gog, & Van Merriënboer, 2013; Leppink, Paas, Van Gog, Van der Vleuten, & Van Merriënboer, 2014) but which I no longer view the same way because after some years of reflection I have reached the conclusion that it is heavily flawed: *cognitive load measurement*. I have been going back and forth between defending and abandoning the practice for a few years for a number of reasons. Firstly, once you are in the middle of a particular practice, you are likely to become at least somewhat biased towards that practice and it becomes hard to let it go. I delivered two Keynote lectures at international conferences largely based on cognitive load theory and measurement: at the International Cognitive Load Theory Conference (June 2016, Bochum, Germany) and at the International Association of Medical Science Educators (June 2017, Burlington, Vermont, United States). During my first talk, I focussed on weak spots in the measurement practice in front of the cognitive load theory 'community' (a bit over 100 scholars from across the world visiting the conference). Weak spots that I had identified and also informally shared at two previous editions of the same conference, but which I had been afraid to share at earlier occasions because I considered myself a 'not yet cited nobody' whose opinion would not matter anyway. In fact, this was not just something I believed; I was told so by some researchers (not from the cognitive load theory community itself, though). However, in June of 2016 some of my work on cognitive load theory and measurement (including Leppink et al., 2013, 2014) had already been cited quite a few times and I thought there was a reason why I was one of the Keynote speakers. I firmly believed in the work I had started some years earlier (resulting in Leppink et al., 2013 after two years of hard labour), and I was confident that I could help the measurement practice forward from there. However, in the year that followed, I lost much of that belief and confidence, and for that reason, my second Keynote lecture turned out to be one of my most difficult presentations in my career thus far. I had my story, but only part of the belief and confidence with which I opened my first Keynote, a year earlier. I still think that *cognitive load theory* (e.g., Sweller, Ayres, & Kalyuga, 2011; Sweller, Van Merriënboer, & Paas, 1998) continues to provide one of many potentially useful frameworks for thinking about the design of education (e.g., Lee, Hanham, & Leppink, 2019), but in my view it will have to start evolving to explain some findings it cannot explain (e.g., Kalyuga & Singh, 2016; Kapur, 2008, 2011, 2014; Kapur & Rummel, 2012; see also the later chapters in Lee et al., 2019) for otherwise at some point it will start to lose its value.

Cognitive load is the core construct of cognitive load theory. Basically, cognitive load theory assumes that information has to be processed within narrow limits of our working memory, and that we have to minimise cognitive load that may hinder learning in order to keep as many working memory resources as possible available for dealing with cognitive load that may stimulate learning. Although there is agreement in the cognitive load theory community that not all cognitive load is 'bad' in the sense that all cognitive load hinders learning, there has been long-standing disagreement on how many types of cognitive load we need and how the types of cognitive load we need ought to be defined (e.g., Leppink & Van den Heuvel, 2015; Sewell et al., 2018). This, together with measurement practices outlined in the following, poses a serious threat to the continued usefulness of cognitive load theory.

To start, cognitive load measurement has largely been based on *self-reports*, above all on a *single mental effort* categorical rating item (Paas, 1992; Sweller, 2018). Core assumptions underlying that practice include respondents' ability to *assess* and *report* on their cognitive load. Knowing that a single item *cannot distinguish* between multiple types of cognitive load, some researchers *attempt to keep constant* all cognitive load types by design except Type X and then interpret differences in self-rated mental effort as differences in X. *Inability* and *bias* of the respondent may undermine the *assessment* and *reporting* assumption, and keeping any type of cognitive load constant may work with *robots* but likely not with *human* learners. Even if we understand keeping something 'constant' as creating randomised groups that on average are the same on a particular type of cognitive load, correlations will still be influenced by scores *not* being constant instead of being constant, and given the sample sizes that are common in cognitive load research (i.e., usually between 10 and 25 participants per condition) considerable differences between groups in that type of cognitive load that is supposed to be 'constant' are well possible.

Studies involving more objective time-based or behaviour-based measures suffer from the same problems and frequently use the mental effort as *gold standard* for *validation*, because the mental effort item is widely believed to be *reliable*. However, estimating the reliability of measurement requires either *multiple items* at a given time or *repeated administration* of an item under the *same circumstances*. For variables such as mental effort, this is impossible, because *measurement error*, *task differences*, and *response shift* due to learning and/or tiredness constitute three *perfectly confounded* sources of variance.

Although multi-item cognitive load measurement instruments may help to distinguish between cognitive load types *in theory*, they rely on the assumption that respondents are able to *differentiate* between and *report* on the different types of cognitive load captured. These assumptions remain *untested* and are heavily *context-dependent*. Moreover, all multi-item instruments available, including my own and variants thereof (e.g., Leppink et al., 2013, 2014), suffer from *question wording effects* that may heavily influence the psychometric structure of the instrument and may invalidate the outcomes. Finally, if researchers cannot agree on the number and definitions of cognitive load types, how can we expect respondents to differentiate between types and report on any specific type of cognitive load?

Theoretical and measurement problems in cognitive load theory reinforce each other and leave cognitive load measures useless in our endeavour to acquire a deeper understanding of *learning processes* and *learning outcomes*. Cognitive load measures are based on untested and perhaps untestable assumptions, and the best they offer is unscientific *after the fact* explanations of processes and outcomes. If outcomes are good, the load must have been 'good'; otherwise, the load must have been 'bad'. Given further that cognitive load measures are *not needed* for explaining learning processes or outcomes, we may as well stop measuring cognitive load altogether.

Example 2: Learning Styles

Another flawed practice, which has been around for much longer than cognitive load theory but—perhaps not in the last place because some have made a career out of it—seems to not go away, is that of measuring *learning styles* in order to tailor education to preferred learning styles (e.g., Veenman, Prins, & Verheij, 2003). This practice is perpetuated despite a complete lack of theoretical (e.g., Kirschner & Van Merriënboer, 2013) and empirical (e.g., Pashler, McDaniel, Rohrer, & Bjork, 2008) support for that practice. In a recent letter to the editor, I summarised the main three reasons to stop this practice (Leppink, 2017b): (1) heavy reliance on questionable self-report measures, (2) leaves all doors open for *after the fact* explanations and has as the only implication for the design of education that *everyone is different*, and (3) it flies in the face of decades of research the findings of which flatly contradict what learning styles theory would predict.

As with cognitive load measures and any other self-report practice, the use of self-report learning styles questionnaires relies on the assumption that learners *are aware* of their learning styles and can reliably and validly report on them. As we have seen, this is a problem with cognitive load measures and is no less of a problem for learning style measures. Empirical support for the assertion that self-report learning style instruments yield any kind of valid and reliably measures of use for educational practice is lacking (Pashler et al., 2008; Veenman et al., 2003). Next, over 70 different learning styles have been identified in the literature and learners can allegedly have all kinds of combinations of learning styles. This quickly results in more combinations of learning styles than there are people on the planet, and if we are to tailor education in that fashion, we would have a different kind of education for every single individual. That would be neither practical nor in line with bodies of literature on research on the design of education (e.g., Kalyuga & Singh, 2016; Kapur, 2008, 2011, 2014; Kapur & Rummel, 2012; Kirschner & Van Merriënboer, 2013).

Theoretical and Measurement Issues

Cognitive load and learning styles have in common that theoretical and measurement issues reinforce each other. Inconsistency in definitions and question wording effects lead to poor measurement, and poor measurement does not allow us to investigate important theoretical statements that are claimed to be of value for educational practice. This leaves us with an *after the fact* explanation of empirical phenomena which, if not changed, may at some point result in a consensus in broader communities that cognitive load theory and theories on learning styles have hardly if any more scientific and practical value than Sigmund Freud's psychoanalysis (e.g., Freud, 1920). At least, contrary to theories on learning styles, cognitive load theory has resulted in actually useful and widely recognised approaches to the design of education (e.g., Leppink & Van den Heuvel, 2015; Sweller et al., 1998, 2011; Van Merriënboer & Kirschner, 2018). However, theoretical inconsistencies and bad measurement practices will need to be revised in order to maintain its perceived value as a useful theory for the design of education. Extensive revisions may help to initiate new research on phenomena it appears to consistently fail to explain (e.g., productive failure: Kalyuga & Singh, 2016; Kapur, 2008, 2011, 2014; Kapur & Rummel, 2012; see also the later chapters in Lee et al., 2019). Moreover, research inspired by cognitive load theory constitutes a good example of researchers attempting to integrate many *small sample size* studies into a theoretical story. Experiments with conditions of hardly over 20 participants, which are common in cognitive load research, leave researchers poorly equipped to test and estimate most effects of interest (i.e., low statistical power and precision) and disable researchers to thoroughly examine the psychometric properties of measurement instruments used. Unless the cognitive load theory community incorporate these revisions, my prediction is that sometime in the next decade we will see *both learning styles and cognitive load* placed in the museum of *history of art and science in education* and will no longer be considered useful or needed for educational research and practice. That said, it appears that there is increasing awareness among cognitive load researchers that theory and measurement practice need to evolve; in a recent special issue on cognitive load theory in the journal called Educational Psychology Review (Issue 2 of 2019), several ways forward are suggested. Good examples of ways forward include an expansion of the theory to account for relations between cognitive load and movement (e.g., tracing, gesturing, eye movement, and body movement), the consideration of motivational change as a response to changes in cognitive load rather than only as a precursor of cognitive load, and accounting for emotion and stress. However, issues related to specific measurement practices (e.g., the continued dominance of single-item mental effort ratings) and frameworks that generate findings and explanations that cannot be accounted for by cognitive load theory (e.g., productive failure) remain unaddressed in this special issue.

The two examples shared in this section illustrate the importance of good theory and good instruments; without these, any attempt to study the reliability or statistical aspects of the validity of an instrument may be nothing more than a useless exercise. If we cannot even agree in a community what we are measuring, there is no way we can expect reliable and valid measurements of participants in our experiments. If a measurement practice already flies in the face of the theory it is supposed to be based on—such as using a single item for complex constructs such as (types of) cognitive load—we may as well stop doing research in that area altogether. Good theory and sound instruments are necessary conditions for a meaningful study of the psychometric properties of our instruments.

Finally, the problems related to reliability estimation of single items are not limited to cognitive load theory; this is a much wider problem in educational and psychological research. To mention another rather influential line of research, especially in Educational Psychology, the one on *judgement of learning* (e.g., Kornell & Metcalfe, 2006; Metcalfe & Kornell, 2005; Thiede, Anderson, & Therriault, 2003; Thiede & Dunlosky, 1999), suffers from exactly the same problem. Regardless of how the item is formulated (there is considerable variety in that), much of the research uses a single item; perfect confounding of *measurement error*, *differences in tasks in which it is used*, and a likely *response shift* due to participants learning, changes in their perspectives of learning, and eventually tiredness, is also a commonly ignored elephant in the room in this line of research. The measures discussed in the remainder of this chapter are built on the assumption that researchers come prepared, with solid theories, appropriate sample sizes, and sound measurement instruments.

Consensus

Which methods to use to estimate the reliability or statistical aspects of the validity of our measurement instruments depends on a number of questions. Firstly, we need to consider the level of measurement of our instrument or, in most cases, of the items that constitute the instrument: nominal, ordinal, interval, or ratio (Stevens, 1946). Secondly, we need to carefully inspect the distributions of the different items. Thirdly, a core assumption in measurement is that a set of items or a team of raters can be considered raters of the same construct of interest, and that the categories of items or ratings have the same meaning across items or ratings.

If we are dealing with two categorical items or raters, several measures of consensus may be useful to a more or lesser extent. A first, straightforward measure of consensus is found in the *percentage of agreement*. This provides an easy understandable, intuitive metric. However, it does not yet correct for a degree of consensus that could be expected just by random response (guessing), and this is especially a problem when there are only few categories (e.g., two categories) and/or when one or some categories occur very frequently or very rarely (e.g., Hayes & Hatch, 1999; Stemler & Tsai, 2010).

A second estimate of consensus can be found in the *odds ratio* (*OR*). This concept is especially easy to use in the case of two-category ratings on two items or by two raters (i.e., 2 × 2 contingency tables) and is useful in many other statistical applications of categorical data analysis (e.g., Agresti, 2002). In the context of interrater agreement, the *OR* is used as follows. Suppose, two professors independently rate the performance of 100 undergraduate medical students in a skills training test as either ‘pass’ or ‘fail’. Professor A lets 90 students pass (odds are 90:10), while Professor B only lets 80 students pass (odds are 80:20). The resulting *OR* is [90/10]/[80/20] = 9/4 = 2.25. If the odds of the two professors were the same, the *OR* would equal 1; the more the *OR* deviates from 1, the larger the discrepancy between professors in their level of consensus.

Larger levels of discrepancy as expressed by the aforementioned *OR* result in lower interrater agreement and lower interrater reliability estimates with it. Two commonly used statistics of interrater reliability that, contrary to the percentage of agreement, correct for agreement due to random response (guessing), are Cohen’s κ (Cohen, 1960, 1968) and Krippendorff’s α (2004). Although *OR* of 1 does not say anything about Cohen’s κ or Krippendorff’s α, the further away the *OR* from 1 the lower the upper bound of possible κ and α values. For instance, in the in practice somewhat unlikely case that Professor A lets 80 students pass while Professor B only lets 50 students pass, the resulting *OR* is 4, the maximum possible percentage of agreement is 70, and the maximum possible Cohen’s κ value is 0.40.

When the items or ratings are not categorical but quantitative, Bland-Altman analysis may constitute a useful approach to examining consensus. In Bland-Altman analysis enables research to estimate bias in measurement (i.e., M_d) as well as dispersion around that bias (e.g., the *SD*) and possible outliers. If the bias is constant across the range of measurement, the bias can be easily accounted for by a simple addition or subtraction. If the bias is not constant across the range of measurement, this indicates that the difference in items or ratings depends on the extent to which participants have the construct of interest. For example, Professor A might give higher exam scores than Professor B for very knowledgeable students, while Professor B might give higher exam scores than Professor A for not so knowledgeable students.

Consistency

In cases where multi-category ordinal or quantitative variables are concerned and researchers are primarily interested in consistency of ratings as in covariation of scores on different items or scores obtained from different raters rather than in absolute agreement per se, several correlation-based approaches can be considered.

Pearson, Spearman, and Kendall

When there are two items or ratings that can be considered of interval or ratio level of measurement and there are no severe departures from normally distributed residuals, Pearson's correlation coefficient *r* (Pearson, 1900) can be used as an estimator of the linear correlation between items or ratings. When there are some dubious departures from normality but still want to use Pearson's *r*, researchers have a number of options. Two easy approaches would be *trimming* the data (i.e., omitting the most extreme scores, for example the 5% or 10% most extreme scores) or *winsorising* the most extreme cases by replacing their values by the nearest score that is not an outlier (e.g., Field, 2018). Another approach is found in the use of *robust* methods (e.g., Wilcox, 2017), which include bootstrapping (Efron, 1979, 1981, 1982, 1983; Efron & Tibshirani, 1993) and other resampling methods (Fisher, 1960; Mosteller & Tukey, 1977; Quenouille, 1949; Rodgers, 1999; Tukey, 1958; Yu, 2010). Transformations (e.g., square root or log of right-skewed distributions; Field, 2018) could constitute another approach but would only work if the same transformation can be applied to both items or ratings.

Researchers who do not want to use any of the aforementioned methods just to use Pearson's *r* can also opt for rescaling the observed values to ranks. For example, the values 10, 7, 4, 3, 2, 1 would then be rescaled to 6, 5, 4, 3, 2, 1. Pearson's *r* can then be calculated based on these rescaled ranks. This is in fact what Spearman's correlation coefficient ρ (Spearman, 1904) does and this is interesting to think about for a bit longer for at least one other reason. Commonly, Spearman's ρ is presented as the default correlation coefficient for ordinal variables. However, by squaring distances based on ranks we assume these ranks to be of interval level of measurement; in other words, we are actually treating ordinal data as if we were dealing with interval data (e.g., Tacq & Nassiri, 2011). Several scholars have used comparisons of Pearson's *r* and Spearman's ρ to justify treating

Likert data (i.e., Likert, 1932) as interval even if they are ordinal (e.g., Norman, 2010). However, this comparison is problematic because both coefficients treat the numbers in the calculations as at least interval (i.e., equal distances).

If we really want to go ordinal, we probably need to use (one of) Kendall's τ (rank) coefficients (Kendall, 1938, 1962). As explained by Tacq and Nassiri (2011), there is an interesting connection between Kendall's approach to ordinal data and the work of a French sociologist that in some fields appears to be used an example of why qualitative is 'more important' than quantitative data: Pierre Bourdieu (e.g., Bourdieu, 1984). Amongst others, Bourdieu distinguishes the dominant class, the middle class, and the working class as three classes based on a relational logic (i.e., power relations, financial means), not in terms of '1', '2', and '3' with equal distances between these labels. Instead of interval distances, τ coefficients are based on the numbers of concordant (i.e., agreeing) and discordant (i.e., disagreeing) pairs and, in some cases, on the number of ties (i.e., pairs that are neither concordant nor discordant) (e.g., Agresti, 2010; Berry, Johnston, Zahran, & Mielke, 2009; Kruskal, 1958).

Correlation Structures

Although the previous paragraph discusses correlation coefficients in a context of two items or two raters, the concept of correlation can be easily generalised to larger numbers of items or raters. This generalised concept is also known as *ICC*. In its simplest form, for any set of two or more items or ratings that are supposed to measure a particular construct of interest, differences between participants in that construct of interest creates a correlation between residuals of different items or raters that is proportional to these differences. When *SD*s of residuals are the same across items or raters and the correlations between residuals are the same across pairs of items or raters as well, we are dealing with a so-called *compound symmetry* (CS) structure (e.g., Field, 2018; Tan, 2010). In the context of mixed-effects aka multilevel analysis, this structure can be accounted for with a *random-intercepts* (RI) model (e.g., Snijders & Bosker, 2012; Tan, 2010; see also Chaps. 14–16 of this book). The *ICC* estimated (using restricted maximum likelihood, REML; e.g., Tan, 2010) from a RI (i.e., CS residual covariance) model, in which the item *M*s may vary as long as the *SD*s are equal across items, is the *ICC* that serves as input for Cronbach's alpha and generalisability theory (Cronbach et al., 1972). Given the number of items *k* and an *ICC*, Cronbach's alpha (α) can be computed via the Spearman–Brown formula (Brown, 1910; Spearman, 1910):

$$\alpha = (k * \mathrm{ICC})/[1 + ((k{-}1) * \mathrm{ICC})].$$

In other words, Cronbach's alpha is a function of the number of items *k* and *ICC* that can be estimated in a two-level (upper level: participant; lower level: item) RI (i.e., at the level of participant) model, using REML for estimation. If $k = 4$ and $ICC = 0.30$, $\alpha = 0.63$. Given *ICC*, to increase α we would need more items.

Cronbach's Alpha and Generalisability Theory

Cronbach's alpha (1951), which is the same as Guttman's lambda-3 (1945) and a generalisation of Kuder-Richardson's KR-20 coefficient for dichotomous items (Kuder & Richardson, 1937), remains the most widely used estimator of 'reliability', 'internal consistency', 'unidimensionality', and 'validity' despite long-standing critique from increasing numbers of scholars. In fact, Lee J. Cronbach himself (1951, 1988) argued that alpha is a poor index of unidimensionality and, in Cronbach and Shavelson (2004, p. 397) literally states: "*It is an embarrassment to me that the formula became conventionally known as Cronbach's* α." The take home message from Cronbach and Shavelson (2004, p. 391) is that "*alpha covers only a small perspective of the range of measurement uses for which reliability information is needed and that it should be viewed within a much larger system of reliability analysis, generalizability theory*." To generalisability theory, we return in a bit. First, let us have a closer look at why Cronbach's alpha is, at least in most educational and psychological research settings, not a good estimator of reliability, internal consistency, unidimensionality, validity or whatever label of scale quality researchers want to give it.

With regard to unidimensionality, we can be short: this is an assumption that *underlies* Cronbach's alpha, not something that can be tested or estimated *by* Cronbach's alpha (e.g., Leppink & Pérez-Fuster, 2017; Peters, 2014). If a set of items cannot be expected to measure the same construct of interest, computing Cronbach's alpha over that set of items is a useless exercise. Next, CS is, at least in most educational and psychological research, an unrealistic assumption (e.g., Dunn, Baguley, & Brunsden, 2014; Peters, 2014). Although small departures from that assumption may not create a lot of distortion, substantial departures can do so, are quite common, and ought to be accounted for by choosing a model that assumes a different residual covariance structure (i.e., unequal *SD*s and/or unequal correlations). For exactly the same reason, in the process of multilevel analysis of longitudinal data, researchers rarely rely on CS but prefer residual covariance structures such as first-order autoregressive (AR1; e.g., Tan, 2010), and in experiments that include series of say 3–5 repeated measurements residual covariance structures like Huynh-Feldt (HF; e.g., Eyduran & Akbaş, 2010; Huynh & Feldt, 1970, 1976) are often more appropriate than CS. In some cases, a slight relaxation of the CS model that allows for unequal *SD*s may already do the trick, and yield testing and estimation outcomes that differ substantially from the restrictive CS model.

Cronbach's alpha constitutes a simplified case of generalisability theory (Brennan, 2001; Cronbach et al., 1972; Cronbach, Rajaratnam, & Gleser, 1963). Generalisability theory is an approach to questions like how many items or raters would be needed to achieve a certain reliability of measurement as expressed by Cronbach's alpha or the *ICC* on which alpha is based. One problem with these kinds of generalisability studies is the same as the use of Cronbach's alpha: CS is a very restrictive assumption that in many cases may be too restrictive, and other models will likely need to be considered for a more appropriate estimation of reliability and numbers of items or raters needed (see also Chaps. 14–16). This is one of the reasons why such generalisability studies, from a practical point of view, rarely make sense. With increasing numbers of items or raters, the likelihood of a violation of CS increases, even if these items or raters can be reasonably assumed to measure the same trait or state of interest. Moreover, with more items or more raters, the likelihood of a violation of undimensionality increases as well. For different raters, even if they provide simultaneous ratings about the same participants independent of other raters, it is unlikely to find ever-increasing numbers of raters that measure the same trait or state of interest to the same extent, and large numbers of raters are in practice usually not feasible anyway. For items, main problems are that respondents tend to become more and more tired with increasing numbers of items to be responded to, that at some point it becomes difficult to formulate more items about the same trait or state of interest that do not parrot other items, and that the repeated use of the same items comes at the risk of items being interpreted and responded to differently at subsequent occasions for a variety of reasons. All these issues together reduce generalisability theory to an approach of very limited potential from a practical point of view, and predictions such as one needing say at least twelve raters instead of the four included in a study at hand, at least ten items instead of the five items in the study in question, or at least eight repeated measurements instead of the two or three in the study just carried out, should not be given too much weight.

Greatest Lower Bound and McDonald's Omega

Several alternatives to Cronbach's alpha have been suggested, some of which are better than others (for a good overview, see for instance: Revelle & Zinbarg, 2009). Two main alternatives that have been mentioned by several scholars are the Greatest Lower Bound (GLB; e.g., Peters, 2014; Sijtsma, 2009) and McDonald's omega (e.g., Crutzen & Peters, 2017; Deng & Chan, 2017; Dunn et al., 2014; Green & Yang, 2009; Peters, 2014; Revelle & Zinbarg, 2009; Trizano-Hermosilla & Alvarado, 2016; Watkins, 2017; Zhang & Yuan, 2016). Dunn et al. (2014) summarize the advantages of omega over alpha as follows (p. 406): "*(1) Omega makes fewer and more realistic assumptions than alpha. (2) Problems associated with inflation and attenuation of internal consistency estimation are far less likely. (3) Employing 'omega if item deleted' in a sample is more likely to reflect the true population estimates of reliability through the removal of a certain scale item. (4) The calculation of Omega alongside a confidence interval reflects much closer the variability in the estimation process, providing a more accurate degree of confidence in the consistency of the administration of a scale*". Sijtsma (2009) suggests that the GLB is the lowest possible value that a scale's reliability can have, meaning that the true reliability should lie somewhere between the value indicated and 1 (perfect reliability). As noted by Ten Berge and Sočan (2004), the GLB has been ignored by many researchers due to a positive sampling bias (i.e., overestimating the reliability), especially when many items are involved and samples are small. Ten Berge and Sočan (2004) also indicate that McDonald's omega has the same bias. For that reason, some recommend to use the GLB and McDonald's omega only in the case of large samples (e.g., $N > 1{,}000$; Lorenzo-Seva & Ferrando, 2013). However, Ten Berge and Sočan (2004) also indicate that other reliability estimates—including Cronbach's alpha—also suffer from positive sampling bias, and that the bias for the GLB and McDonald's omega is especially a problem when the number of items involved is large and samples are small. In a recent study that compared Cronbach's alpha, McDonald's omega, and the GLB under realistic conditions, Trizano-Hermosilla and Alvarado (2016) concluded that when item distributions are approximately Normal, omega should be the first choice followed by alpha because they avoid the overestimation that GLB suffers from, but that in the case of moderate skewness the GLB or an adjusted version thereof may be preferred over omega and alpha.

In short, it appears that there is no coefficient which always works better than other coefficients. In the case of larger series of items, all coefficients may be problematic. However, in many questionnaires, the number of items over which reliability coefficients can reasonably be calculated usually varies from three to around six, because that is the set of items that measures for instance a particular aspect of motivation. A Cronbach's alpha over a set of dozens of test items can generally be expected to have little meaning, not in the last place because with the number of items increasing the likelihood of a violation of unidimensionality increases as well, and with such numbers of items the alternatives discussed here will have their problems as well. In the case of substantial departure from CS, it appears that either omega or the GLB are to be preferred, and in the absence of substantial deviations from normality omega. When departures from CS are minimal, Cronbach's alpha and omega should give very similar results. When in doubt, reporting both alpha and omega—or perhaps: alpha, omega, and GLB—is also an option.

Sample Size Issues

We already noted that reliability estimates tend to suffer from positive sampling bias and this is especially an issue when sample sizes are small. Moreover, in an experiment, for variables that result from psychometric measurement after the start of treatment, reliability estimation is preferably done *per condition*. Calculating a reliability estimate for different conditions together when these conditions differ from each other as a function of treatment will likely result in artificially inflated reliability estimates. For variables that result from psychometric measurement *before* the start of treatment, reliability estimates may well be calculated for the different conditions together, since variables measured before the start of the treatment are not affected by treatment. In other words, in an experiment with two groups of $n = 125$ participants each, we could calculate alpha, omega, and the GLB across $N = 250$ for data obtained from a questionnaire or other type of psychometric instrument (e.g., a pre-test or a prior knowledge test) measured before the start of treatment, but we would need to calculate these statistics per group of $n = 125$ participants for all instruments administered after the start of treatment. This is another reason to try and go beyond small samples when possible. In experiments that host $N = 50$ participants altogether, it will already be very difficult to obtain accurate estimates of reliability (i.e., the CIs around these estimates will be very wide); numbers like $n = 25$ for instruments after the start of treatment simply cannot be recommended for reliability estimation. Such cases force researchers to make the strong assumption that an instrument that demonstrated good reliability in past studies with similar participants is also reliable in a new experiment at hand. It is better to check that assumption whenever we can, and for that we need sufficiently large samples.

Latent Variable Approaches

Coefficient omega is based on a one-factor model in which the loadings of items on the factor are allowed to vary (e.g., Deng & Chan, 2017). Cronbach's alpha can also be represented in terms of a one-factor model, but one in which the loadings are equal. In cases of minimal deviation from equal loadings, alpha and omega should yield similar results.

The factor from the one-factor model is a *latent variable*, a variable that is not experienced empirically directly but only indirectly through *manifest variables* aka observables such as items or ratings. The basic idea behind a one-factor model is that if the latent variable of interest constitutes a common cause for a set of manifest variables, we can combine that set of manifest variables to measure that latent variable (unidimensionality; e.g., Sijtsma & Molenaar, 2002). If this assumption holds, once we specify a model in which the three items are represented as indicators of the latent variable, the correlations between the residuals of different manifest variables should be (approximately) zero. This assumption is also called *local independence* (e.g., Iramaneerat et al., 2010), constitutes a core idea of psychometrics, and allows us to fit models. When there are two uncorrelated or moderately correlated latent variables, and there is one set of items for each of these two latent variables, a one-factor model will probably indicate that unidimensionality and local independence are violated. In such a case, local independence can be achieved by specifying a two-factor model, where the correct sets of items are specified as two different sets of items that each measure one of the two latent variables (that may or may not be correlated).

Another important assumption here is *monotonicity* of the item-response function (e.g., Iramaneerat et al., 2010; Sijtsma & Molenaar, 2002), meaning a monotonous increase in expected score on quantitative manifest variables or a monotonous increase in probability of a higher category on a dichotomous or multicategory ordinal variable with an increase in the latent variable. An easy example of monotonicity is found in test items that are scored 'correct' or 'incorrect': monotonicity implies that the probability of a correct response on a given item monotonously increases with the respondent's knowledge or skill level tested.

Different types of latent variable methods exist, and which one to choose depends on the nature of the manifest variables and the (assumed) nature of the latent variable(s) of interest. Furthermore, although it is possible to test and estimate group differences in an experiment using latent variable methods, the extent to which any of these methods is feasible in an experimental study heavily depends on the sample size of the experiment. Generally, we are talking about sample sizes of several hundreds of participants or more.

Factor Analysis and Item Response Theory

A commonly used approach to latent variables is that of factor-analytic methods. Factor analysis constitutes a good approach to examining which quantitative items, ratings or manifest variables otherwise can be grouped together, provided that we are dealing with sufficiently large samples.

With regard to sample size guidelines, there are different approaches. On the one hand, there are scholars that recommend minimum required sample sizes irrespective of how many items we intend to include in our factor analysis. For example, Comfrey and Lee (1992, p. 217) suggest that "*the adequacy of sample size might be evaluated very roughly on the following scale: 50—very poor; 100—poor; 200—fair; 300—good; 500—very good; 1000 or more—excellent*." On the other hand, there are scholars who recommend minimum required sample sizes based on the number of items. More specifically, these scholars recommend a minimum participant-to-item ratio: at least 5:1 (in exploratory factor analysis; Gorsuch, 1983; Hatcher, 1994), at least 10:1 (in exploratory factor analysis; Nunnally, 1978), or at least 20:1 (Osborne, Costello, & Kellow, 2010). That said, no ratio will work in all cases, as the number of items per factor as well as communalities and item loading magnitudes also influence the ratio, but large samples are needed unless you are dealing with very strong data (e.g., MacCallum, Widaman, Preacher, & Hong, 2001; Osborne et al., 2010). What is clear is that sample sizes of $N < 100$ rarely provide a good scenario for factor analysis, unless perhaps in very exceptional cases where one wants to do a confirmatory factor analysis to test a one-factor model with three or four items, or to test a two-factor model with two sets of items, that fits very well.

Although factor analysis has constituted an important approach across fields in the context of Likert scales and categorical rating scales alike, recent research provides evidence for a tendency towards over-dimensionalisation (i.e., distinguishing more factors than should be distinguished)—of both exploratory and confirmatory factor analysis—when applied to such rating scale data (Van der Eijk & Rose, 2015). Van der Eijk and Rose (2015) therefore recommend extreme caution with factor analysis on this type of data and suggest item response theory (e.g., Hambleton, Swaminathan, & Rogers, 1991) models—such as the Rasch model (Andrich, 2004; Bond & Fox, 2007; Rasch, 1960; Wright & Stone, 1979) and the Mokken model (Mokken, 1971; Van Schuur, 2011)—as alternatives.

There are quite a few different item response theory models, and which one to use depends on the type of categorical variables one is dealing with (dichotomous, multicategory nominal, or multicategory ordinal) as well as on the type of data, what assumptions we are willing to make, what we want to do with our models, and what is our sample size. Item response theory models are widely used in educational and psychological research as well as in medical and health care assessment (e.g., Embretson & Reise, 2000; Hambleton, 2000; Hays, Morales, & Reise, 2000). Especially in experimental settings where sample sizes are usually limited (typically $N < 200$) and sets of items can be expected to measure a latent variable of interest (e.g., knowledge, skill, effort, motivation) to more or less the same extent, Rasch modelling may constitute the best approach. In the Rasch model, fewer parameters are estimated than in more complex item response theory models. Although the Rasch model is commonly criticised by proponents of more complex models that in many cases may provide a better fit of the data, the Rasch model has strong mathematical properties that provide researchers with a measurement model as a tool to make sense of a particular theoretical framework (Iramaneerat et al., 2010). The goal of the Rasch model is not to best fit the data but to provide *invariant measures* (e.g., Engelhard, 1994) of the degree of latent trait of *both* participants *and* items. This invariance allows researchers to model the measurement system in such a way that the order of participants in terms of increasing degree of the latent variable and the order of items in terms of increasing difficulty are invariant. This implies two things. Firstly, a participant with a higher degree on the latent variable should always have a higher probability of correct item response in a test than a participant with a lower degree on the latent variable, regardless of which items these participants encounter. Secondly, the probability of correct item response should always be lower for a more difficult than for an easier item, regardless of the degree of the latent variable of the participants who respond to those items (Rasch, 1960). For multicategory items, the *rating scale model* (Andrich, 1978; Embretson & Reise, 2000; Wright & Masters, 1982) and the *partial credit model* (Embretson & Reise, 2000; Masters, 1982; Masters & Wright, 1996) are widely used Rasch models. Other extensions of the Rasch model include the *binomial trials model* (Wright & Masters, 1982), the *Poisson model* (Rasch, 1960), the *Saltus model* (Wilson, 1989), the *many-faceted Rasch measurement model* (Linacre, 1989), the *linear logistic test model* (Fischer, 1973), and the *mixed Rasch model* (Rost, 1990, 1991). The latter combines principles of the Rasch model and *latent class analysis* (see next paragraph), meaning that the Rasch model is applied within each class and the parameters obtained from the Rasch model are allowed to vary across classes.

Latent Class Analysis and Latent Profile Analysis

Factor-analytic models and item response theory models have in common that the latent variables of interest are typically assumed to be *continuous*. Latent class and latent profile analysis provide alternatives when latent variables of interest are assumed to be *discontinuous*. Where factor analysis and item response theory models can provide researchers with scores on a continuous scale for the participants in a study, latent class and profile analysis enable researchers to estimate probabilities of participants being part of either of competing classes (Goodman, 1974; Lazarsfeld & Henry, 1968). Latent class analysis is used when manifest variables are categorical, while latent profile analysis is used when manifest variables are quantitative. That is, the two methods have in common that different classes can be distinguished in terms of their response patterns across items; in latent class analysis these response patterns concern different categories of nominal or ordinal items, while in latent profile analysis the response patterns are about differences in quantitative items. Latent class or profile analysis may be useful in studies that include at least a few hundreds of participants, and the numbers of participants required increase with the number of items included and number of classes expected, and become even more stringent (e.g., $N > 1{,}000$) when at least one of the classes has a low probability of occurrence. In experiments of modest sample size ($N < 200$), these methods are often difficult to use.

Latent Growth Analysis and Latent Transition Analysis

When changes in latent variables over time are concerned, *latent growth analysis* (McArdle & Nesselroade, 2003; Meredith & Tisak, 1990; Rao, 1958; Scher, Young, & Meredith, 1960; Tucker, 1958) and *latent transition analysis* (Collins & Lanza, 2010; Lanza & Collins, 2008) provide researchers with options with changes in quantitative and categorical variables, respectively. Latent growth modelling is part of the larger factor analysis and structural equation modelling framework, while latent transition analysis is conceptually related to latent class analysis; the term *growth* refers to changes in a continuous latent variable, while the term *transition* is used for movements from one latent class to another across time. In large-sample experiments (e.g., several 100s or more participants) that include a longitudinal or repeated- measures component with psychometric measurement, these latent change models may be useful.

Network Analysis as an Alternative Approach to Measurement

An alternative to the aforementioned latent variable models is found in *network analysis*. Succinctly put, given a collection of items, each fully connected subnetwork or *clique* of items generates a latent variable. A swarm of animals, for example, may move in a certain direction because there is local interaction between cliques (i.e., subgroups) of animals. These and most other forms of general intelligence may be achieved with mutualism (i.e., reciprocal causation, interaction; Van der Maas et al., 2006), and no hidden latent variables appear needed to explain what we see. If two sets of items in an instrument measure two different aspects of a construct (e.g., intrinsic and extrinsic motivation), this may appear in network analysis as two fully connected cliques, one for each set of items.

Network analysis approach provides new ways of thinking about items and constructs to be measured, and Open Source software such as JASP provides researchers with tools to perform network analysis. However, the testing and estimation procedures available for network analysis are still under research. Besides, conceptually, network analysis is quite a bit more complex than most if not all of the methods discussed earlier in this chapter. Finally, where factor-analytic and item response theory models provide straightforward ways to estimate reliability coefficients, it is not fully clear how network analysis can assist researchers in this enterprise. For these reasons, we have not yet seen many practical applications of network analysis published in educational and psychological research thus far. Perhaps this will change in near future, but for now some of the other methods discussed in this chapter appear to be preferred.

A Pragmatic Approach for Moderate Sample Sizes

Meehl (1990) discusses ten factors that together make most narrative summaries of research in social science research more or less uninterpretable. One of these factors is the *crud factor* mentioned in Chap. 2 of this book (everything correlates with everything). This crud factor is one of several factors discussed by Meehl as factors that tend to make bad theories look good. Other factors in this genre are selective bias in favour of submitting reports rejecting a point null hypothesis and selective bias by reviewers and editors towards accepting such papers, as well as the use of small-sample *pilot studies* to draw conclusions (e.g., on the existence of an effect) that cannot be drawn from pilot studies. As discussed in Chap. 2, small samples come with relatively low estimation precision and (severely) limited statistical power. Two pilot studies might well yield very different results about the same phenomenon of interest. This does not hold only for Ms, proportions, M_ds and differences in proportions; with small-sample studies, item-factor loadings and reliability estimates can vary wildly from one study to the next. Some researchers have used factor analysis in studies with $N < 50$ where the numbers of items were so large that they (almost) ended up estimating more parameters than they had participants in their samples. Calculations of Cronbach's alpha in samples of $n = 20$ are not uncommon in educational and psychological research. These practices, whichever outcomes they yield, are not recommended. Whether we want to estimate reliability using Cronbach's alpha, one of its more viable alternatives McDonald's omega or the GLB, through factor analysis, Rasch modelling, or otherwise, to obtain accurate estimates we need sufficiently large samples, and sample size requirements are generally more stringent with increasing complexity of design and models needed to account for that design. Whichever are the reasons why we cannot go beyond a certain sample size for a given experiment, the sample size limitations have implications for which analytic methods we can reasonably use. Besides, our questions, design features, and nature of the data should also direct our analytic choices.

If we have good reasons to expect two subgroups of 3–4 items in a psychometric instrument, each of which ought to measure one of two constructs of interest, departing from a 20:1 participant-to-item ratio, we would do well taking 20 times 6–8 items (i.e., two sets) or N = 120–160, meaning n = 60–80 in a two-group experiment. These numbers are well in line with the numbers discussed in Chap. 2 for experiments to ensure sufficient statistical power for medium size differences. Note that items that come from different instruments usually do not have to be included in the same factor analysis. Sometimes, researchers merge up to a few dozens of items into a single factor analysis, while these dozens of items come from a number of different instruments. Doing a factor analysis per instrument is in that case more appropriate. In a pub, chairs, tables, and crutches should in terms of their properties form separate factors. Likewise, for items from completely different instruments, the scales have different meaning even if they use the same number of categories and same numerical or verbal labels; putting them in one factor analysis is like comparing chairs, tables, and crutches in a pub. Suppose that the aforementioned two sets of 3–4 items each came from two different psychometric instruments, running a one-factor model on each of these two sets of items is okay.

The recent work by Van der Eijk and Rose (2015) gives us reasons to rethink our common practice of using factor analysis as a default approach to assessing the structure of a psychometric instrument: for Likert and other forms of categorical scales, which are likely candidates for factor analysis in educational and psychological research, both exploratory and confirmatory factor analysis tend to result in over-dimensionalisation. For such data, we may want to consider an increased use of item response theory models such as the Rasch model.

That said, whether we use factor analysis, item response theory models, or other latent variable methods to examine the structure of a psychometric instrument, whether we need reliability estimates additional to the outcomes of these latent variable methods remains a question of debate. Some, including myself, argue that if a factor analysis, a Rasch model or a latent class/profile model indicates good fit, there is no need for further reliability estimates; the key outcomes of the factor analysis, Rasch model or latent class/profile model usually provide sufficient information to understand the data and allow those interested to calculate additional reliability estimates. Others may argue it will still be good to also report additional reliability estimates. If so, I hope this chapter has provided reasons to revise our common practice of Cronbach's alpha as the default option and to consider reporting McDonald's omega and the GLB instead of or along with Cronbach's alpha. Also, Chaps. 14–16 of this book place Cronbach's alpha and its restrictive CS assumption in a broader mixed-effects modelling perspective and provide other residual covariance structure-based alternatives to CS.

Dealing with Missing Data

4

Abstract

A topic not yet touched in the previous chapters of this book is that of how to deal with missing data. Although missing data is commonly associated with other types of research and large proportions of missing data are uncommon in most of experimental educational and psychological research, the topic is equally relevant to experimental research. Frequently encountered methods for dealing with missing data are mean imputation, listwise deletion, pairwise deletion and, in repeated-measures and longitudinal studies, last observation carried forward. Although these methods are very easy to implement, they are usually wrong and may substantially distort our view of effects of interest for reasons discussed in this chapter. Four somewhat more complex yet generally more appropriate approaches to missing data are matching, regression imputation, FIML, and MI. After a comparison of these four methods in terms of their pros and cons, this chapter provides a pragmatic approach to dealing with missing data.

Introduction

For the sake of simplicity of the introduction of concepts, Chaps. 1, 2 and 3 do not deal with missing data situations. Barnard and Meng (1999), Cole (2010) summarize three major problems with incomplete data due to missing response: loss of information and statistical power, complications in data management and analysis, and the risk of biased estimation and testing outcomes with regard to effects of interest. How to deal with missing data depends on the expected nature of the missingness. More than four decades ago, Rubin (1976) proposed a framework of three types of missing data: MCAR, MAR, and MNAR.

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_4

Under MCAR, the probability of missing response is unrelated to our observed variables of interest and is unrelated to unobserved variables that might affect our variables of interest, because the missingness occurs for completely *unsystematic*, random reasons (Abraham & Russell, 2004). In experiments that include an online component, for example, a random disturbance may result in a (partial) loss of data for some participants. In such cases, the group of participants with complete data is considered a random subsample of all participants in the experiment (e.g., Cole, 2010). Although the assumption of no relation between missingness and unobserved variables cannot be tested, several software programmes include Little's MCAR test (Little, 1988) for as far as the observed variables are concerned. Rationale behind this test is that apart from the missingness, the two groups should not differ significantly from each other in the observed variables of interest; otherwise, the MCAR assumption can be rejected.

In short, if the reason of missingness can reasonably be expected to be random, such as in the experiment where a random disturbance results in random omissions in a dataset, MCAR may be assumed, and when in doubt, Little's MCAR test provides an objective way to test the MCAR assumption. If MCAR does not hold, MAR is a next candidate. MAR occurs when the probability of missingness on variable *X* is related to one or more other observed variables in the dataset but not to the value of *X* itself (e.g., Acock, 2005). In longitudinal research, which is especially susceptible to missing data, within MAR a distinction can be made based on whether the probability of missingness depends on the *previous response* (MAR1) or on the *previous two responses* (MAR2; De Rooij, 2018; Rubin, 1976). Contrary to MCAR, MAR cannot really be tested, because there is no way to verify that the probability of missing data on *X* is related only to observed variables and not to unobserved variables as well. Some of the observed variables that may correlate with the occurrence of missingness of *X* can be used as explanatory or *auxiliary* variables (Collins, Schafer, & Kam, 2001) to establish correlates of missingness to decide on how to deal with the missing data. Misclassifications of missingness as MCAR or MAR often have a rather minor impact on point estimates and *SE*s (Cole, 2010; Collins et al., 2001), and a major advantage of MCAR and MAR over MNAR is that we do not need to model the *mechanism* of missingness. As such, data are often said to be 'ignorable' if they are either MCAR or MAR and "*the parameters that govern the missingness mechanism are unrelated to the processes to be estimated*" (Cole, 2010, p. 217).

The most troublesome missing data mechanism is MNAR, because the probability of missingness on variable *X* depends on the actual value for the non-responding participant(s) on *X*, even after controlling for other observed variables. Basically, it is impossible to even verify MNAR without knowing the missing values, and the latter is in practice usually the case. Under MNAR, in contrast to MCAR and MAR also referred to as *nonignorable data* (Cole, 2010), we need sophisticated methods to model the missingness mechanism (e.g., De Rooij, 2018; Molenberghs & Verbeke, 2005).

In well-controlled and carefully managed experiments with single or small series of repeated measurements, missingness occurs much less frequently than in experimental or non-experimental longitudinal studies. The control and careful management that makes an experiment a good experiment often helps to avoid missingness or otherwise to minimise it (i.e., both participant and item non-response remaining well below 10%) and reasonably establish the mechanism where it occurs. While MNAR may be quite likely in many non-experimental studies, MCAR and MAR are generally much more likely in well-controlled and carefully managed experiments. Fallout of a server or failure of email are likely forms of MCAR. Participants not being able to respond to all items of a final questionnaire in a fixed-time experimental session due to a lack of time may be considered MAR.

As becomes clear later on in this chapter, how to deal with missing data not only depends on the type or mechanism of missingness and percentages of missing (per participant and per item); how many times a variable of interest is measured (e.g., De Rooij, 2018) and whether that variable serves as an outcome variable or as a covariate (Horowitz & Manski, 2000; Janssen et al., 2010; White & Carlin, 2010) are also questions to be considered.

Simple Missing Data Methods

Despite over five decades of research on missingness, the use of several simple but wrong missing data methods continues to be widespread across fields: mean imputation, listwise deletion, pairwise deletion, and last observation carried forward.

A first commonly encountered simple missing data method is that of mean imputation. In this case, all participants whose scores on *Y* are missing are replaced by the mean of *Y* based on the participants that do have a score on *Y*. This method is known to result in underestimated *SD*s and biased estimates of correlations with it (Haitovksy, 1968). Moreover, especially under MAR and MNAR, this method tends to result in biased estimates of treatment effects (e.g., Eekhout et al., 2014). Although in a two-way variant on mean imputation in questionnaires, the score to be imputed is a function of both the mean of *Y* and the mean of the participant on other items that are supposed to measure the same construct (Van Ginkel, Sijtsma, Van der Ark, & Vermunt, 2010), the risk of biased estimates remains.

A second approach that is commonly used, perhaps because it has been the default option in frequently used statistical packages, is that of listwise deletion aka casewise deletion (Abraham & Russell, 2004; Acock, 2005; Cole, 2010). In this method, all participants with any missing data are removed from the analysis, regardless of the proportion of that missingness. Even under MCAR, this approach tends to result in an unnecessary loss of statistical power and precision, unless we are dealing with larger samples and the proportion of missingness is less than 5%

(Cole, 2010). Moreover, under MAR and MNAR, it also tends to result in biased estimates (Enders, 2010), even in the case of smaller proportions of missingness.

Another deletion approach that is sometimes used is that of pairwise deletion aka available-case analysis (e.g., Cole, 2010). In this case, if a participant has missing on variable *X* but not on variables *Y* and *Z*, that participant is included where *Y*-*Z* relations are concerned but not where *X*-*Y* or *X*-*Z* relations are concerned. Main problems of this approach lie in loss of statistical power and inconsistency of *SE*s and other statistics across comparisons (e.g., Little & Rubin, 2002; Schafer & Graham, 2002) as well as, under MAR and MNAR, expectedly biased estimates (Baraldi & Enders, 2010). In cases where a variable *X* that has missing data is considered as a predictor in some models for the explanation of *Y*—because *X* is a treatment factor or a relevant covariate—but not in some competing model(s) of *Y* (e.g., the 'no difference' model), pairwise deletion comes with the problem that the data included is not the same across competing models. This undermines any kind of comparison between competing models in terms of *p*-values, information criteria or BFs.

Finally, in longitudinal studies, *last observation carried forward* (Peto et al., 1977) is sometimes applied. In this approach, the missing value of a participant on variable *X* at occasion O_t is imputed with last observation (O_{t-1}) on the same variable *X* for that participant. This assumption only makes sense in longitudinal studies where change does not occur frequently or where changes are generally small, and this is something that is unrealistic in most longitudinal studies (Wood, White, & Thompson, 2004). In other words, underestimated *SD*s and biased correlation estimates are a main concern in this method as well.

More Complex Approaches to Missing Data

This section discusses four approaches to missing data that are more complex but are generally better in one several ways compared to the methods discussed in the previous section: matching imputation, regression imputation, FIML, and MI.

Matching imputation is a method in which non-responding participants are matched to participants that in terms of other observed variables are similar, and scores on variable *X* of matched participants who did respond are given to the matched participants whose scores on *X* are missing. Two variants of hot-deck imputation are the distance function approach and the matching pattern approach. In the case of distance function, the imputation score is found directly from the 'nearest neighbour' participant (i.e., the participant who based on the other observed variables has the smallest squared distance statistic to the participant whose score on *X* is missing). In the case of matching pattern, the total sample is stratified into a limited number of homogenous subgroups, and the imputation score for the missing participant is a random draw from cases in the same homogenous subgroup (e.g., Fox-Wasylyshyn & El-Masri, 2005). Compared to the methods discussed in the previous section, this approach results in imputations that are more realistic and somewhat more respect the distribution of observed *X* scores. However, risks of underestimated *SE*s remain (e.g., Roth, 1994).

In a second approach, regression imputation, we perform regression analysis with other observed variables to obtain predicted values where there is missingness. This approach is more robust than the methods discussed in the previous section, because it enables researchers to incorporate a variety of indicators to obtain realistic imputation values that are consistent with observed relations between variables (Cole, 2010). However, when several variables have missing data, the process can become quite complex. Moreover, imputed scores are made to fit an observed straight line, and this artificial reduction of residual variance is likely to result in somewhat underestimated *SE*s (Enders, 2010).

Matching and regression imputation have in common with the methods discussed in the previous section that no additional datasets are created. In the case of MI, several imputed datasets are created. Basically, MI involves three stages: *imputation*, *analysis*, and *pooling*. In the imputation phase, different versions of the dataset are created in which the observed values are always the same but the imputed values differ based on an iterative regression approach. This yields a limited number of (e.g., 5–10) after imputation 'complete' datasets. Each of these datasets is analysed in the second stage (i.e., analysis stage) with the statistical methods one would normally use on complete datasets. Finally, in the pooling stage, the parameter estimates and *SE*s obtained in the different datasets are pooled to obtain outcomes that respect the distributions of the different variables involved, enable researchers to model the bias in and between imputed datasets, and can—at least under MAR (e.g., Eekhout et al., 2014)—produce unbiased estimates. In the case of MNAR, the robustness of the imputations and outcomes may be examined via the consideration of a range of auxiliary variables, variables that correlate with the variable that needs imputation and/or with the probability of missing response on that variable (Collins et al., 2001). Auxiliary variables can also be useful in the aforementioned regression imputation and can be useful in the approach discussed next as well. Examples of commercial software packages that include MI are *SPSS* and *Stata*. A good package from *R* is the *multivariate imputation by chained equations* (MICE) package (e.g., Azur, Stuart, Frangakis, & Leaf, 2011; Horton & Lipsitz, 2001; Luo, Szolovits, Dighe, & Baron, 2017; Van Buuren & Groothuis-Oudshoorn, 2007).

In all methods and approaches discussed in this chapter thus far, missing values are imputed. In FIML, missing data is accounted for in the analysis without any kind of imputation taking place, by using all available information without pairwise deletion. FIML estimation of multilevel models, confirmatory factor models (e.g., *Jamovi*) and structural equation and latent growth models (e.g., *Mplus*) are good examples of this approach.

Collins et al. (2001) demonstrated that in many practical situations, MI and FIML will produce similar or very similar results. Some studies suggest that MI may be preferred over FIML when sample sizes are small (e.g., Graham & Schafer, 1999; Schafer & Graham, 2002), although for MI a participant-to-variable ratio of 10:1 appears a recommendable lower bound (Cole, 2010). MI appears to have more potential than FIML in cases of MNAR through the inclusion of more auxiliary variables, but more auxiliary variables also elevates sample size requirements. FIML estimation is attractive to researchers who prefer not to impute missing values, is easier to carry out than MI, and provides straightforward model indices for a single dataset. Moreover, some work indicates that MI in small samples can produce biased outcomes (Hayes & McArdle, 2017) and that FIML may be better than MI in small samples under MAR although in small samples FIML may yield substantial bias as well (Yuan, Yang-Wallentin, & Bentler, 2012).

A Pragmatic Approach to Dealing with Missing Data in Experiments

This final section of this chapter provides some pragmatic guidelines for how to deal with missing data in an experiment depending on whether or not variables are measured repeatedly, what type (mechanism) of missingness is expected, how many times a variable of interest is measured, and whether that variable serves as an outcome variable or as a predictor variable. Generally speaking, the problem with including unimputed predictor variables with missing data in our models is that comparisons between models—some of which include and some of which exclude one or several of the predictor variables with missing data—cannot really be made because they do not use exactly the same data. For instance, take an experiment with one treatment factor, one covariate measured prior to treatment, and one outcome variable. Even if there is only *one* case missing on the covariate, doing nothing and just including treatment factor and covariate in different competing models—which do or do not include the treatment factor and/or covariate—results in the following model. For models that do *not* include the covariate (i.e., the null model and the treatment-factor-only model), data from N participants is used. However, for all models that *do* include the covariate, data from N – 1 number of participants is used. This affects *SE*s and all that is based on it, including CIs and p-values. Besides, all model comparisons involving BFs or information criteria like AIC and BIC are based on the premise that *exactly* the same data is used in each of the models. For missing in an outcome variable, the situation is different: given missing on the outcome variable but not on any of the predictor variables, the data used for model comparison is *always* the same.

Single-Time Measurement of Variables

In many experiments in educational and psychological research, variables of interest are measured once in time, even though the timing of that measurement may differ for different variables: outcome variables are measured after the start of treatment, and predictor variables may be measured any time before (i.e., possible covariates and moderators) or after the start of treatment (i.e., possible mediators and moderators) but before the final outcome variables. Moreover, well-controlled and carefully managed experiments may result in only a small percentage of missing (if any at all), say 5% or up to 10%. In such situations, the loss of statistical power due to listwise deletion may be small. For instance, consider an experiment of $N = 128$ that starts with two groups of $n = 64$ each. Suppose, in both groups, six participants have some missing data on a questionnaire or test (i.e., 9.4% in each group). With listwise deletion, we are left with two groups of $n = 58$ each. For Cohen's $d = 0.50$ and a two-sided test at $\alpha = 0.05$, that means a reduction in statistical power of 0.80–0.76. Given that listwise deletion tends to result in biased estimates under MAR and MNAR but not under MCAR, under MCAR (e.g., random fallout of a server or mail failure) with such small proportions of missing listwise deletion may be defendable. Besides, in some odd cases, it happens that due to failure of a server we lose the information with regard to one or a few participants with regard to what condition they were in. With a solid experimental setup and good data processing and registration equipment, this kind of missing should in most cases be zero and otherwise constitute less than 5% or in any case less than 10% of the cases. In such cases, data may well be MCAR (i.e., failure of a server affecting a random participant at time point T), and listwise deletion may be considered. The problem with FIML is then that across different competing models—some of which may and some of which may not include the treatment factor of interest—not exactly the same data may be used (cf. the aforementioned N vs. $N - 1$ example with a missing data point on a covariate). MI might be considered but should be done carefully, and perhaps with somewhat larger numbers of imputations, for there may be a considerable chance of misclassifying (at least) one participant in terms of condition membership.

In the case of well-designed and carefully managed experiments that do not involve a repeated-measures or longitudinal component, more than 20% of participants having missing data is rather uncommon. When the group of participants who have some missing data is somewhere in the 10–20% range and missing is only on one or more outcome but not on any predictor variables, under MCAR and MAR both MI and FIML can be considered, and under MNAR one is likely better off with MI. When combinations of missing on predictor and outcome variables are concerned and the problem cannot be resolved by deleting the one or few participants whose data is missing on *all* or almost all of these variables, MI is recommended. Besides, for questionnaires and instruments alike that consist of series of items, how to use MI also depends on how much missing individual participants have on a questionnaire or test of interest. When participants have fairly small percentages of items missing (e.g., 1 out of 5 items in a scale or 20%, or 3 out of 10 items in a scale or 30%), MI can be done at the item level (i.e., imputing item scores). However, if participants have most of the items missing (e.g., 4 out of 5 items or 80%), MI on the total score of a set of items is probably better.

MI may still be useful if percentages of missingness exceed 20%, and perhaps in some cases FIML estimation may under MAR also provide a satisfactory solution, but the larger the percentages of missingness the more worrisome things get. I have rarely seen an experiment where over 50% of participants had missing data, but if the occurrence of missingness goes beyond such a level one may wonder what is the use of analysing the data at all—especially in cases of MNAR and/or where the percentages of missingness per participant are also high—unless missingness results from a dropout that is in itself considered an outcome variable of interest in the context at hand.

Multiple-Time Measurement of Variables

Sometimes, the same outcome variable of interest is measured repeatedly in a short time interval (i.e., repeated-measures design) or two or more times over a longer time interval (i.e., longitudinal design). The same guidelines as for single-time measurement of variables can be applied when it comes to missing in predictor variables, and whichever method we sue, we should always explicitly report which method we have used, how we have used that method, and why we have done so.

Apart from the guidelines for single-time measurement of variables, when outcome variables are measured repeatedly, larger proportions of missing data may become more likely, unless we have a very solid setup and context which help to minimise the occurrence of missing data. MI may provide useful solutions under MCAR, MAR or MNAR as long as the proportions of missingness do not exceed such levels that data analysis altogether becomes questionable (e.g., 80% of participants having missing data on the same outcome variable at several occasions, especially under MNAR) and/or the effective sample size due to missingness becomes so small that MI can no longer be expected to provide meaningful estimates (see also the previous note on FIML providing somewhat less biased estimates than MI under MAR in smaller samples). Finally, under MCAR and MAR, missingness in outcome variables can be handled with FIML, unless the percentages of missing are very high (e.g., 80% of participants having missing data on the same outcome variable at several occasions) and/or the effective sample size due to missingness becomes so small that FIML is no longer a reasonable option.

Part II
Types of Outcome Variables

5 Dichotomous Outcome Variables

Abstract

The main take home message from Part I of this book is that whether we deal with simple group comparisons (Chap. 2), measurement issues (Chap. 3) or missing data (Chap. 4), data-analytic choices ought to be driven by the questions that led us to do the experiment, by the features of the experimental design that resulted from our questions, and by the nature of the data acquired in the experiment (the QDA bridge from Chap. 1). In this second part of the book, this approach is applied to different types of outcome variables. In this first chapter of Part II, we focus on dichotomous outcome variables. Examples of dichotomous variables are pass/fail decisions in tests, recover/failure to recover distinctions in mental health-related contexts, and event occurrence/event absence. This chapter discusses different plots and statistics for experiments in which a dichotomous outcome variable is measured once in time as well as for experiments in which the outcome variable is a dichotomous variable in the form of event occurrence/absence in a particular time period. Although the latter is commonly associated with survival analysis in hospitals, (simulated) traffic research for example may focus on the occurrence or absence of accidents in different groups of participants studied.

Introduction

Although many experiments in educational and psychological research involve quantitative variables or multi-category ordinal variables that are treated as if they were of interval or ratio level of measurement, there are experiments in which at least one of the outcome variables of interest is a dichotomous, two-category variable. In some experiments, that two-category variable results from dichotomising a quantitative or multicategory variable, which is rarely recommendable because it usually results in unnecessary information loss and a loss of statistical power and precision

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_5

with it. However, where dichotomous variables arise from qualitative judgements or from the occurrence/absence of a phenomenon of interest (e.g., a participant having a missing response on variable Y or not, see Chap. 4), we need appropriate analytic methods for dichotomous variables. Researchers often report a χ^2-test of a 'no effect' null hypothesis with its p-value and leave it there. This is a pity, because we learn little from just a χ^2- and p-value, but if researchers at least report the observed frequencies in a contingency table they allow other researchers to compute statistics that may provide us with much more useful information about effects of interest than χ^2- and p-values. Three examples of a simple two-group experiment with a dichotomous outcome variable are discussed in this chapter: two examples with pass/fail decisions arising from a qualitative judgment on a skills test at the end of an experiment, and the occurrence of an accident in an experiment on different ways of training driving skills in a simulated environment.

Experiment 1: Effect of Treatment on the Probability of Passing a Test

Suppose, a group of medical education researchers have developed a new technique for training communication skills with virtual patients among undergraduate medical students. They want to compare this technique with a conventional technique for that skills training, and develop a practice scenario and a test scenario for an experiment. With the conventional technique, the pass rate at the first attempt is about 65%. Both scenarios are presented in the same online environment. The practice scenario constitutes the learning stage or practice period in preparation for the test scenario. In the control condition, participants approach the practice scenario with the conventional techniques. In the experimental treatment condition, participants approach that same practice scenario with the new technique. After the practice scenario, participants take the test (i.e., same test scenario) without any kind of the help that was included in the conventional or new technique.

The researchers randomly recruit $N = 300$ undergraduate medical students in the United States, because this is about the largest possible sample they can draw with the logistic possibilities they have, and this number can guarantee a statistical power of about 0.80 for a two-sided test at $\alpha = 0.05$ for substantially better (80% or higher) or substantially worse (48% or lower) pass rates with the new technique (*GPower*), knowing that the $N = 300$ students are going to be allocated randomly to the two conditions such that each condition hosts $n = 150$ different students. Pass/fail decisions are made by teachers who have extensive experience with skills training in the context at hand but who are blind to which participants were trained with which technique and do not yet know of the new technique let alone have any expectations with regard to how that technique might affect pass/fail rates.

A Treatment Effect?

In the control condition, 100 of the n = 150 participants (about 66.7%) pass the test. In the experimental treatment condition, 107 of the n = 150 participants (about 71.3%) pass the test. Figure 5.1 (from JASP) depicts the conditional estimates plot aka *estimated marginal means* (*EMM*) plot for the control condition (left: $X = 0$) and for the experimental treatment condition (right: $X = 1$): the observed pass rates with 95% CIs around them.

Figure 5.1 indicates that the treatment effect observed in the sample is small. Given the absence of an a priori one-sided hypothesis with regard to the direction of an eventual treatment effect, two-sided testing is performed. Using the LR test for $\boldsymbol{H_0}$ stating 'no treatment effect' against $\boldsymbol{H_1}$ stating 'treatment effect', we find $\chi_1^2 = 0.764$, $p = 0.382$. AIC is 373.460 for Model 0 ($\boldsymbol{H_0}$: no treatment effect) and 374.696 for Model 1 ($\boldsymbol{H_1}$: treatment effect). BIC is 377.164 for Model 0 and 382.104 for Model 1. Thus, AIC and BIC both prefer Model 0.

When it comes to estimating R^2 (see also Chap. 2), quite a few different R^2-statistics have been proposed in the literature in the context of logistic regression analysis, some of which appear better than others (e.g., Menard 2000; Mittlbock and Schemper 1996). One of the most reported and perhaps also still one of the best R^2-statistics comes from McFadden (1974). McFadden's R^2_{McF} corresponds to a *proportional reduction or error variance* similar to the R^2-statistic for quantitative outcome variables. It can be calculated from the *deviance* (-2LL) of Model 0 (null model, no treatment effect) and that of Model 1 (treatment effect):

$$R^2_{\text{McF}} = 1 - ([\text{-2LL}\,\underline{\text{Model 1}}]/[\text{-2LL}\,\underline{\text{Model 0}}]).$$

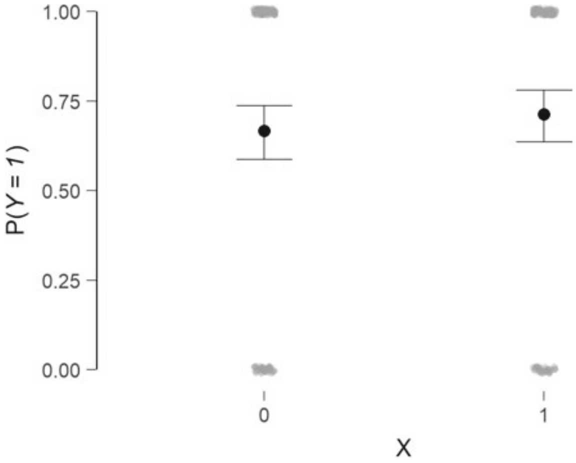


Fig. 5.1 *EMM* plot of Experiment 1: (observed) point and interval (95% CI) estimates of the probability of pass $P(Y = 1)$ for the control ($X = 0$) and treatment ($X = 1$) condition (JASP)

In Experiment 1, -2LL Model 1 = 370.696, and -2LL Model 0 = 371.460, hence:

$$R^2_{\mathrm{McF}} = 1 - (370.696/371.460) \approx 0.002.$$

Another more recently proposed R^2-statistic that is conceptually closely related to the R^2-statistic for quantitative outcome variables, because it is based on predicted probabilities (proportions), is that of Tjur (R^2_{T}; 2009). However, contrary to R^2_{McF}, R^2_{T} is not easily applicable in the case of multicategory nominal (Chap. 6) or ordinal outcome (Chap. 7) variables. Moreover, at least when R^2_{McF} is high, the behaviour of R^2_{T} is somewhat odd, and we will see what is meant by that after discussing two other commonly encountered R^2-statistics that are applicable to dichotomous, multicategory nominal, and ordinal outcome variables: Cox and Snell's R^2_{CS} (Cox and Snell 1989; Cragg and Uhler 1970; Maddala 1983) and Nagelkerke's R^2_{N} (Nagelkerke 1991). However, these two coefficients are computationally more complex and are quite dependent on the marginal probability of an event (e.g., the overall probability of pass). More specifically, R^2_{CS} has the undesirable feature that, contrary to what an R^2-measure should be able to do, its highest possible value (i.e., upper bound) cannot be 1; for marginal probabilities around 0.5, its upper bound is about 0.75, while for marginal probabilities further away from 0.5 (e.g., 0.1 or 0.9) its upper bound can go below 0.50. R^2_{N} corrects for that by dividing R^2_{CS} by its upper bound. Although that correction yields an R^2-statistic that has an upper bound of 1 and remains with a lower bound of 0, the R^2_{N} may more often than not be somewhat exaggerated, especially when the upper bound of R^2_{CS} is fairly low (e.g., 0.50).

If you find the discourse about R^2_{CS} and R^2_{N} difficult to follow, you can simulate an example for yourself. Open any statistical software package that provides R^2_{McF}, R^2_{T}, R^2_{CS}, and R^2_{N}, for instance *JASP*. Create say 40 cases, 20 of which score '1' on two variables X and Y and 20 of which score '0' on these two variables. A good R^2-statistic should yield 1.000 as outcome for this dataset, because there is a *perfect* relation between X and Y. The -2LL of Model 0 (i.e., no relation aka statistical independence) is 55.452, while that of Model 1 (i.e., relation) is *exactly zero* because Model 1 fits the data perfectly. Consequently, R^2_{McF} = 1.000. However, *JASP* and *Jamovi* also return the following outcomes: R^2_{T} = 0.513 (i.e., only in *JASP*, not available in *Jamovi*, at the time), R^2_{CS} = 0.750, and R^2_{N} = 1.000. The outcomes of R^2_{T} and R^2_{CS} simply do not make sense.

The marginal probability in the aforementioned exercise is 0.50, and we see that the upper bound—that is: the maximum possible value of the statistic, which we find in the case of *perfect* relation—is 0.750 for R^2_{CS} and 0.513 for R^2_{T}. Now let us redo the exercise with a marginal probability of 0.80 (i.e., 0.20 would yield the same outcomes), which is somewhat further away from the 0.50 value that might be considered 'ideal' in the sense that we have as many cases with '0' as we have cases '1' on the outcome variable of interest. For our exercise with a marginal probability of 0.80, we create 40 cases, 32 of which score ‘1’ and the other 8 score ‘0’ on the outcome variable of interest. The -2LL of Model 0 is 40.032, while that of Model 1 is 0. *JASP* and *Jamovi* return the following values: R^2_{McF} = 1.000, R^2_{T} = 0.821 (i.e., only in *JASP*, not available in *Jamovi*, at the time), R^2_{CS} = 0.632, and R^2_{N} = 1.000.

These two exercises, with marginal probabilities of 0.50 and 0.80 (or 0.20), clearly demonstrate that some caution is needed when choosing R^2-statistics for categorical variables. R^2_{T} and R^2_{CS} suffer from the same problem of upper bounds substantially lower than 1, and the correction applied by R^2_{N} may be too heavy, especially for marginal probabilities somewhat further away from 0.50. Although several other R^2-statistics have been proposed, R^2_{McF} is probably still the best R^2-statistic for categorical outcome variables (Kvålseth 1985; Menard 2000). Thus, if we want to report R^2-statistics for dichotomous outcome variables (logistic regression) that is also easily applicable to multicategory nominal and ordinal outcome variables and that *can* cover the full 0–1 range, the best option for now is R^2_{McF}. R^2_{McF} is available in virtually all statistical packages, including *Jamovi*, *JASP*, *SPSS*, and *Stata*. In the latter, it is the default R^2 for regression models with categorical outcome variables and is found under the name ‘pseudo R^2’. In Experiment 1, $R^2_{\mathrm{McF}} \approx 0.002$. In other words, the effect observed in the sample is small.

Other useful statistics with regard to the effect of interest are Cramér’s V (Cramér 1946) and the *OR* and ln(*OR*) (the latter is also called log *OR* or *logit* and is found by taking the natural logarithm, ln, of the *OR*; Agresti 2002). In the case of two dichotomous variables—as is the case for the experiments discussed in this chapter: a dichotomous treatment variable (treatment vs. control) and a dichotomous outcome variable (pass vs. fail)—Cramér’s V yields the same point estimate as Pearson’s r, Spearman’s ρ, Kendall’s τ coefficients, and coefficient φ (Agresti 2002; Guilford 1936). In Experiment 1, we find $V = 0.050$, which corresponds with small differences (Cohen 1988; Lipsey and Wilson 2001). The interpretation of V-values in terms of ‘small’, ‘medium’ and ‘large’, like for any effect size statistic, depend on the context; what is ‘large’ in one context may be considered ‘modest’ in another context and vice versa. Moreover, in the case of V, small-medium-large interpretations of V depend on the number of categories of the variables involved. As long as one of the two variables involved is a dichotomous variable, 0.1, 0.3, and 0.5 are commonly interpreted as ‘small’, ‘medium’, and ‘large’ effects, but these labels are associated with larger values when both variables involved have three or more categories.

The point estimate of the *OR* in Experiment 1 is 1.244 (p = 0.383), and the 95% CI of the *OR* extends from 0.762 to 2.032. The ln(*OR*)—also denoted as b because it can be interpreted as a regression coefficient (though not in the same way as in a linear model!)—is obtained by taken the *natural logarithm* (ln) from the *OR*. In Experiment 1, we find b = 0.218, and the 95% CI extends from −0.272 to 0.709. *OR*s are very useful in many statistical applications for categorical variables, including in meta-analyses that include categorical variables. The b of 0.218 in Experiment 1 is the slope in logistic regression, analogous to the regression coefficient of a treatment effect in linear regression (though not the same!).

Relatively Equivalent?

The statistics discussed thus far indicate that the effect observed in Experiment 1 is small. However, as discussed in Chap. 2, from effects being small and not statistically significant at $\alpha = 0.05$, we cannot conclude that the conditions compared in an experiment are equal or equivalent in a population of interest, even if other criteria such as AIC and BIC also indicate a preference towards the 'no treatment effect' model. TOST equivalence testing and the Bayesian ROPE approach provide better ways to establish relative equivalence of the treatment and control condition. Both approaches require researchers to reach consensus about a range of values on a given statistic in which two conditions can be considered relatively or practically equivalent. In the context of Experiment 1, researchers in the field may agree that a difference between conditions within a range of 10% is not of substantial practical interest and hence the region of −10% (negative treatment effect) to +10% constitutes the region of practical equivalence. The 95% CI for the difference in pass rate (71.3% in the treatment condition, 66.7% in the control condition) ranges from −0.058 to 0.151 and the 90% CI ranges from −0.0411 to 0.134 (*Jamovi*). Although we can reject $\boldsymbol{H_{0.1}}$: $\pi_{\text{treat}} - \pi_{\text{control}} < -0.10$ (i.e., the difference being more negative than 0.10; $p = 0.003$), we fail to reject $\boldsymbol{H_{0.2}}$: $\pi_{\text{treat}} - \pi_{\text{control}} > 0.10$ (i.e., the difference being more positive than 0.10; $p = 0.159$). In terms of FOST, our evidence appears *inconclusive*: we have insufficient evidence for relative equivalence (i.e., the 90% CI would have to be within [−0.10; 0.10]) and we have insufficient evidence to reject relative equivalence as well (i.e., the 90% CI would have to be fully outside the −0.10 to 0.10 range).

Note that significance tests for differences in proportions are *z*-tests, not *t*-tests. Reason for this is that in the case of dichotomous variables the variance σ^2 is a direct function of the proportion π:

$$\sigma^2 = \pi * (1 - \pi).$$

Thus, given π, σ is known. When dealing with quantitative outcome variables, we normally use *t*-tests that *estimate* the population standard deviation σ from the sample *SD*; in situations where σ is known, *z*-tests can be used. That said, in Experiment 1, the findings leave us inconclusive; we fail to find evidence for a treatment effect *and* fail to find evidence to assume relative equivalence. More experiments will be needed to accumulate more information, so that perhaps a meta-analysis on a series of experiments can help us establish point and interval estimates that do allow us to more safely draw conclusions with regard to the treatment effect of interest.

Experiment 2: Effect of Treatment on the Probability of Successful Completion

Suppose, the conventional technique actually yields much lower (first-time) pass rates and that was the reason why the group of medical education researchers developed a new technique. They run the experiment under the same conditions as explained for Experiment 1 and the findings are as follows. In the control condition, 43 of the $n = 150$ participants (about 28.7%) pass the test. In the experimental treatment condition, 115 of the $n = 150$ participants (about 76.7%) pass the test. Figure 5.2 (from JASP) depicts the *EMM* plot for the control condition (left: $X = 0$) and for the experimental treatment condition (right: $X = 1$): the observed pass rates with 95% CIs around them. Contrary to Fig. 5.1 (Experiment 1), Fig. 5.2 illustrates a very clear positive treatment effect.

Given the absence of an a priori one-sided hypothesis with regard to the direction of an eventual treatment effect, two-sided testing is performed. Using the LR test for $\boldsymbol{H_0}$ stating 'no treatment effect' against $\boldsymbol{H_1}$ stating 'treatment effect', we find $\chi^2_1 = 72.311$, $p < 0.001$. For Model 0 ($\boldsymbol{H_0}$: no treatment effect), AIC = 417.035, and BIC = 420.738. For Model 1 ($\boldsymbol{H_1}$: treatment effect), AIC = 346.724, and BIC = 354.131. Thus, AIC and BIC both prefer Model 1. The R^2-statistics are now much higher than in Experiment 1: $R^2_{\text{McF}} = 0.174$, $R^2_{\text{N}} = 0.286$, $R^2_{\text{CS}} = 0.214$, and $R^2_{\text{T}} = 0.241$. Cramér's V now equals 0.481, which indicates a medium to large effect. For the *OR*, we find a point estimate of 8.176 ($p < 0.001$) and a 95% CI from 4.870 to 13.726. For the b, we find a point estimate of 2.101 and a 95% CI from 1.583 to 2.619. The 95% CI of the difference in pass rate now extends from 0.381 to 0.579, which is fairly far outside the [−0.10; 0.10] range. Although even in this kind of cases it is recommended to do follow-up experiments to examine the replicability, Experiment 2 provides initial evidence for a medium to large positive

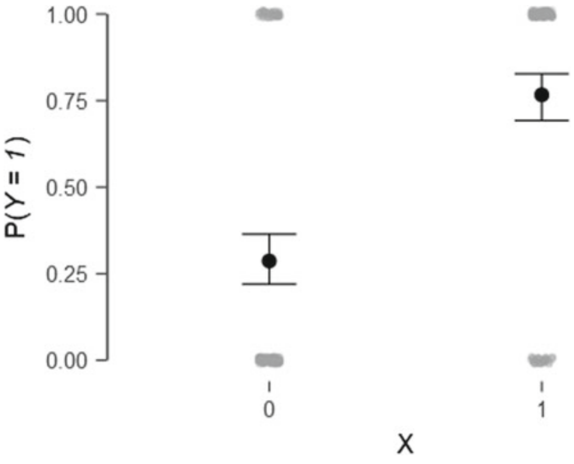


Fig. 5.2 *EMM* plot of Experiment 2: (observed) point and interval (95% CI) estimates of the probability of pass $P(Y = 1)$ for the control ($X = 0$) and treatment ($X = 1$) condition (JASP)

treatment effect. In FOST, we reject the relative equivalence region, and with such a strong finding from an initial experiment, one-sided testing anticipating a positive treatment effect would be defendable in a future experiment.

Experiment 3: Effect of Treatment on the Probability and Timing of Accidents

Suppose, a group of psychologists is interested in the effect of a type of hay fever treatment on the probability and timing of accidents when driving a car in a busy city. They randomly sample $N = 320$ people in the age group of 25–40 years who have experience with the hay fever treatment at hand but are not taking any other medications. These 320 participants are randomly allocated to either treatment ($n = 160$) or control ($n = 160$) condition. In the control condition, participants are instructed to drive a car in a driving simulator during 120 min unless an accident stops them. In the treatment condition, participants receive the same instructions but receive a standard recommended dosage of the treatment half an hour before the start of the driving session. In both conditions, the experiment stops either after 120 min or immediately after they are involved in an accident regardless of the severity of that accident. The literature with regard to possible effects of the treatment on driving behaviour is mixed, leaving the researchers with no direct reason to engage in one-sided testing. They are interested in differences between conditions in the *proportion* of accidents and *timing* of accidents. The occurrence of accidents is a dichotomous variable; participant A either has an accident or not.

The Proportion of Accidents

In the control condition, 10 of the $n = 160$ participants (6.25%) have an accident. In the experimental treatment condition, 32 of the $n = 160$ participants (20%) have an accident. Figure 5.3 (from JASP) depicts the *EMM* plot for the control condition (left: $X = 0$) and for the treatment condition (right: $X = 1$): the observed accident rates with 95% CIs around them.

There are 3.2 times as many accidents in the treatment condition than in the control condition. Using the LR test for $\boldsymbol{H_0}$ stating 'no treatment effect' against $\boldsymbol{H_1}$ stating 'treatment effect', we find $\chi^2_1 = 13.862$, $p < 0.001$. For Model 0 ($\boldsymbol{H_0}$: no treatment effect), AIC = 250.804, and BIC = 254.572. For Model 1 ($\boldsymbol{H_1}$: treatment effect), AIC = 238.942, and BIC = 246.479. Thus, AIC and BIC both prefer Model 1. The R^2-statistics are as follows: $R^2_{\text{McF}} = 0.056$, $R^2_{\text{N}} = 0.078$, $R^2_{\text{CS}} = 0.042$, and $R^2_{\text{T}} = 0.069$. Cramér's $V = 0.204$. The *OR* is 3.750 ($p < 0.001$) and the 95% CI ranges from 1.775 to 7.924. The 95% CI of the difference in accident rate (point estimate = 20 – 6.25 = 13.75%) is [0.065; 0.210], and the 90% CI is [0.077; 0.198]. It appears that we are dealing with a small to medium effect. Note that although the point estimate and CI are positive, positive numbers in this case indicate a negative treatment effect (i.e., higher accident rate in the treatment condition).

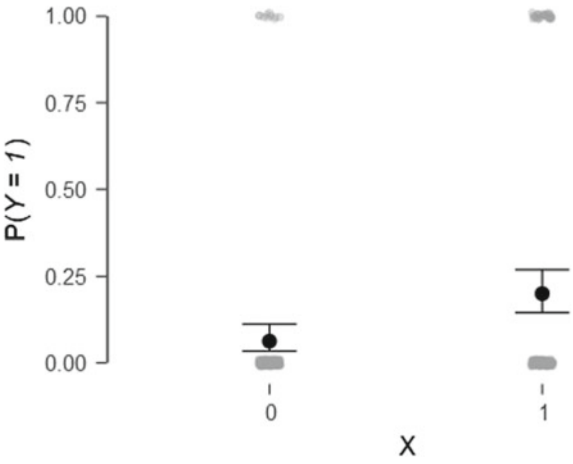


Fig. 5.3 *EMM* plot of Experiment 3: (observed) point and interval (95% CI) estimates of the probability of an accident $P(Y = 1)$ for the control ($X = 0$) and treatment ($X = 1$) condition (*JASP*)

Survival Analysis

Note that the statistics presented thus far do tell us something about the proportion of accidents but do not tell us anything about the occurrence of accidents over time. For the latter, we need *survival analysis*. Figure 5.4 displays the *survival curve* (Kaplan and Meier 1958) of the survival rates in the two conditions (‘0’ = control condition, upper curve; ‘1’ = treatment condition, lower curve) with their 95% CIs (*Jamovi*). The survival rate in condition X is found by dividing the number of participants surviving longer than time point t by the total number of participants in condition X.

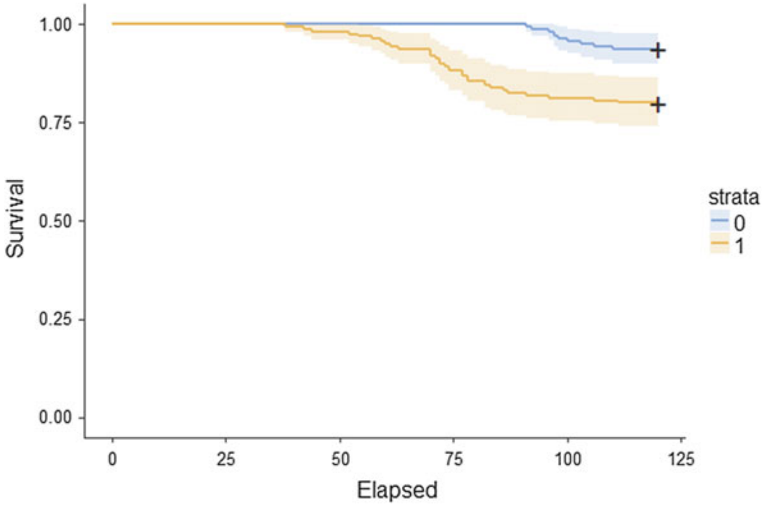


Fig. 5.4 Survival plot of Experiment 3: (observed) point and interval (95% CI) estimates of the proportion of survival for the control ($X = 0$) and treatment ($X = 1$) condition (*Jamovi*)

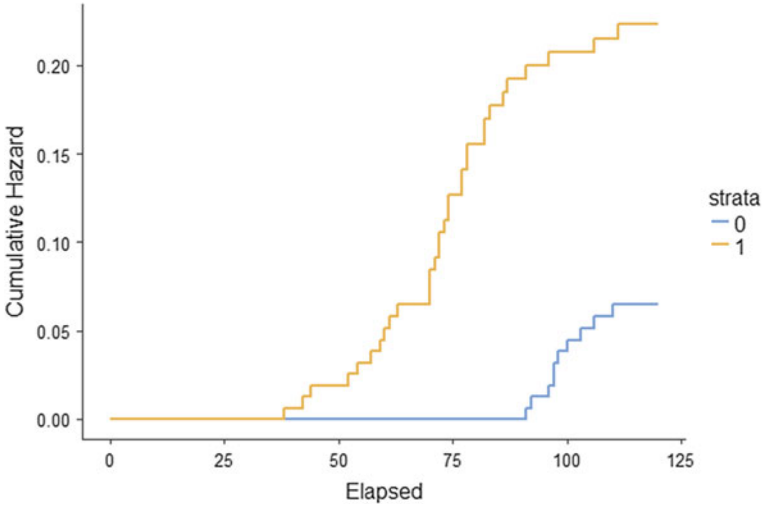


Fig. 5.5 Cumulative hazard function of Experiment 3 for the control ($X = 0$) and treatment ($X = 1$) condition (*Jamovi*)

Figure 5.4 indicates an earlier start of accidents (the horizontal axis indicates the time in minutes over the 0–120 min duration of the experiment) and a consistently lower survival rate in the treatment condition compared to the control condition. Another way to plot the difference between conditions is in terms of *cumulative hazard* (Peterson 1977). As long as the survival is 1 (100%), the cumulative hazard equals 0. Once we start to see accidents, the survival starts to go down from 1, while the cumulative hazard starts to go up from 0. The cumulative hazard, which is also known as the *conditional failure rate*, can be expressed as the event rate at time point *t* conditional on surviving up to or beyond time *t*. Figure 5.5 depicts the cumulative hazard function for Experiment 3.

Different tests can be used to test if the difference between conditions in this trajectory of accident occurrence is statistically significant at a given significance level. To start, there is the classical nonparametric *log-rank test* (Mantel 1966; Mantel and Haenszel 1959) which assumes that the hazard ratio of the two conditions is proportional across time (i.e., proportional hazard assumption). When the hazard ratio is not constant, as in Experiment 3, at least three alternatives based on the Wilcoxon *signed-rank test* (Wilcoxon 1945)—which for that reason are also called *generalised Wilcoxon tests*—provide alternatives to the classical log-rank test: the Gehan-Wilcoxon test (Gehan 1965), the Peto-Peto test (Peto and Peto 1972), and the Tarone-Ware test (Tarone and Ware 1977). These three alternatives differ from the classical test in that they deal differently with multiple accidents at a given time point *t* and give more weight to earlier accidents than later accidents (i.e., in the log-rank tests, all accidents have the same weight). The classical test is more efficient than its alternatives whenever the proportional hazard assumption is (more or less) met, whereas the alternatives are more efficient in the case of (substantial) departures from that assumption (Harrington and Fleming 1982). When in doubt, reporting all four tests is always an option. Doing so, in Experiment 3, we find (*Jamovi*): log-rank $z = 3.785$, Peto-Peto $z = 3.917$, Gehan-Wilcoxon $z = 3.917$, and Tarone-Ware $z = 3.852$; for all four tests, $p < 0.001$. In other words, all four tests provide evidence for the hypothesis that the two conditions differ in trajectory.

Additionally, we can take a closer look at the *distribution* of accidents over time —excluding the cases who had *no* accident at all—in the two conditions, as displayed in the histograms in Fig. 5.6 and the boxplots in Fig. 5.7 (*Jamovi*).

In the control condition, the *M* time of accident occurrence is 99.000 min and the *SD* is 5.944 (range: 91–110 min). In the treatment condition, the *M* time of accident occurrence is 72.188 and the *SD* is 16.851 (range: 38–111 min).

Welch's two-samples *t*-test for unequal *SD*s yields $t_{39.20} = -7.612$, $p < 0.001$, and Cohen's $d = -2.122$ with a 95% CI of [−3.682; −1.812] (*JASP*).

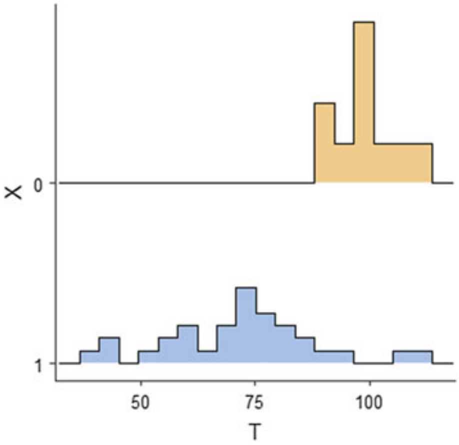


Fig. 5.6 Histogram of the distribution of accident times per condition in Experiment 3 ($X = 0$: control; $X = 1$: treatment) (*Jamovi*)

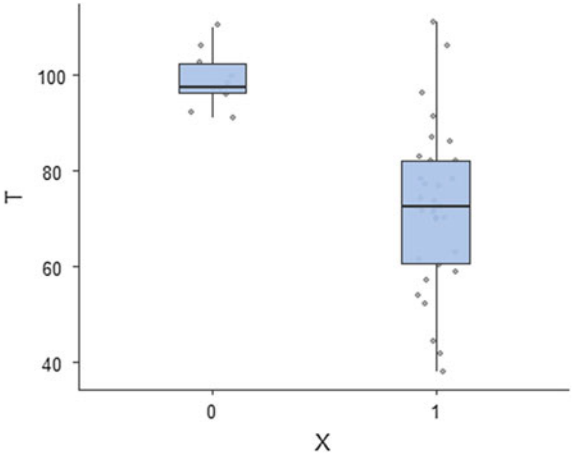


Fig. 5.7 Boxplot of the distribution of accident times per condition in Experiment 3 ($X = 0$: control; $X = 1$: treatment) (*Jamovi*)

Mann-Whitney's nonparametric U test (Fay and Proschan 2010; Mann and Whitney 1947) yields $U = 21.500$, $p < 0.001$, and a rank biserial correlation (Glass 1966; Willson 1976) of −0.866 with a 95% CI ranging from −0.939 to −0.717.

Note that an alternative approach is to compare the distribution of *all* participants including those who had no accident. However, the researchers in Experiment 3 are interested in the *timing* of *accidents*, and in that light a comparison of the conditions in terms of their distribution of *accidents* (i.e., excluding those cases that had no accident) makes more sense.

Final Note on Questions and Outcome

While all three example experiments presented in this chapter have in common that the key outcome variable is a dichotomous one, Experiment 3 differs from Experiments 1–2 in that the question(s) and nature of the outcome variable in Experiment 3 require researchers to inspect graphs and report statistics additional to the ones reported in Experiments 1–2. That is, contrary to Experiments 1–2, Experiment 3 involves a time variable. Extensions of Experiments 1–2 involving a time variable are found in post-tests that involve more than trial, be it multiple scenarios in Experiments 1–2 or multiple correctly/incorrectly responded items in a post-test. With such extensions (see also Chaps. 14–16 of this book), multilevel aka mixed-effects logistic models (Molenberghs and Verbeke 2005) can provide appropriate tests and estimates of the treatment effect of interest, and Rasch modelling may help to provide measurement reliability estimates as well.

6 Multicategory Nominal Outcome Variables

Abstract

The concepts discussed in Chap. 5 in the context of dichotomous outcome variables are also useful for multicategory nominal outcome variables. Although this type of outcome variable remains less common than dichotomous, ordinal (Chap. 7) or quantitative outcome variables (Chap. 8), it may constitute a main outcome variable in for example experiments that focus on choice behaviour or association as a function of treatment. For instance, in research on emotion or attitudes, different content or different formats of presenting content may trigger different types of emotions and stimulate different attitudes with it. These emotions and attitudes may at best be considered nominal rather than ordinal categories. Statistics for association as well as model performance statistics and plots for multicategory nominal outcome variables are discussed in this chapter.

Introduction

Suppose, some psychologists are interested in factors that influence European Union (EU) citizens' attitudes towards EU politics, because recent polls revealed that over 30% of the EU citizens are somewhat indifferent towards EU politics. The psychologists decide to write a research proposal for a series of experiments on factors that may influence EU citizens' attitudes towards EU politics for the better or for the worse. For reasons of feasibility, the researchers decide to carry out this first series of experiments in their own EU member state. They realise that this may limit the generalisability to their country rather than to the whole EU (even though the aforementioned percentages are quite accurate for their country), but they hope that researchers with similar interests in other EU member states will invest in similar experiments in their countries.

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_6

In each experiment, they randomly assign $N = 240$ citizens of age group 18–75 years old to either of two conditions. In the control condition ($n = 120$), participants see a 10-min YouTube video on a specific contemporary question in EU politics. In the treatment condition ($n = 120$), participants see a 10-min video that covers the same content as the video in the control condition but presents that content in a slightly different way. In both conditions, immediately after the video, participants are asked to choose which of four different words describes best how they feel about the EU after the video: *indifferent*, *embarrassed*, *surprised* or *disappointed*. Suppose that the psychologists have chosen these categories because previous research indicates that these four words are by far the most likely candidates in the context of the video content. In other experiments, which focus on other contents, different words constitute the categories to choose from.

Exchangeable Categories

Although the four categories represent different states, they cannot be ordered as in an ordinal or interval level of measurement variable; the categories are exchangeable. As in Chap. 5, the treatment variable is a dichotomous variable. However, the outcome variable now has four instead of two categories. Binary logistic regression, the core analytic method in Chap. 5, can no longer be used, at least not as a starter. However, there is an extension of binary logistic regression called *multinomial logistic regression*. Most of the plots, information criteria, as well as R^2-statistics and other effect size estimates discussed in the binary logistic regression context can also be used in multinomial logistic regression.

Estimates Per Condition

In the control condition, the choices are as follows: 44 participants indifferent (36.7%), 12 participants embarrassed (10.0%), 20 participants surprised (16.7%), and 44 participants disappointed (36.7%). In the treatment condition, we find: 23 participants indifferent (19.2%), 22 participants embarrassed (18.3%), 17 participants surprised (14.2%), and 58 participants disappointed (48.3%). Figure 6.1 presents an *EMM* plot (*Jamovi*), which displays the aforementioned observed proportions of participants per category for each condition ($X = 0$: control; $X = 1$: treatment) and 95% CIs around them. These *EMM* plots are what in some other software packages (e.g., *JASP*) are called *conditional estimates plots* (see Chap. 5).

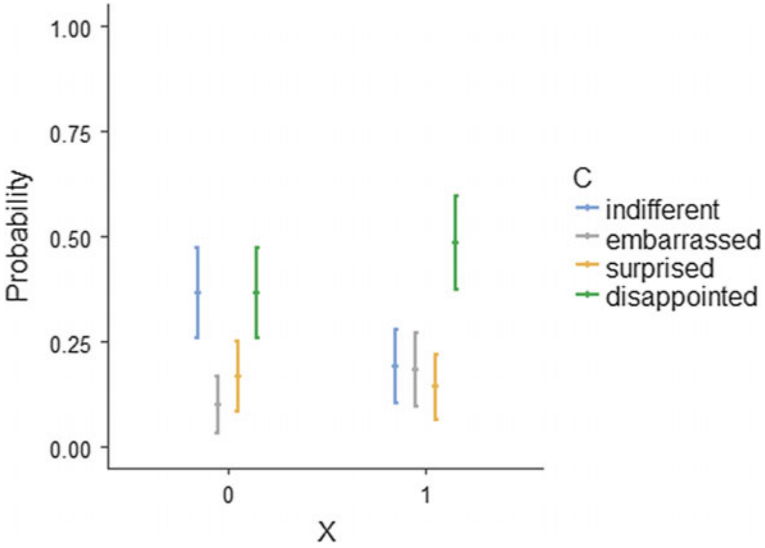


Fig. 6.1 *EMM* plot of the experiment: observed proportions per category per condition (X = 0: control; X = 1: treatment) and 95% CIs (*Jamovi*)

Multinomial Logistic Regression

Testing $\boldsymbol{H_0}$: 'no difference' against $\boldsymbol{H_1}$: 'difference' at $\alpha = 0.05$ with a LR test, we find $\chi^2_3 = 11.851, p = 0.008$. Note that the df is now 3 not 1. Given c columns and r rows, df can be computed as follows:

$$df = (c-1) * (r-1).$$

Therefore, the LR ratio test is now a test with $df = 3$, not $df = 1$. AIC is 622.783 for Model 0 ($\boldsymbol{H_0}$: no treatment effect) and 616.932 for Model 1 ($\boldsymbol{H_1}$: treatment effect). BIC is 633.225 for Model 0 and 637.816 for Model 1. In other words, BIC indicates a preference for Model 0, while AIC indicates a preference for Model 1. The BF for Model 1 versus Model 0 (BF_{10}) using the *JASP* default prior (based on Gunel & Dickey, 1974) is 3.292, indicating moderate evidence in favour of Model 1. In other words, although there is some disagreement, most criteria appear to hint at Model 1.

Statistical packages in which multinomial logistic regression analysis is included present R^2_{McF}, R^2_{CS}, R^2_{N}, or combinations thereof. For example, *SPSS* and *Jamovi* provide all three, while *Stata* (by default) provides R^2_{McF} (under pseudo R^2). However, interestingly, different packages provide different outcomes for R^2_{McF} as well as for R^2_{CS} and R^2_{N}. Let us do an exercise similar to that in Chap. 5 with different marginal probabilities (i.e., 0.50, and 0.80 casu quo 0.20). First, let us take the example experiment and take a hypothetical case where the two groups could be told apart *perfectly* in terms of which of four different emotions they choose. In the very extreme case that *all* participants in the control condition chose *one* specific emotion and *all* participants in the treatment condition chose *one other* specific emotion, we would effectively be back to the exercise in Chap. 5 with marginal probability 0.50, and we would find horrors with regard to R^2_{T} and R^2_{CS} (i.e., values well below 1). Now, suppose that in the control condition half of the participants would embrace emotion '1' and half of the participants would embrace emotion '2', while in the treatment condition half of the participants would embrace emotion '3' and half of the participants would embrace emotion '4'. We now calculate $R^2_{\mathrm{McF}}, R^2_{\mathrm{CS}},$ and R^2_{N} in different packages and find 'magic'. For R^2_{McF}, we find 1.000 in *SPSS* and 0.500 in *Stata* and *Jamovi*. *SPSS* computes R^2_{McF} to be 1.000 because −2LL of Model 0 is 55.452 while that of Model 1 is 0. In *Stata* and *Jamovi*, $R^2_{\mathrm{McF}} = 0.500$ because even though the conditions can be separated perfectly based on the emotions they embrace, within group there is still uncertainty, and hence, in that respect, the relation is not perfect. Consider an equivalent with a quantitative outcome variable where the range of scores in the control condition is 0-10 while that in the treatment condition is 15–25. The conditions can be told apart *perfectly* from one another in terms of score on the outcome variable, but within conditions there is still variation and hence R^2 would probably be large but would not be 1. In that respect, in the example at hand, $R^2_{\mathrm{McF}} = 0.500$ (i.e., −2LL of Model 1 being not 0 but being *half* of −2LL of Model 0, i.e., 55.452 for Model 1 and 110.904 for Model 0) makes more sense than $R^2_{\mathrm{McF}} = 1.000$. Likewise, *SPSS* reports $R^2_{\mathrm{CS}} = 0.750 R^2_{\mathrm{N}} = 0.652$ and $R^2_{\mathrm{N}} = 1.000$, while *Jamovi* reports $R^2_{\mathrm{CS}} = 0.293$ and $R^2_{\mathrm{N}} = 0.586$.

Now let us do the same with marginal probabilities of 0.20 and 0.80 for emotions '1' (4 of the 20 participants) and '2' (16 of the 20 participants) in the control condition and marginal probabilities of 0.20 and 0.80 for emotions '3' (4 of the 20 participants) and '4' (16 of the 20 participants) in the treatment condition. *SPSS* returns the same numbers as in the aforementioned 50%/50% example: $R^2_{\mathrm{McF}} = 1.000$, $R^2_{\mathrm{CS}} = 0.750$, and $R^2_{\mathrm{N}} = 1.000$. *Stata* and *Jamovi* return $R^2_{\mathrm{McF}} = 0.581$ (−2LL Model 0 = 40.032, and −2LL Model 1 = 95.484), and we find $R^2_{\mathrm{CS}} = 0.293$ and $R^2_{\mathrm{N}} = 0.652$. Again, given the heterogeneity within conditions, 0.581 and 0.652 make more sense than 1.000. The value 1.000 in SPSS considers only differences between conditions but not differences within conditions, while 0.581 and 0.652 account for both types of differences.

For the dataset at hand, where we deal with effects of a for educational and psychological research much more realistic size, *SPSS*, *Stata*, and *Jamovi* return the same outcome for R^2_{McF} : 0.019 (i.e., the outcome is the same on the first six decimals). However, *SPSS* returns $R^2_{\mathrm{CS}} = 0.048$ and $R^2_{\mathrm{N}} = 0.052$ while *Jamovi* returns $R^2_{\mathrm{CS}} = 0.012$ and $R^2_{\mathrm{N}} = 0.026$. Again, the difference in R^2_{CS} and R^2_{N} lies in *Jamovi* taking within-condition heterogeneity into account while the numbers in *SPSS* are based on between-condition heterogeneity only. The way R^2_{CS} and R^2_{N} are computed in *Jamovi* is consistent with our notions of R^2 for quantitative outcome variables, and consequently the numbers are situated somewhere around R^2_{McF} (i.e., in *Jamovi*) instead of quite a bit above R^2_{McF} (i.e., in *SPSS*).

As we recall from Chap. 5, a nice feature of R^2_{McF} is that even if it is not provided by a software package but we do have correct −2LL estimates for Model 0 and Model 1, we can compute R^2_{McF} from the *deviance* (−2LL) of Model 0 (null model, no treatment effect) and that of Model 1 (treatment effect):

$$R^2_{\mathrm{McF}} = 1 - ([-2\mathrm{LL}\ \underline{\mathrm{Model\ 1}}]/[-2\mathrm{LL}\ \underline{\mathrm{Model\ 0}}]).$$

In the experiment at hand, $2\mathrm{LL}\ \underline{\mathrm{Model\ 1}} = 604.932$, and $-2\mathrm{LL}\ \underline{\mathrm{Model\ 0}} = 616.783$ (i.e., recall from Chap. 2 that the difference between these two deviances is the χ^2—value of 11.851 used for the LR test), hence:

$$R^2_{\mathrm{McF}} = 1-(604.932/616.783) \approx 0.019.$$

Some readers may have noticed that another way of calculating R^2_{McF} is as follows:

$$R^2_{\mathrm{McF}} = \chi^2/[-2\mathrm{LL}\ \underline{\mathrm{Model\ 0}}]) = 11.851/616.783 \approx 0.019.$$

Cramér's V for this dataset is: $V = 0.221$. In other words, we appear to be dealing with a fairly small effect. Researchers who prefer to base model selection on BIC may prefer to stop here, point at the low R^2-statistics and fairly low Cramér's V, and leave it there. However, in my view, it is always better to use a combination of criteria, and small R^2- or V-values by themselves do not imply relative equivalence of treatment and control condition. This is even more so because model fit and effect size estimates for multicategory nominal outcome variables do not indicate *what* an effect looks like: are there small differences between conditions across categories of the outcome variable, or are some differences between conditions (i.e., for some of the outcome variable categories) larger than other differences?

Follow-Up Analysis (A): Simultaneous Estimation

From the criteria discussed thus far, we do not learn *what* an (eventual) effect looks like. Figure 6.1 gives us an initial idea. What we can do next, after the outcomes of statistical significance testing (p = 0.008) and AIC and BF_{10} (both in favour of Model 1), is a follow-up analysis that uses the information depicted in Fig. 6.1. If next to BIC, one or two of the other criteria—p-value, AIC or BF_{10}—also indicated no reason to go beyond Model 0, doing this follow-up analysis would not make sense. In that case, one would not expect meaningful differences in the follow-up analysis and any differences that did arise might well be false alarms.

The goal of the follow-up analysis in this kind of choice experiment is to acquire a more accurate picture of what a treatment effect looks like. However, we should

Table 6.1 Follow-up analysis of multinomial logistic regression: *b* point estimates and 95% CIs (95% LB and UB) for each of three comparisons (*Jamovi*, *Stata*)

Comparison	Term	*b* estimate	95% LB	95% UB
Embar.–Ind.	Intercept	−1.299	−1.938	−0.661
	Treat–Control	1.255	0.389	2.120
Surpr.–Ind.	Intercept	−0.788	−1.317	−0.260
	Treat–Control	0.486	−0.334	1.306
Disap.–Ind.	Intercept	≈0.000	−0.418	0.418
	Treat–Control	0.925	0.286	1.564

preferably do that follow-up analysis with the smallest number of statistical tests possible. Frequently, the questions and design already inform which tests we should carry out. In the case at hand, the psychologists designed the experiment from the recent observation that over 30% of the EU citizens are somewhat indifferent towards EU politics. Of the four words, *indifference* is probably the most *laissez faire* attitude towards politics; the other three words appear to indicate at least some kind of interest or emotional involvement. In this context, we can choose *indifferent* as a *reference category* and compare the proportions of the three other choices with those of indifference. With regard to the treatment effect of interest, this results in three statistical tests: *embarrassed* versus *indifferent*, *surprised* versus *indifferent*, and *disappointed* versus *indifferent*. Software packages like *Stata* and *Jamovi* provide an easy way to carry out this follow-up analysis: they provide *b* and *OR* with CI and statistical significance test for each of the aforementioned three comparisons. Table 6.1 presents the outcomes of this follow-up analysis as per *Stata* and *Jamovi*.

With regard to the treatment effect, we need the outcomes of the *Treat–Control* rows in Table 6.1. For all three comparisons, we find a positive *b*: 1.255 for embarrassment versus indifference (p = 0.004), 0.486 for surprise versus indifference (p = 0.245), and 0.925 for disappointment versus indifference (p = 0.005). Some may suggest to apply a correction for multiple testing, for instance a Bonferroni correction (see also Chaps. 2 and 9), since we are testing three times at $\alpha = 0.05$. Others may argue that the comparisons arise naturally from the origin of the experiment and moreover all *b*s are estimated simultaneously and logically relate to each other and therefore such a correction is not needed per se. That said, whether we apply a Bonferroni correction or not, two of the three comparisons yield a statistically significant outcome: embarrassment versus indifference, and disappointment versus indifference.

Now, what do we mean by *b*s logically relating to each other? If we had no natural reference category, we might as well list all possible comparisons of pairs of categories C_p:

$$C_p = [k * (k{-}1)]/2.$$

For $k = 3, C_p = 3;$ for $k = 4, C_p = 6;$ for $k = 5, C_p = 10$. In the presence of a natural reference category, the number of natural comparisons C_n (i.e., the comparisons that make sense given the reference category) equals:

$$C_n = k-1.$$

Given *indifference* as a reference category in the experiment of this chapter, we only list the three comparisons (i.e., $k-1 = 4-1 = 3$) that involve anything else versus *indifference*. However, we can compute the outcomes for the other possible comparisons using the numbers presented in Table 6.1. To start, the *Treat–Control b* for *embarrassed* versus *indifferent* (B_{EI}) equals 1.255, and that for *surprised* versus *indifferent* (B_{SI}) is 0.486. From this, the *b* for *embarrassed* versus *surprised* (B_{ES}) follows naturally:

$$B_{\mathrm{ES}} = B_{\mathrm{EI}} - B_{\mathrm{SI}} = 1.255 - 0.486 = 0.769.$$

Likewise, the *b* for *embarrassed* versus *disappointed* (B_{ED}) can be calculated from the numbers in Table 6.1 as well:

$$B_{\mathrm{ED}} = B_{\mathrm{EI}} - B_{\mathrm{DI}} = 1.255 - 0.925 = 0.330.$$

To complete the exercise, the *b* for *surprised* versus *disappointed*, following the numbers from Table 6.1, is:

$$B_{\mathrm{SD}} = B_{\mathrm{SI}} - B_{\mathrm{DI}} = 0.486 - 0.925 = -0.439.$$

Note that the latter also directly follows from the previous two:

$$\begin{aligned}
&B_{\mathrm{SE}} = -B_{\mathrm{ES}}, \text{ and} \\
&B_{\mathrm{DE}} = -B_{\mathrm{ED}}, \text{ and hence,} \\
&B_{\mathrm{SD}} = B_{\mathrm{SE}} - B_{\mathrm{DE}} = -0.769 - (-0.330) = -0.769 + 0.330 = -0.439.
\end{aligned}$$

To understand what these outcomes mean, we should also check again Fig. 6.1. For embarrassment, we find a slightly lower proportion than for indifference in the treatment condition but a much lower proportion than for indifference in the control condition. This difference is statistically significant. For surprise, we find a slightly lower proportion than for indifference in the treatment condition but a considerably lower proportion than for indifference in the control condition. This difference, however, is not statistically significant. Finally, for disappointment, we find a somewhat higher proportion than for indifference in the treatment condition but the same proportion (36.7%) in the control condition. This difference is statistically significant.

Figure 6.1 and the outcomes of Table 6.1 appear to indicate that the two conditions differ in terms of two balances: embarrassment versus indifference and disappointment versus indifference; the treatment under study may create (or contribute to) some shift in these two balances from indifference towards embarrassment (for some citizens) or disappointment (for some other citizens).

Follow-Up Analysis (B): Differences in Proportions Per Outcome Variable Category

The aforementioned follow-up analysis makes sense in the light of the origin of the experiment and the omnibus tests ($p = 0.008$, and AIC and BF_{10} in favour of Model 1). A second possible follow-up analysis is found in a direct comparison of the conditions in terms of proportions for each of the categories of the outcome variable. This approach provides straightforward testing and estimation outcomes for condition difference per category of the outcome variable, instead of for differences in differences (i.e., the aforementioned approach is about differences between conditions in differences between categories of the outcome variable). If we are to apply a correction for multiple testing, which probably researchers would argue for, this correction will be slightly stronger than in the aforementioned approach, since we have one more test to carry out. Using the α_{total}-formula discussed in Chap. 2, for three tests $\alpha_{\mathrm{total}} \approx 0.1426$ while for four tests we find $\alpha_{\mathrm{total}} \approx 0.1855$. The latter would come down to a corrected alpha of about 0.0127 per test to keep α_{total} at 0.05 (compared to about 0.0169 in the case of three tests).

Table 6.2 presents 90% CIs (default uncorrected intervals for two one-sided tests equivalence testing; e.g., Lakens, 2017), 95% CIs (default uncorrected intervals for two-sided testing of a ‘no difference’ null hypothesis), and 99% CIs (i.e., slightly more conservative than the corrected alpha) for the difference in proportion between conditions for each of the four categories of the outcome variable.

In Table 6.2, positive differences are indicative of higher proportions in the treatment condition compared to the control condition. The outcomes of Table 6.2 indicate that whether we test two-sided at $\alpha = 0.05$ or at $\alpha = 0.01$, the only difference that is significantly different from zero is that of *indifference*: the proportion of indifference is significantly *lower* in the *treatment* condition than in the control condition (see Fig. 6.1). However, this follow-up analysis does not really help us to acquire a somewhat more accurate picture of how an eventual reduction of a tendency towards indifference is reflected one or more other categories.

Table 6.2 Alternative follow-up analysis of multinomial logistic regression: 90, 95, and 99% CIs (LB and UB) for the condition difference in proportion per category of the outcome variable (*Jamovi*)

Category	90% LB	90% UB	95% LB	95% UB	99% LB	99% UB
Ind.	−0.268	−0.082	−0.286	−0.064	−0.321	−0.029
Embar.	0.010	0.157	−0.004	0.171	−0.032	0.198
Surpr.	−0.102	0.052	−0.116	0.066	−0.145	0.095
Disap.	0.012	0.221	−0.008	0.241	−0.047	0.280

Follow-Up Analysis (C): Outcome Variable Distribution Per Condition

Having a closer look at Fig. 6.1, some readers may wonder if there is not a third possible follow-up analysis: to compare the distribution of proportions for each of the two conditions separately. Most software packages on the market, including free ones such as *JASP* and *Jamovi*, include a *multinomial test* that allows researchers to test whether an observed distribution differs significantly from an expected distribution, be it equal proportions or unequal proportions based on previous research. In this third follow-up analysis, that multinomial test would then be done *per condition*. However, just like omnibus tests in a multinomial logistic regression do not tell us *what* a treatment effect looks like, an omnibus multinomial test per condition tells us only *if* there is *any difference* but not where these differences are. For the latter, we would need additional *binomial tests* for local differences. With such an approach, we run into serious multiple testing problems; we will either face a heavy inflation of Type I error probability or need to apply a severe correction to counter that inflation, and the latter comes with a considerable loss of statistical power. Moreover, this approach—which focusses on differences *within* conditions —tells us little if anything meaningful about the *treatment effect* of interest which gave rise to the experiment in the first place.

A Pragmatic Approach to Multicategory Nominal Outcome Variables

This chapter started with the statement that binary logistic regression (the core method in Chap. 5) cannot provide an omnibus test for a treatment effect of interest when the outcome variable of interest consists of multiple instead of two categories. Instead, we start with multinomial logistic regression. If the multinomial logistic regression analysis yields insufficient evidence to go beyond Model 0 (i.e., we fail to reject the null hypothesis of 'no treatment effect'), there is no reason to perform any follow-up analysis. However, if the multinomial logistic regression analysis does yield such evidence to go beyond Model 0, some follow-up analysis is needed to gain a deeper understanding of the treatment effect of interest. Given that the second approach to follow-up analysis discussed in this chapter (i.e., proportion differences between conditions per category of the outcome variable) involves only a *subset* of the data used in the first approach (i.e., simultaneously estimated *b*s), the first approach has at least three advantages over the second approach.

To start, using only a subset of the data, where using a full set of data is possible, comes with an unnecessary loss of statistical power and precision. Besides, while there is a direct logical connection between the different *b*s in the first approach, that relation is missing in the second approach. That is, although given the *b*s for all comparisons involving a given reference category (here: *indifferent*) all other *b*s are known, we cannot calculate the difference in proportions between conditions in any of the other categories just by knowing the difference in proportion in a reference category (e.g., the difference in proportion for *surprised* is not yet known given the difference in proportion for *indifferent*). Finally, the first approach comes with the advantage that—based on theory, previous research, and the data at hand—specific *equality constraints* (i.e., fixing two *b*s as equal) can be applied. This is especially useful when outcome variables include many possible categories; a gain in *df* resulting from equality constraints may in such cases greatly facilitate the estimation of other parameters.

The only thing we learn from the second (i.e., proportion differences) follow-up approach is that there is a statistically significant difference between conditions in the proportion of *indifference*. Proponents of the first approach may argue that this specific outcome of the second approach adds weight to the interpretation of Fig. 6.1 and Table 6.1 in terms of a shift in a certain tendency of difference towards more embarrassment and disappointment, because the difference in proportion of indifference (Table 6.2) is significantly lower in the treatment condition. It is to be noted, however, that this specific difference is rather strong; differences of just a bit less of a magnitude would not be statistically significant. The category-specific follow-up approach comes with a substantial loss of information and statistical power and may in more than a few cases result in none of the categories yielding a statistically significant difference even if there are substantial differences and the *b*-approach detects (some of) these differences. Therefore, the category-specific (i.e., second) approach might be used *in addition to* the *b* (i.e., first) approach—especially in cases where the outcomes of the *b*-model appear somewhat more difficult to understand—but *should not be used instead* of the *b*-approach.

Final Note on Questions and Outcome

For the ease of introduction, the experiment discussed in this chapter has a multicategory nominal outcome variable that is measured once in time. When a time variable is involved, or when several questions with multiple nominal level choices are involved, we need more complex models. Depending on the questions driving the experiment, the nature of the choice variables and data acquired, and the sample size at hand, one may have one or several alternatives to the simple model discussed in this chapter. However, all alternatives have in common that they require much larger samples than the sample in the experiment discussed in this chapter. In fact, even for simple experiments like this, sample size is a tricky issue. The smaller the sample, the higher the risk of ‘zero’ cells, contingency table cells that remain with 0 observations. The occurrence of ‘zero’ cells can seriously affect outcomes of testing and estimation with regard to effects of interest, and in some cases the testing and estimation of an effect of interest may not even be possible. Therefore, when small proportions are expected in at least some of the cells—because some phenomena naturally occur rather rarely or because there are many competing categories in an experiment—we will need larger samples for otherwise we may not be able to (accurately) test and estimate our treatment effect of interest. For the experiment in this study, two conditions with n = 120 participants each works fairly well, because the smallest proportion observed in a condition is 10% (in the control condition, n = 12 for 'embarrassed'). However, even in this kind of experiments more is better; with small numbers, the difference between one more or one less in Cell B may constitute a more than trivial change in *OR*s in which the number of Cell B is part of the input.

That said, when the same choice question is asked repeatedly or different choice questions are presented at the same occasion (e.g., in the same experiment), and we have access to much larger samples (e.g., choice experiments done online), we may have several options. A first option may be multilevel aka mixed-effects multinomial logistic regression models (e.g., Dey, Raheem, & Lu, 2016; Hedeker, 2003). This kind of mixed-effects models applies the same logic as mixed-effects models for other types of outcome variables: participant (level 2, upper level) and question, item or occasion (level 1, lower level) constitute the hierarchical levels. Other approaches, which involve latent variables (e.g., for a comparison of non-latent and latent-variable models, see: Hox, Moerbeek, & Van de Schoot, 2017), can be found in generalised item response theory (IRT) models (e.g., Jeon & De Boeck, 2016), the multidimensional random coefficient multinomial logit model (Adams, Wilson, & Wang, 1997; Briggs & Wilson, 2004), latent class analysis, and mixed Rasch models.

Ordinal Outcome Variables

7

Abstract

A commonly undervalued and mistreated type of outcome variable is the ordinal one. Two common types of mistreatment are treating ordinal variables as interval/ratio level outcome variables (frequently in linear models) and, in other cases, as multicategory nominal outcome variables. Multicategory nominal outcome variables are covered in Chap. 6 and quantitative outcome variables in Chap. 8 of this book. What these two types of mistreatment have in common is that they more often than not may result in outcomes that do not make sense. Where we treat ordinal variables as if they were (at least) interval, we may see linear relations where they do not make sense. Where we treat ordinal variables as if they were nominal, we treat all categories as exchangeable and lose the information with regard to the order of categories and, to a large extent, the meaning of outcomes with it. Although many of the concepts discussed in the context of dichotomous (Chap. 5) and multicategory nominal outcome variables also have their use, we need to take an additional step to respect the ordinality information when dealing with ordinal outcome variables. Differences between approaches to ordinal data encountered in the literature are discussed first in terms of animal comparisons (mice, hedgehogs, cats, bears, and elephants) and then in the form two experiments each of which indicates a different type of treatment effect.

Introduction

Once upon a time, there were five healthy adult animals: a *bear*, a *cat*, an *elephant*, a *hedgehog*, and a *mouse*. This may constitute one way of ordering the animals, in alphabetical order, yet the order depends on the language we use. In Spanish, for example, the alphabetical ordering would be as follows: *elefante*, *erizo*, *gato*, *oso*,

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_7

ratón (elephant, hedgehog, cat, bear, mouse). In other words, this kind of sorting is *nominal*: although the animals can be distinguished in terms of their labels in either language, alphabetic ordering does not imply some kind of increase in size, weight or ability to run, and people who prefer different languages may come up with different orders. It is a bit like the states discussed in Chap. 6: indifferent, embarrassed, surprised, disappointed. Is there one natural way of ordering these states in terms of valence or intensity? Likely not, so we appear to be dealing with nominal categories.

If we decide to order the adult animals in terms of their weight, it appears we are able to rank them in only one way from light to heavy: *mouse* (likely in the 3–35 g range depending on the type), *hedgehog* (likely in the range of 0.25–1.25 kg depending on the type), *cat* (likely in the 3–5 kg range), *bear* (likely in the 60–600 kg depending on the type), and *elephant* (several 1,000s of kilos). Although there is variation in weight among each of the five animal populations, it is safe to say that a fat mouse will weigh less than a lightweight hedgehog, that a heavy hedgehog will be lighter than a meagre cat, that a round cat will weigh less than a hungry bear, and that a lazy fat bear has quite a few kilograms less than an elephant in the lowest weight category. Relative to one another, we could order the five animals in terms of their weight categories as follows:

very little (mouse)
little (hedgehog)
neither little nor much (cat)
much (bear)
very much (elephant)

We could safely do the same for these five healthy animals in terms of their height or size:

very small (mouse)
small (hedgehog)
neither small nor big (cat)
big (bear)
very big (elephant)

One might also argue that this ordering fits either of the items *This animal is heavy* and *This animal is big*:

strongly disagree (mouse)
disagree (hedgehog)
neutral (cat)
agree (bear)
strongly agree (elephant)

Let us now assign values to these five animals based on the aforementioned ordering:

1 (*mouse*) – 2 (*hedgehog*) – 3 (*cat*) – 4 (*bear*) – 5 (*elephant*)

Next, we ask 100 random people spread out over a big zoo that hosts these five animals—in smaller and bigger sizes, in slimmer and rounder shapes—to respond to this five-point scale with the first animal they encounter. This yields $N = 100$ single responses, and the outcomes (i.e., frequencies of responses per category) are as follows:

mouse (1, *strongly disagree*): 3
hedgehog (2, *disagree*): 11
cat (3, *neutral*): 35
bear (4, *agree*): 35
elephant (5, *strongly agree*): 16

We now calculate the *M* and *SD* of this random sample from this population of way over 5,000 people in the zoo, and we find: $M = 3.50$, $SD = 0.99$. Let us also calculate the skewness (*S*) and kurtosis (*K*) of the distribution, to examine how much the observed distribution deviates from a Normal one: $S = -0.32$, $K = -0.21$. In other words, the average response is a *bearish cat* or a *cattish bear*, one *SD* down we find the *hedgehogcat* or *cathedgehog* and one *SD* up we find the *bearelephant* or *elephantbear*. The negative skewness indicates a slight skew to the left (mind though that the $n = 3$ that responded 'mouse' are not outliers!), and the negative kurtosis indicates that the observed distribution has somewhat fatter tails than a Normal distribution. The median response is *bear* and the interquartile range (IQR), which extends from percentiles 25 (first quarter, Q1, *cat*) to 75 (third quarter, Q3, *bear*), is |*cat-bear*|.

Relational Meaning

At this point, you may be wondering what I was smoking when I was writing this chapter. And that would be a legitimate question. Note, however, that this is how we have been treating categorical rating data for many decades, including myself for the time I have been around. We are often told it is okay to treat categorical rating data as *interval* level of measurement data, that we can safely calculate *M*s and *SD*s, and can comfortably compute linear correlations between scores of different items. Although there may be cases where that is indeed the case, even in such cases it is probably more appropriate to treat the data as *ordinal*. In the world of truly ordinal data, *M*s, *SD*s, and Pearson's *r* (which uses *M*s and *SD*s as input) tend to make less sense than when dealing with interval or ratio level data.

Calculating *M*s, *SD*s, and Pearson's *r*, we assume that we can reasonably call the ordered categories *equidistant*. Consider a river with heavy stones that pop up such that you could step on them to get from one side to the other side of a river in a straight line without getting your feet wet; if the distance between the stones is such that every subsequent step you take is (about) the same distance, it is fair to treat the stones as (more or less) equidistant. If the distance between stones varies considerably, taking equidistant steps we would get our feet wet. The same holds for *M*s and *SD*s, and statistics that are based on them, such as linear correlations: when there are (substantial) departures from equidistance, they fall in the water and lose their meaning. Adding numerical labels (e.g., '1' … '5') to a scale does not mean we get rid of departures from equidistance. Whether we take *weight* or *height* (*size*) of the five animals, we cannot draw a straight line connecting all animals in terms equidistant increases. The differences between animals in terms of weight and height are non-linear by nature; differences in weight or height between an average hedgehog and an average cat are of completely different orders than differences between an average cat and an average bear, et cetera. Elephant minus cat does not equal hedgehog, and mouse plus cat does not equal bear either. Besides, even within animal groups, there is quite some fluctuation, just like the difference between *disagree* and *neutral* in a questionnaire may be quite distinct for different items. Finally, a multinomial approach does not help either because it ignores the information that hedgehogs are heavier and taller than mice, cats are heavier and taller than hedgehogs, et cetera; in a multinomial model, mice might as well be heavier than elephants and the height of a hedgehog might well exceed that of a bear.

In Chap. 3, I briefly discuss an example taken from Tacq and Nassiri (2011), who take the work from Pierre Bourdieu as an example of the nature of ordinal variables: distinctions in terms of dominant class, middle class, and working class are based on a relational logic (i.e., power relations, financial means) but not in terms of 1–2–3 equidistant categories. The same holds for the animals, and probably for most if not all other categorical ratings, whether the variable that underlies the ratings is weight or height (in the case of the animals), motivation or effort (in quite some educational and psychological research), financial means (often broader categories rather than single equidistant values), or something else. As discussed by Tacq and Nassiri (2011) and in Chap. 3 of this book, using coefficients such as Spearman's ρ does not resolve the problem of non-equidistance but at best pushes the problem forward to another level. In Spearman's ρ, the ranks are treated as if they were equidistant, but pasting numbers like '1', '2', '3', '4', and '5' to ranks does not magically turn an ordinal variable into an interval one. This is why Pearson's *r* and Spearman's ρ yielding almost the same outcomes on some categorical rating data cannot be understood as a justification for treating ordinal variables as interval ones (see also Chap. 3).

Ordinal variables come with information that is not present in nominal variables: a natural and consistent order of categories (i.e., strongly disagree, disagree, neutral, agree, strongly agree; or: mouse, hedgehog, cat, bear, elephant). If categories do not come in a consistent order, we are back to nominal variables. Moreover, depending on how the categories are formulated, there may be a risk that categories cannot be clearly separated. This may be a problem especially for items that include a perhaps excessive number of categories, such as the mental effort item discussed in Chap. 3: "*In solving or studying the preceding problem I invested: 1. very, very low mental effort; 2. very low mental effort; 3. low mental effort; 4. rather low mental effort; 5. neither low nor high mental effort; 6. rather high mental effort; 7. high mental effort; 8. very high mental effort; 9. very, very high mental effort.*" Respondents are supposed to choose the categories that applies to them for the activity that just finished. However, can a respondent distinguish between *very, very low* and *very low*, between *low* and *rather low*, between *rather high* and *high*, between *high* and *very high*, and between *very high* and *very, very high*? Surprisingly, although more or less since its introduction in 1992 the mental effort item has constituted the dominant method of cognitive load measurement (Sweller, 2018), this question remains untested. Yet, the answer to this question is *crucial* to determine the (most likely) level of measurement and can be tested with, among others, item response theory models including the Mokken model and variants of the Rasch model called the *rating scale model* and the *partial credit model*.

The procedure can be fairly straightforward: to administer the mental effort item with a group of $N > 300$ participants (to obtain good estimates and reduce the likelihood of zero cells) under the same circumstances a few (e.g., 5–10) times in a row, that is: with tasks and participants that could guarantee sufficient variation in response and the tasks being as similar as possible in terms of content and difficulty. If participants *can* make the distinction and the aforementioned psychometric methods provide reasonable support for the equidistance assumption, treating mental effort data as interval may make sense. However, if the categories can be distinguished in a consistent order but psychometric analysis does *not* provide support for the equidistance assumption, the different categories in ordinal variables likely only have a *relational meaning*. In the latter case, they cannot be summed, subtracted, averaged, or captured in neat linear equations. Indeed, ordinal data analysis is supposed to be non-linear. Furthermore, there is a third and perhaps a fourth option. Until empirically demonstrated otherwise, a possible outcome of the proposed study is that the order of the nine categories of the mental effort item is not consistent but deviates in at least one of the administrations. If so, the categories constitute *a multicategory nominal variable at best*. At best, because in the case of such a shift, we may find evidence for a phenomenon that would degrade the mental effort to *not even nominal*: respondents who actually differ in their mental effort provide the same response and/or respondents with the same mental effort respond differently.

Experiment 1: Different Qualities of Action

Some pharmacy education researchers are investigating the effectiveness of different approaches to skills training with simulated patients. In one of their experiments, they want to study the effect of providing a particular type of additional instruction prior to seeing a simulated patient. They do so because they are searching for ways to increase students' performance during this type of simulation training, but are not sure if the type of instruction they have in mind likely affects that performance positively or negatively (i.e., $\boldsymbol{H_0}$: 'neither positive nor negative effect'; $\boldsymbol{H_1}$: 'positive or negative effect').

They randomly sample $N = 160$ undergraduate pharmacy students from different universities in the region and randomly allocate them to treatment (n = 80) or control (n = 80) condition. The only way in which the two conditions differ is that participants in the treatment condition receive the type of instruction prior to seeing the simulated patient in addition to the usual instruction (i.e., the instruction that is *always* provided in this type of training), whereas in the control condition participants only receive the usual instruction. The simulated patient is the same for the two conditions. In both conditions, sessions are video-recorded and are evaluated by a team of pharmacy educators from the different universities sampled from who are blind to whether a participant received treatment or not and who are not aware of possible effects of the treatment on the outcome they are coding. For each participant, the team members discuss and come to consensus with regard to the performance: *appropriate* performance (label: '2'), *inappropriate* performance without damaging the simulated patient (label: '1'), and *inappropriate performance with damage* to the simulated patient (label: '0'). Note that the labels '0', '1', and '2' are just codes, not pretended interval level of measurement values. These three categories are to be treated as *ordinal*, with damaging the simulated patient representing the worst outcome, appropriate performance the best outcome, and inappropriate action without damage being somewhere in between. The coding and order of categories is in line with standard practice for assessing performance in this type of simulation training and for providing feedback after the training.

Towards Higher-Quality Performance

In the control condition (n = 80), 19 participants (23.8%) are coded '0' (damage), 28 participants (35.0%) are coded '1' (inappropriate but no damage), and 33 participants (41.3%) are coded '2' (appropriate). In the treatment condition (n = 80), 13 participants (16.3%) are coded '0', 20 participants (25.0%) are coded '1', and 47 participants (58.8%) are coded '2'. Hence, we observe slightly fewer '0' and '1' but somewhat more '2' codes in the treatment condition than in the control condition. Figure 7.1 presents the aforementioned percentages with their 95% CIs.

Testing $\boldsymbol{H_0}$: 'no effect' against $\boldsymbol{H_1}$: 'effect' at $\alpha = 0.05$ with a LR test, we find χ^2_2 = 4.536, $p = 0.033$. The *df* equals two, because we have one dichotomous variable

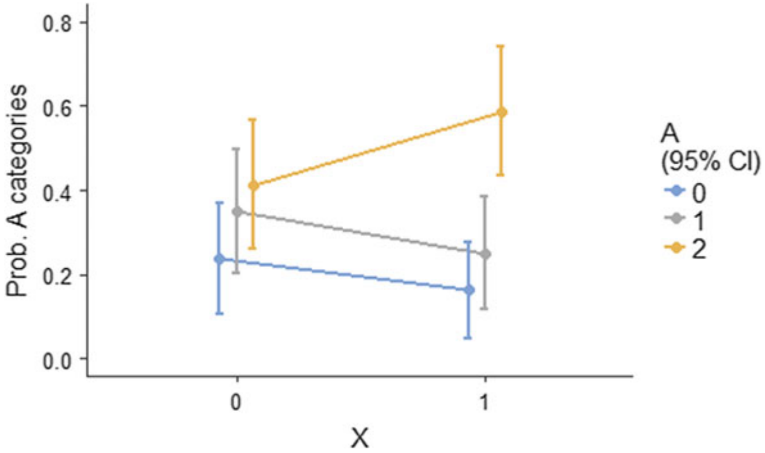


Fig. 7.1 Effects plot aka *EMM* plot aka conditional estimates plot of the findings in Experiment 1: observed proportions per category (A = 0: inappropriate with damage; A = 1: inappropriate without damage; A = 2: appropriate) per condition (X = 0: control; X = 1: treatment) and 95% CIs (*Jamovi*)

and one three-category variable. The χ^2-value of 4.536 is the difference in deviance (−2LL) of Model 1 (324.953) and Model 0 (324.953 + 4.536 = 329.489). Thus, we can now calculate R^2_{McF} (see also Chaps. 5 and 6):

$$R^2_{\text{McF}} = 1-([-2\text{LL}\,\underline{\text{Model 1}}]/[-2\text{LL}\,\underline{\text{Model 0}}]).$$

In Experiment 1, we find:

$$R^2_{\text{McF}} = 1-(324.953/329.489) \approx 0.014.$$

AIC is 333.489 for Model 0 ($\boldsymbol{H_0}$: no treatment effect) and 330.953 for Model 1 ($\boldsymbol{H_1}$: treatment effect). BIC is 339.639 for Model 0 and 340.178 for Model 1. *JASP* provides BFs for Kendall's τ coefficient (b; Kendall, 1962, in this Experiment, τ = 0.159, p = 0.034, 95% CI = [0.077; 0.242]). Using a default beta prior with width 1 (*JASP*; Ly, Verhagen, & Wagenmakers, 2016; Rouder & Morey, 2012; Wetzels & Wagenmakers, 2012), we find BF_{10} = 8.801. In other words, BIC indicates a slight preference towards Model 0, but the other criteria ($p < 0.05$, $\text{AIC}_{\text{Model1}} < \text{AIC}_{\text{Model0}}$, BF_{10} = 8.801) indicate a preference for Model 1.

Proportional Odds

Figure 7.1 graphically depicts that there are (proportionally) fewer 0s and 1s and (consequently) more 2s in the treatment condition. From the frequencies or proportions per category (0, 1, 2) per condition we can calculate *OR*s. One type of *OR* we can calculate is for each pair of categories, as in the multinomial logistic regression approach in Chap. 6. Even though in Chap. 6 I convert *b*s, we can also convert *OR*s. If we take the lowest category, 0, as reference category and calculate the *OR* of category 1 versus category 0 for the two conditions, we find:

$$OR_{1\,\text{versus}\,0} = \text{odds of '1' for treatment/odds of '1' for control} = (20/13)/(28/19) \approx 1.044.$$

Likewise, for *OR* of category 2 versus category 0 for the two conditions, we find:

$$OR_{2\,\text{versus}\,0} = \text{odds of '2' for treatment/odds of '2' for control} = (47/13)/(33/19) \approx 2.082.$$

Now, given only three categories, *OR* for category 2 versus category 1 for the two conditions follows from $OR_{2\ \text{versus}\ 0}$ and $OR_{1\ \text{versus}\ 0}$:

$$OR_{2\,\text{versus}\,1} = OR_{2\,\text{versus}\,0}/OR_{1\,\text{versus}\,0} \approx 1.994.$$

In other words, all *OR*s are in favour of the treatment condition. Although there are *both* fewer 0s *and* fewer 1s in the treatment condition than in the control condition, the *odds* of being in '1' instead of '0' are higher in the treatment condition than in the control condition.

A second type of *OR* we can compute when dealing with ordinal outcome variables is that of one category versus all the others. For a trichotomous (i.e., three-category) outcome variable, that comes down to two *ORs*, namely *OR* of being *at least* in category 1 (i.e., 1 or 2) versus being in category 0 ($OR_{12\ \text{versus}\ 0}$) and *OR* of being in the highest category (2) versus being in one of the lower categories ($OR_{2\ \text{versus}\ 10}$):

$$OR_{12\,\text{versus}\,0} = (67/13)/(61/19) \approx 1.605, \text{and}$$
$$OR_{2\,\text{versus}\,10} = (47/33)/(33/47) \approx 2.028.$$

Contrary to the aforementioned simple *OR*s, $OR_{12\ \text{versus}\ 0}$ and $OR_{2\ \text{versus}\ 10}$ are so-called *cumulative ORs*: they are ratios of *cumulative odds* or *cumulative probabilities* (note though that odds themselves are not probabilities!), because probabilities of different categories are accumulated here (Agresti, 2002; McCullagh, 1980). When we take the natural logarithm (ln) from an *OR*, we obtain *b*, which can be interpreted as a regression coefficient. In Experiment 1, we find $b_{12\ \text{versus}\ 0} \approx 0.473$ and $b_{2\ \text{versus}\ 10} \approx 0.707$. Especially when dealing with smaller sample sizes, to avoid zeros and to reduce bias some recommend to add 0.5 to all numerators and denominators in the above formulas of $OR_{12\ \text{versus}\ 0}$ and $OR_{2\ \text{versus}\ 10}$. Doing so in Experiment 1, we find

$$OR_{12\,\text{versus}\,0} = (67.5/13.5)/(61.5/19.5) \approx 1.585, \text{ and}$$
$$OR_{2\,\text{versus}\,10} = (47.5/33.5)/(33.5/47.5) \approx 2.010.$$

For the adjusted logits, we then find $b_{12 \text{ versus } 0} \approx 0.461$ and $b_{2 \text{ versus } 10} \approx 0.698$. Either way, from a practical standpoint, the logits are in the same direction and of a similar magnitude. In such cases, it makes sense to calculate and interpret one overall *OR* (or log *OR*) for the treatment effect of interest (OR_{treat}); this model is also known as the *proportional odds model* (e.g., Agresti, 2002; McCullagh, 1980) and is the default ordinal logistic regression model estimated by *Stata*, *SPSS*, and *Jamovi*. Moreover, *Stata* and *SPSS* also include a test of the *proportional odds assumption* (i.e., that $OR_{12 \text{ versus } 0}$ and $OR_{2 \text{ versus } 10}$ have the same sign and are of [about] the same magnitude). For example, *SPSS* returns a deviance (−2LL) value for the proportional odds (Model 0) and non-proportional odds (Model 1) model and performs a LR (i.e., χ^2) test on the difference:

$$\chi^2 = [-2\,\text{LL}\,\underline{\text{Model}\,0}] - [-2\,\text{LL}\,\underline{\text{Model}\,1}].$$

In Experiment 1, the difference between Model 0 and Model 1 in $df = 1$, hence our χ^2-test will be a χ^2-test with $df = 1$, χ^2_1. We find: $\chi^2_1 = 0.398$, $p = 0.528$. With such an outcome, all criteria including AIC will indicate a preference for Model 0, here: the proportional odds model. *Stata*, *SPSS*, and *Jamovi* return $OR_{\text{treat}} = 1.900$, with a 95% CI from 1.052 to 3.463 ($p = 0.034$). On a *logit* scale (b), we find: $b_{\text{treat}} = 0.642$, with a 95% CI from 0.051 to 1.242. It is to be noted that this regression coefficient does not imply we are treating ordinal data as interval and that this coefficient *should not be interpreted as a linear coefficient* that we could derive through linear regression. In fact, if we were to perform a linear regression here (wrong and not recommended), we would find $B = 0.250$, quite a bit different from the coefficient obtained from ordinal logistic regression.

Experiment 2: Degrees of Identification

Although the proportional odds model constitutes the default option in software like *Stata*, *SPSS*, and *Jamovi* and is attractive in that a *single* coefficient can be obtained with regard to a treatment effect of interest, it is always recommended to first inspect separate category-to-category *OR*s (in our case: OR_{10}, OR_{20}, and OR_{21}) and cumulative *OR*s (in our case: $OR_{12 \text{ versus } 0}$ and $OR_{2 \text{ versus } 10}$) to check if merging outcomes across the full range of the outcome variable into a single coefficient (cf. proportional odds model) makes sense. In Experiment 2, we are taking a look at a case where the proportional odds assumption clearly does not hold and where the proportional odds model can surely still provide a single coefficient for the treatment effect but does not provide an accurate picture of what is actually going on.

Experiment 2 ($N = 160$) is carried out by a group of radiologists who want to compare two approaches of training—incorporated in the treatment ($n = 80$) and control condition ($n = 80$), respectively—in terms of fostering students' skills of identifying abnormalities in an x-ray. In both conditions, students are given the same x-ray with the same instructions, but the *timing* of some of the instruction differs between the two conditions (but is the same within conditions). In both conditions, participants are instructed to mark the area in the x-ray that shows an abnormality. These individually marked x-rays are evaluated by a team of fellow-radiologists who were not present during the experiment, do not know which participants were in which condition, and are blind to the researchers' reasons to expect a difference between conditions, positive or negative. For each marked x-ray, the team members come to a single, consensus-based verdict about the performance (i.e., quality of identification): 0 = not identified, 1 = acceptable, and 2 = excellent. Although this kind of categorisation is not rarely dealt with in linear models by researchers, there is no guarantee that the difference between 'not identified' and 'acceptable' is comparable to the difference between 'acceptable' and 'excellent'.

More Excellent and Poor Performance

In the control condition (n = 80), 17 participants (21.3%) are coded '0' (not identified), 32 participants (40.0%) are coded '1' (acceptable identification), and 31 participants (38.8%) are coded '2' (excellent identification). In the treatment condition (n = 80), 24 participants (30.0%) are coded '0', 13 participants (16.2%) are coded '1', and 43 participants (53.7%) are coded '2'. Hence, we observe slightly more '0' and '2' but somewhat fewer '1' codes in the treatment condition than in the control condition. Figure 7.2 presents the aforementioned percentages with their 95% CIs.

Testing $\boldsymbol{H_0}$: 'no effect' against $\boldsymbol{H_1}$: 'effect' at $\alpha = 0.05$ with a LR test, we find χ^2_2 = 0.585, p = 0.444. For R^2_{McF}, we find a value of about 0.002. For Model 0 (no

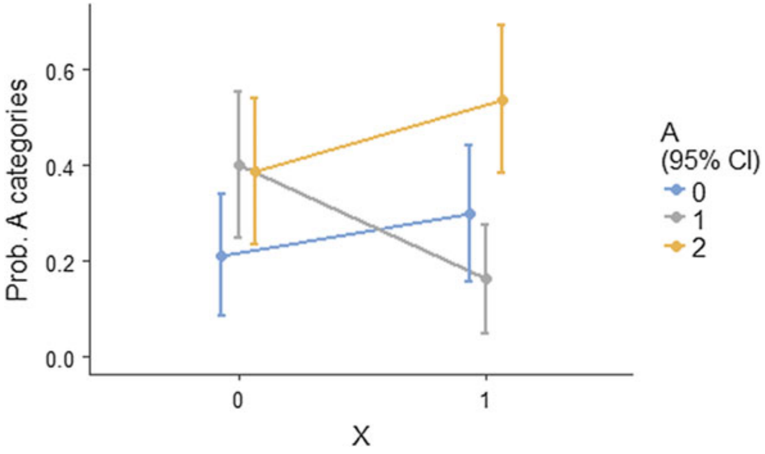


Fig. 7.2 Effects plot aka *EMM* plot aka conditional estimates plot of the findings in Experiment 2: observed proportions per category (A = 0: not identified; A = 1: acceptable identification; A = 2: excellent identification) per condition (X = 0: control; X = 1: treatment) and 95% CIs (*Jamovi*)

treatment effect), AIC = 343.941; for Model 1 (treatment effect), AIC = 345.356. For Model 0, BIC = 350.091; for Model 1, BIC = 354.581. Kendall's τ (b) = 0.057 (p = 0.448; 95% CI = [−0.029; 0.143]), and BF_{10} for Kendall's τ, using the same default prior as in Experiment 1, is 0.182. In other words, all criteria appear to point in the direction of Model 0.

When we calculate a single OR_{treat} based on the data from Experiment 2, we find OR_{treat} = 1.255, with a 95% CI ranging from 0.702 to 2.253 (p = 0.445). On a logit scale, we find: b_{treat} = 0.227, and 95% CI = [−0.354; 0.812]. Now, how do these findings make sense in the light of Fig. 7.2?

Contrary to what we find in Experiment 1, in Experiment 2 there is not just a higher proportion of '2' in the treatment condition but a higher proportion of '0' as well! Figure 7.2 indicates that *only* the proportion of '1' is lower in the treatment than in the control condition. This hints at a violation of the *proportional odds assumption*, and *SPSS* yields the following outcomes of a test of that assumption: the −2LL of Model 0 (proportional odds) is 28.714, and the −2LL of Model 1 (disproportional odds) is 17.864, hence χ^2_1 = 10.850, p = 0.001. To understand how serious the departure from proportional odds really is, we can take a closer look at the category-to-category and cumulative *OR*s. For the category-to-category *OR*s, we find:

$$OR_{1\,\text{versus}\,0} = (13/24)/(32/17) \approx 0.288, \text{ and}$$
$$OR_{2\,\text{versus}\,1} = (43/13)/(31/32) \approx 3.414.$$

This corresponds with $b_{1\ \text{versus}\ 0} \approx -1.246$ and $b_{2\ \text{versus}\ 1} \approx 1.228$. Next, for the cumulative *OR*s, we find:

$$OR_{21\,\text{versus}\,0} = (56/24)/(63/17) \approx 0.630, \text{ and}$$
$$OR_{2\,\text{versus}\,10} = (43/37)/(31/39) \approx 1.837.$$

This corresponds with $b_{1\ \text{versus}\ 0} \approx -0.463$ and $b_{2\ \text{versus}\ 1} \approx 0.608$. Clearly, these cumulative logits point in distinct directions, and the category-to-category (*OR*- and) b-values (and Fig. 7.1) help us understand why. The treatment appears to create a shift from acceptable performance towards both *more cases of excellent identification*, $b_{2\ \text{versus}\ 1}$ = 1.228, 95% CI = [0.435; 2.021] (p = 0.002), *and more cases of no identification*, $b_{0\ \text{versus}\ 1}$ = 1.246, 95% CI = [0.350; 2.141] (p = 0.006). In other words, we fail to find an overall positive or overall negative treatment effect, but the findings do indicate that the treatment may have a *positive* effect for one subpopulation of students but a *negative* effect for another subpopulation. Future experiments may shed light on factors that may influence the direction and magnitude of treatment effects for different types (subpopulations) of students.

Note that again the b-coefficients shared here ought *not* to be interpreted as b-coefficients that can be obtained with linear regression. If we were to do linear regression analysis (wrong and not recommended), instead of $b_{1\ \text{versus}\ 0} \approx -1.246$ and $b_{2\ \text{versus}\ 1} \approx 1.228$ we would find B = −0.302 and B = 0.276, respectively; the signs are the same, the magnitude and interpretation are different.

Ordinality and Different Types of Treatment Effects

Experiments 1 and 2 each constitute an example of a different type of treatment effect. In Experiment 1, the proportional odds assumption is reasonable, and we can express the treatment effect of interest in a single OR_{treat}- or b_{treat}-value. The b-value from the resulting ordinal logistic regression based on the proportional odds assumption (i.e., proportional odds model) cannot and should not be interpreted as a linear coefficient we could achieve with linear regression as well. Although the linear and ordinal model will normally agree in terms of sign, the magnitude and interpretation of the coefficients is different (e.g., Agresti, 2002). If we were to do a multinomial logistic regression instead of an ordinal logistic regression, we would lose the ordinality information. In Experiment 1, a LR test treating the outcome variable as multicategory nominal—instead of as ordinal—would yield: $\chi^2_2 = 4.934$, $p = 0.085$, and $R^2_{\text{McF}} \approx 0.015$. The problem is that with a multinomial model we do not recognise different types of performance can be ranked in terms of their quality and hence we just test if there is any difference between conditions instead of whether there is a fairly constant or at least clearly monotonous relation between the conditions that we may as well express in a single coefficient.

In Experiment 2, researchers who perform multinomial logistic regression thinking that this model tests what actually should require an ordinal logistic regression may erroneously conclude an overall treatment effect (i.e., an overall tendency of improvement or decline): $\chi^2_2 = 11.435$, $p = 0.003$, and $R^2_{\text{McF}} \approx 0.034$. That is quite a difference from the $\chi^2_2 = 0.585$, $p = 0.444$, and $R^2_{\text{McF}} \approx 0.002$ we find in a model that treats the data as ordinal. Now, it *is* true that $b_{2\ \text{versus}\ 1}$ and $b_{0\ \text{versus}\ 1}$ can be obtained via a multinomial logistic regression model (see also Chap. 6). For instance, when in *Jamovi* we define '1' as the reference category, the programme returns both $b_{2\ \text{versus}\ 1}$ and $b_{0\ \text{versus}\ 1}$. However, the fact that *OR*s and logits have their use in binary logistic, multinomial logistic, and ordinal logistic regression should not be taken as a justification to treat our data as nominal by performing multinomial instead of ordinal logistic regression. Questions drive methods, and the question that normally underlies a two-group experiment with ordinal or quantitative outcome variables is whether there is an *overall increase or decrease* in performance. With quantitative outcome variables, that usually calls for comparisons of *M*s (or sometimes: medians); with ordinal outcome variables, that usually calls for comparisons of observed (frequencies or) proportions for different ordered categories through *b*s or *OR*s. When the proportional odds assumption is reasonable, a single OR_{treat}- or b_{treat}-value provides an accurate and parsimonious estimate of a difference between two conditions. When the LR test and other criteria indicate a preference towards a 'no treatment effect' model (Model 0), this can be interpreted as insufficient evidence to assume an overall treatment effect. This may mean either of two things: the differences are overall too small to be detected, or a treatment creates more heterogeneity because its effect differs (substantially) across subpopulations.

That said, either way, as introduced in Chap. 1, analytic choices are driven by questions and the features of our design and data acquired, and when dealing with ordinal data that means that an overall test based on an ordinal logistic model constitutes a natural starting point. If there is an expectation of different treatment effects (i.e., the sign or magnitude differs substantially for different subpopulations), a two-group experiment (i.e., one-way comparison) is not the best way to start in the first place; in such cases, a two-way design with a possible moderator as second factor (if the moderator is categorical) or covariate (if it is quantitative) makes more sense (see also Chap. 11).

Final Note on Questions and Outcome

For the sake of introduction of new concepts, the experiments discussed in this chapter have an ordinal outcome variable that is measured only once in time. When series of ordinal items (e.g., several x-rays administered in a session or spread out over different sessions) are involved, we need more complex models. Depending on the questions that drive the experiment, the nature of the choice variables and data acquired, and the sample size at hand, one may have one or several alternatives to the simple approach discussed in this chapter. However, all alternatives have in common that they require much larger sample sizes than the sample in the two experiments discussed in this chapter. In fact, even for simple experiments like this, larger would be better. The smaller the sample, the higher the risk of 'zero' cells which may make the estimation of some effects impossible, and the higher the risk of cells with very small numbers (e.g., 1 or 2 per cell) that may result in fairly imprecise estimates (e.g., Agresti, 2002; McCullagh, 1980). This is also one of the reasons why in practice ordinal data are often treated as if they were interval; samples of $n = 10$ or $n = 20$ per condition make it sheer impossible to meaningfully use ordinal models. This is not to say that small samples justify treating ordinal data as interval; ordinal and quantitative outcome variable models provide different types of coefficients, and with small samples such as the ones just mentioned quantitative outcome variable models tend to yield inaccurate estimation outcomes as well.

That said, when the same ordinal item is administered repeatedly or different ordinal items are presented at the same occasion (e.g., in the same experiment), and we have access to much larger samples (e.g., experiments online), we may have several options. A first option may be multilevel aka mixed-effects ordinal logistic regression models (e.g., Bauer & Sterba, 2011; Hedeker & Gibbons, 1994, 1996; Hox, Moerbeek, & Van de Schoot, 2017; Liu, 2016). This kind of mixed-effects models applies the same logic as mixed-effects models for other types of outcome variables: participant (level 2, upper level) and question, item or occasion (level 1, lower level) constitute the hierarchical levels. Other approaches, which involve latent variables (e.g., for a comparison of non-latent and latent-variable models, see: Hox et al., 2017), can be found in latent class analysis, extensions of the Rasch model for multicategory ordinal data (e.g., Embretson & Reise, 2000; Linacre, 1989; Masters & Wright, 1996; Wilson, 1989; Wright & Masters, 1982), the Mokken model or mixed Rasch models.

Quantitative Outcome Variables

8

Abstract

After a brief summary of Chaps. 1–7, this eighth chapter delves into some important questions when dealing with quantitative outcome variables. Since an example with a quantitative outcome variable and no substantial departures from assumptions is already covered in Chap. 2, this chapter focusses on three types of somewhat more difficult situations: considerable skewness in a time outcome variable, skewness inherent to the nature of a count outcome variable for perhaps not so frequent (or rare) events, and non-linearity. For each of these three types of situations, different ways of dealing with departure from 'the typical' (i.e., Normal and linear) are presented with their advantages and disadvantages.

Introduction

The main take away from Part I (Chaps. 1–4) of this book is that data-analytic choices should be driven by the questions that led us to do the experiment, and should also be informed by the features of the experimental design that resulted from our questions as well as by the nature of the data acquired in the experiment. The first three chapters of this second part of the book (Chaps. 5–7) focus on different types of categorical variables. Although there are differences in the best ways of dealing with these different types of variables, there is also some common ground.

To start, categorical outcome variables ought to be treated as categorical, not as variables of interval or ratio level of measurement. Recent work by Van der Eijk and Rose (2015) illustrates that common practices such as factor analysis on categorical rating data may need a revision. Where we deal with multiple or repeated dichotomous or polytomous ordinal outcome variables and assumptions of continuous latent variables are reasonable, various item response theory models

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_8

mentioned in Chaps. 3, 5, and 7 provide useful alternatives to factor analysis. When dealing with multiple or repeated multicategory nominal items (Chap. 6), generalised item response tree models can be considered. Finally, when some kind of discontinuity in latent variables is likely, researchers may opt for latent class analysis or latent profile analysis as well as mixed Rasch analysis.

Treating categorical outcome variables as categorical also has implications for the statistics we use for single outcome variables, such as in experiments that focus on a difference between conditions in a single categorical outcome variable (not measured repeatedly). When dealing with categorical outcome variables, we deal with *OR*s and *b*s to estimate and express relations and effects of interest. For overall model fit, we can use a pseudo-R^2 known as R^2_{McF}. Although many software packages also report R^2_{CS} and R^2_{N}, R^2_{McF} is generally to be preferred (see Chap. 5). Moreover, some software packages (e.g., *JASP*) also provide R^2_{T} for binary logistic regression, but an advantage of R^2_{McF} is that its generalisation to multicategory nominal and ordinal outcome variables is much more straightforward than that of R^2_{T} (see Chap. 5). Moreover, R^2_{T} suffers from the same problem as R^2_{CS} of having an upper bound substantially lower than 1. In any case, R^2_{McF} (or any of the other pseudo-R^2 alternatives) cannot and should not be equated with the R^2-statistic for quantitative outcome variables. Thus, researchers should *not* run a linear regression on a categorical outcome variable and interpret the R^2-statistic from that linear regression as R^2_{McF}.

Finally, the example experiments discussed in Chaps. 5–7 demonstrate that the pragmatic approach to statistical testing and estimation (PASTE) outlined in Chap. 2 with quantitative outcome variable examples also applies to categorical outcome variables: different Frequentist outcomes, information criteria, and Bayesian outcomes can be combined to make informed and reasonable decisions. Whether we deal with *M*s, proportions or *OR*s, where CIs and CRIs can be computed, we can have a meaningful discourse with regard to the kind of effect we are dealing with: likely substantial, likely not substantial or not yet clear; see Chaps. 2 and 5 for a rationale behind these qualifications using FOST (introduction in Chap. 2, another example in Chap. 5), which can be done with TOST equivalence testing and/or the Bayesian ROPE. In this chapter, we repeat this exercise for different quantitative outcome variable scenarios: a skewed time variable, a count variable, and an outcome variable that holds a non-linear relation with a treatment factor.

Experiment 1: Skewed Time Outcome Variable

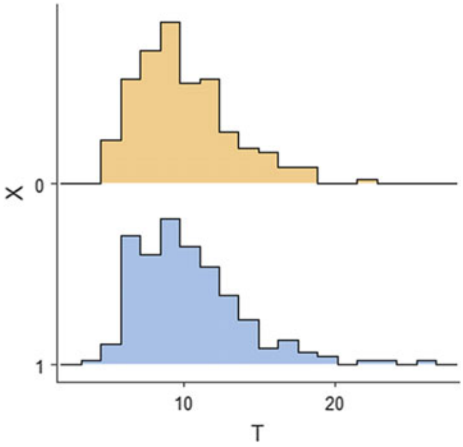
Some psychologists have been working on an online critical thinking test for Bachelor of Science students. They have developed two variants with somewhat different content scenarios but with the same set of tasks. The psychologists want to carry out a series of experiments to examine if these two variants can be considered (relatively or practically) equivalent variants of the same ‘critical thinking task’.

Since this is an online test, time from start (i.e., pressing ‘start’) to completion (i.e., pressing ‘complete’) can be measured accurately along with task performance and a few other variables. For this example, we only focus on the time outcome variable in one of the experiments carried out by this group of psychologists. In this experiment, a random sample of $N = 400$ students from different Bachelor of Science in Psychology programmes from different countries participate. They are randomly allocated to the two variants: $n = 200$ to Variant ‘0’ and $n = 200$ to Variant ‘1’. The rationale behind these numbers is as follows. Lakens’ (2017, 2018) *R* package *TOSTER* (see also Chap. 2) indicates that, using $\alpha = 0.05$ for TOST equivalence testing, with equivalence bounds $d = -0.3$ and $d = 0.3$, for a statistical power of 0.80 we would need 190.3077 participants per condition ($n = 191$), hence 382 participants in total ($N = 382$). They anticipate that even though they randomly contact $N = 400$ students, some of the students contacted may not participate. Given random allocation to condition, there is no reason to assume unequal sample size for the two conditions unless the final N is an odd instead of an even number, and a final N of 382 instead of 400 would correspond with a 4.5% non-response.

Figure 8.1 presents the histograms of the time distribution in the two conditions (variants 0 and 1), and Fig. 8.2 presents the boxplots of the time distribution in the two conditions.

Time ranges from 4.719 to 22.612 min in variant 0 and from 4.236 to 26.319 min in variant 1. The median time is 9.240 min in variant 0 and 9.730 min in variant 1. The interquartile range (i.e., from the 25th to the 75th percentile) is [7.720; 11.926] in variant 0 and [8.001; 12.186] in variant 1. The *M* time is 9.971 min in variant 0 and 10.330 in variant 1, and the *SD* is 3.156 in variant 0 and 3.453 in variant 1. As seen Fig. 8.1, time is skewed to the right in both conditions: the skewness statistic is 0.907 in variant 0 and 1.357 in variant 1 (i.e., positive values indicate skewness to the right, whereas negative values indicate skewness to the left). The kurtosis statistic is 0.934 in variant 0 and 2.945 in variant 1 (i.e., positive values indicate thinner tails than a Normal distribution, and negative values indicate thicker tails than a Normal distribution).

Fig. 8.1 Histograms of the distribution of time in minutes in variant 0 ($n = 200$) and variant 1 ($n = 200$) (*Jamovi*)



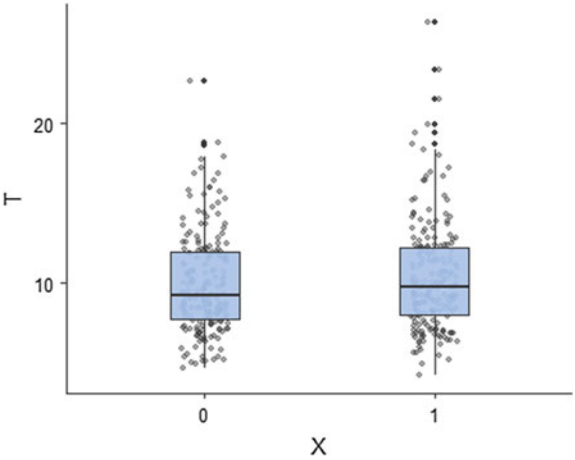


Fig. 8.2 Boxplots of the distribution of time in minutes in variant 0 (n = 200) and variant 1 (n = 200) (*Jamovi*)

Due to the skew to the right, the two conditions have a *M* time that is about half a minute higher than their median time. With this kind of distributions, different teams of researchers may prefer different routes for estimating and reporting on the time difference between conditions. Whichever route is chosen, *transparency* or *full disclosure* of steps taken is key.

Some researchers may argue that some of the cases are outliers or extreme cases and should therefore be deleted. They then either delete these cases or apply either *trimming* (i.e., omitting the most extreme scores, for example the 5% or 10% most extreme scores) or *winsorising* the most extreme cases by replacing their values by the nearest score that is not an outlier (e.g., Field, 2018). Some others might consider *robust* methods which include bootstrapping and other resampling methods. Nonparametric procedures such as Mann–Whitney's *U* test and the rank biserial correlation could also be considered. Finally, again others may consider *transforming* the outcome variable (e.g., Field, 2018).

Of these approaches, I generally prefer the ones that do *not* involve loss of data (i.e., deleting cases) or edits of data (i.e., winsorising), do not require resampling (e.g., bootstrapping is about taking repeated samples with replacement from the data set at hand), and stay as closely as possible to the original features of the data. However, which approach to choose partly depends on the nature of abnormalities in observed data. Therefore, let us compare several approaches to this right-skewed time outcome variable, some of which are encountered frequently and some of which are encountered less frequently but may be promising. The approaches compared in the following are: *analysing the data as is* (i.e., not deleting, editing or transforming anything), a *nonparametric* approach using Mann–Whitney's *U* test and the rank biserial correlation, a *robust t-test* approach based on Wilcox (2017), and a *transformation* approach with two types of transformation: square root and logarithmic. Advantages and disadvantages of these approaches are discussed.

Approach (a): Unchanged

Some researchers may argue that there appear to be some relative outliers that are not really ‘extreme’ cases; they are like the small islands above the mainland in the North of the Netherlands (i.e., Texel, Vlieland, Terschelling, and a few more) which are a short ferry trip distance from the mainland of the Netherlands rather than Hawaii in relation to the mainland of the United States of America (i.e., a 5–6 h trip by plane from Los Angeles). Moreover, although there is rather clear skewness, the central limit theorem kicks in; with samples of this size, it is more than reasonable to assume an approximately Normal sampling distribution. When sample sizes are smaller and in doubt about the sampling distribution, *R* packages like *userfriendlyscience* allow researchers to estimate the sampling distribution (see Chap. 2).

In the example experiment in Chap. 2, the largest *SD* is about 1.295 times the smallest *SD*, and the *t*-test for equal variances and that for unequal variances yield almost identical results. In Experiment 1 in this chapter, the largest *SD* is 3.453/3.156 ≈ 1.094 times the smallest *SD*. The smaller the difference in *SD*s, the smaller the difference in the two variants of the *t*-test in outcomes. Therefore, let us proceed assuming equal variances. Using *Mplus*, we find that AIC is 2095.397 for Model 0 ($\boldsymbol{H_0}$: no treatment effect) and 2096.217 for Model 1 ($\boldsymbol{H_1}$: treatment effect). BIC is 2103.380 for Model 0 and 2108.191 for Model 1. Thus, AIC and BIC both prefer Model 0. The R^2-statistic is 0.003, and the adjusted R^2-statistic is smaller than 0.001. The 95% CI of Cohen’s *d* is [−0.088; 0.305] and the 90% CI is [−0.056; 0.273] (*JASP*). The observed $d = 0.108$, and for a two-sided *t*-test of $\boldsymbol{H_0}$: ‘no treatment effect’, we find: $t_{398} = 1.084$, $p = 0.279$ (for Welch’s test assuming unequal variances, we find the same 90 and 95% CI and $t_{394.832} = 1.084$, $p = 0.279$; in short, in the first three decimals both tests yield the same outcomes). Using $d = -0.3$ and $d = 0.3$ as equivalence bounds, TOST equivalence testing (*Jamovi*; Lakens, 2017) yields the following outcomes: $\boldsymbol{H_{0.1}}$: $d < -0.3$, $p < 0.001$, and $\boldsymbol{H_{0.2}}$: $d > 0.3$, $p = 0.028$. Using a default prior (see also Chap. 2) for the *t*-test (*JASP*; Rouder, Speckman, Sun, Morey, & Iverson, 2012), we find $\text{BF}_{10} = 0.195$ ($\text{BF}_{01} = 5.121$) and a 95% CRI of [−0.086; 0.294]. In other words, applying FOST, we learn that *both* TOST *and* ROPE provide evidence in favour of relative or practical equivalence. Whether one prefers Frequentist, one prefers Bayesian, or one is open to both, these outcomes provide reasonable evidence in favour or relative or practical equivalence.

Approach (b): Nonparametric

Some researchers might prefer a nonparametric approach instead of a parametric one such as the *t*-test. For Experiment 1, Mann–Whitney’s test yields: $U = 20153$, $p = 0.363$, BF_{10} (default settings; *JASP*) = 0.199, and the rank biserial correlation is 0.053 with a 95% CI of [−0.042; 0.147]. Contrary to Approach (a), this nonparametric approach is based on *ranks* instead of *actual scores*. This comes with the clear advantage that it can be used with distributions that show severe departures from normality where Approach (a) could well be considered inappropriate. However, a main drawback of the nonparametric approach is that we pretend to be dealing with ordinal data instead of with data of interval or ratio level of measurement. To use TOST or ROPE, we need to convert bounds of a straightforward scale such as Cohen's *d* into reasonable bounds for ranks and that may seem easier than it is.

Approach (c): Robust

An alternative to Approach (a) that does *not* involve thinking in terms of ranks instead of actual scores is found in the robust independent samples *t*-test (Algina, Keselman, & Penfield, 2005; *Jamovi*; Mair & Wilcox, 2018; Wilcox, 2017; Wilcox & Tian, 2011; Yuen, 1974). In this *t*-test, and robust Cohen's *d* estimation, observed *M*s are replaced with 20% trimmed *M*s and *SD*s with the square root of a 20% winsorised variance (e.g., Algina et al., 2005; Yuen, 1974). The *R* package *WRS*2 (Mair & Wilcox, 2018), which is also incorporated in *Jamovi* includes this test, as well as alternatives to it based on bootstrapping or other robust estimators. The trim proportion (default 20%) can be adjusted to higher or lower levels (and the number of bootstrap samples, default 599, as well). Running Yuen's *t*-test with the default settings (*Jamovi*; Mair & Wilcox, 2018), we find: $t_{237.886} = 0.849$, $p = 0.397$, and (robust variant of Cohen's *d*; Algina et al., 2005) $\xi = 0.064$ with a 95% CI of [0.000; 0.215]. In line with Approach (b), the robust approach may be promising in cases where extreme departures from normality make Approach (a) inappropriate. The trimming and winsoring proportion can be adjusted to higher or lower levels in order to deal with extreme cases or one or two extremely long and thin tails, leaving the bulk of the data unchanged. In my view, the latter constitutes a clear advantage of the robust approach to the nonparametric approach; although some of the original information is lost, we are still comparing scores not ranks. In the case of only a few extreme scores in an otherwise fairly symmetric and unimodal sample distribution, the trim proportion can be reduced to 5 of 10% to deal with these extreme scores and leave the rest as is. However, in my view, the default of 20% trimming is already quite something, and in cases where we need that or a higher trim proportion, one may want to consider either Approach (b) or Approach (d).

Approach (d): Transformation

Although the nature of the changes is different, the use of nonparametric and robust methods comes with a change in either the scale of comparison (nonparametric: ranks instead of actual scores) or the tails of sample distributions (robust: trim proportion). These changes may have researchers better equipped than the *analyse as is* Approach (a) to several types of fairly extreme departures from normality, but

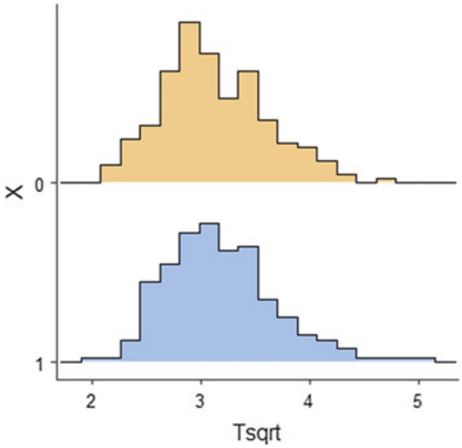


Fig. 8.3 Histograms of the square root transformed time distribution in variant 0 (n = 200) and variant 1 (n = 200) (*Jamovi*)

also come with some information loss. In the final, fourth approach discussed in this section, data transformation is applied *to all data* following a simple mathematical function that helps us to keep a link with the original scale. Two common transformations to deal with non-zero time outcome variables, which—like in Experiment 1—are often skewed to the right are *square root* transformation and *logarithmic* ($\log_{10}$) transformation. These two transformations have in common that they reduce the right-skew tendency and can as such yield transformed outcome variables that are somewhat closer to Normal.

Figures 8.3 and 8.4 present the histograms of the *square root transformed* time and the *logarithmically transformed* time per condition, respectively.

With the square root transformation, we obtain a distribution with a skewness of 0.493 and a kurtosis of 0.060 in the control condition and a distribution with a skewness of 0.792 and a kurtosis of 1.038 in the treatment condition. With the logarithmic transformation, we obtain a distribution with a skewness of 0.098 and a kurtosis of −0.282 in the control condition and a distribution with a skewness of 0.293 and a kurtosis of 0.151 in the treatment condition.

Using square root transformation, we find the following *M*s and *SD*s per condition: M = 3.120, SD = 0.485 in the control condition; M = 3.174, SD = 0.508 in the treatment condition. Using logarithmic transformation, we find: M = 0.978, SD = 0.133 in the control condition and M = 0.993, SD = 0.135 in the treatment condition. Square root and logarithmic transformation affect the *M*s, *SD*s, and shapes of the distribution; both M_ds and differences in *SD*s between conditions are smaller after than before transformation.

Using square root transformation, we find AIC = 578.627 for Model 0 and AIC = 579.472 for Model 1, and we find BIC = 586.610 for Model 0 and BIC = 591.446 for Model 1. Using the default (equal variances) *t*-test, we find: t_{398} = 1.073, p = 0.284. The 95% CI of Cohen's d is [−0.089; 0.303] and the 90% CI is [−0.057; 0.272]. These intervals very closely resemble the ones from Approach (a) (i.e., [−0.088; 0.305] and [−0.056; 0.273], respectively). Using d = −0.3 and d = 0.3 as

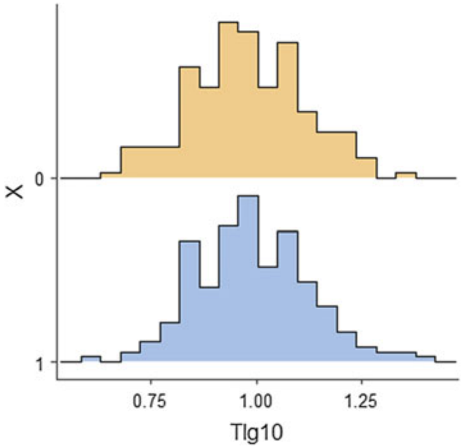


Fig. 8.4 Histograms of the logarithmically transformed time distribution in variant 0 (n = 200) and variant 1 (n = 200) (*Jamovi*)

equivalence bounds, TOST equivalence testing (*Jamovi*; Lakens, 2017) yields the following outcomes: $\boldsymbol{H}_{\mathbf{0.1}}: d < -0.3$, $p < 0.001$, and $\boldsymbol{H}_{\mathbf{0.1}}: d > 0.3$, $p = 0.027$. Using a default prior for the t-test (*JASP*; Rouder et al., 2012), we find $\mathrm{BF}_{10} = 0.193$ ($\mathrm{BF}_{01} = 5.184$) and a 95% CRI of [−0.090; 0.296]. Hence, findings very similar to Approach (a), in favour of relative/practical equivalence.

Using logarithmic transformation, we find AIC = −467.742 for Model 0 and AIC = −466.884 for Model 1, and we find BIC = −459.759 for Model 0 and BIC = −454.910 for Model 1. Comparisons of AIC and BIC values in negative territory are the same as for positive territory; the lower (i.e., the less positive or the more negative) the better. Hence, both AIC and BIC indicate a preference for Model 0. Using the default (equal variances) t-test, we find: $t_{398} = 1.067$, $p = 0.287$. The 95% CI of Cohen's d is [−0.090; 0.303] and the 90% CI is [−0.058; 0.271]. Using $d = -0.3$ and $d = 0.3$ as equivalence bounds, TOST equivalence testing (*Jamovi*; Lakens, 2017) yields the following outcomes: $\boldsymbol{H}_{\mathbf{0.1}} : d < -0.3$, $p < 0.001$, and $\boldsymbol{H}_{\mathbf{0.1}} : d > 0.3$, $p = 0.027$. Using a default prior for the t-test (*JASP*; Rouder et al., 2012), we find $\mathrm{BF}_{10} = 0.192$ ($\mathrm{BF}_{01} = 5.215$) and a 95% CRI of [−0.090; 0.294]. Again, findings very similar to Approach (a), in favour of relative/practical equivalence.

Comparison of Approaches

Although the skewness in the time outcome variable in Experiment 1 is not minimal, approaches (a) and (d) yield almost the same outcomes. These minimal differences in outcomes between Approaches (a) and (d) may provide an argument in favour of researchers who prefer to stick to Approach (a). However, when departures from normality are more severe (e.g., skewness > 2 and kurtosis > 5 in one or both conditions), differences between Approaches (a) and (d) in outcomes may well be more substantial, and we may want to report both (a) and (d). A clear advantage of approaches (a) and (d) over (b) and (c) is no loss of information due to ranking (b) or trimming (c) data. However, in some cases working with ranks or trimmed data may provide a more sensible way of dealing with departures from normality than ignoring the departures (a) or getting around them via a transformation (d). Approach (d), for example, works well when the outcome variable is skewed in the same direction such that a single transformation can be applied to reduce normality problems (cf. Experiment 1). For time outcome variables, right skew is quite common and can be reduced via a square root or a logarithmic transformation. When departures from normality are such that a single straightforward mathematical transformation does not really help us, approaches (b) and (c) are likely to provide better alternatives to (a) than (d).

Experiment 2: Skewed Count Outcome Variable

Time outcome variables constitute an example of an outcome variable that may well be skewed to the right. Another type of outcome variable in which right skew is common is that of *counts*: the outcome variable may have any non-negative integer value in a particular interval of possible or likely values. In Chap. 5, we have an example of an experiment on the effect of treatment on accident occurrence. However, in that example, the outcome variable is dichotomous: a participant either has or has no accident in the (*max*) 120 min session in the simulated environment. Suppose, we deal with a simulated driving environment where the session is not ended when there is an accident, but we count the number of errors (with or without consequences such as other drivers getting angry or ending in an accident) for each participant in a given interval. We have $N = 140$ participants randomly divided over control ($n = 70$) and treatment ($n = 70$) condition complete a 60-min session in a simulator and count the number of errors made (e.g., not giving priority to other drivers where needed, too close of a distance to the car in front) for each participant. Figure 8.5 presents the histograms of the distribution of errors per condition ($X = 0$: control, $X = 1$: treatment).

The frequencies are as follows. In the control condition ($M = 1.39$, $SD = 1.231$, variance = 1.516): 16 participants without error, 28 participants with 1 error, 16 participants with 2 errors, 6 participants with 3 errors, 2 participants with 4 errors, 1 participant with 5 errors, and 1 participant with 6 errors. In the treatment condition ($M = 1.07$, $SD = 1.108$, variance = 1.227): 25 participants without error, 28 participants with 1 error, 6 participants with 2 errors, 9 participants with 3 errors, and 2 participants with 4 errors.

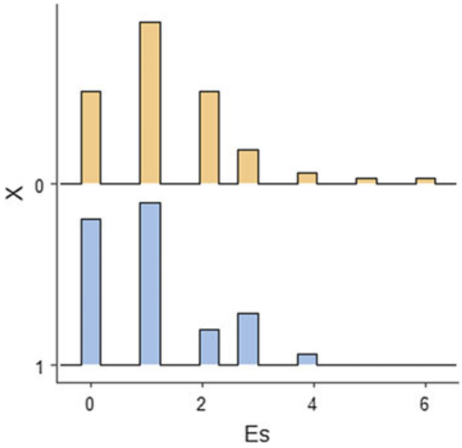


Fig. 8.5 Histograms of the distribution of errors in the control ($X = 0$) and treatment ($X = 1$) condition (Jamovi)

Poisson Distribution

Many researchers who are familiar with *t*-tests and other kinds of linear analysis are inclined to approach this kind of data with some of the approaches (a), (b), (c) and/or (d) discussed in the context of Experiment 1. However, each of these four approaches fails to appreciate a basic feature of counts, namely that they tend to follow a so-called *Poisson distribution* (i.e., the Poisson distribution is named after Siméon-Denis Poisson, 1790–1840; e.g., Nussbaum, Elsadat, & Khago, 2010). When event occurrence is so frequent that zero counts are unlikely and it is not restricted by a stringent upper limit, the Poisson distribution may somewhat resemble a Normal distribution. However, for events like in Experiment 2, low counts (i.e., 0, 1, 2) are much more likely than higher counts, and hence the distribution of counts is clearly skewed to the right. For instance, in Experiment 2, we find skewness values of 1.327 in the control condition and 0.975 in the treatment condition. Logarithmic transformation cannot be applied with zeros on the outcome variable, and square root transformation is not a solution either because the distance between counts larger than 1 becomes smaller while the distance between 0 and 1 remains the same.

Our count variable is in essence of ratio level of measurement: '0' is a natural zero, and hence 2 errors is twice as many as 1 error just like 4 errors is twice as many as 2 errors. Nonparametric approaches convert our ratio count variable to an ordinal variable, and we lose the Poisson information with it. The proportional odds model discussed in Chap. 7 suffers from the same problem and has an additional problem: several zero cells (i.e., no participants with 5 or 6 errors in the treatment condition) and several low frequency cells (i.e., in the control condition, only 1 participant with 5 errors and only 1 participant with 6 errors, and in both conditions only 2 participants with 4 errors) undermine testing and estimation.

Some may argue that although the population distribution of errors might be *Poisson* distributed (Nussbaum et al., 2010), the sampling distribution of M is approximately *Normal* with the sample size at hand. However, the violation of normally distributed residuals comes with an inflation of *SD*s and *SE*s, and that makes linear regression a substantially less powerful approach to count data than Poisson regression; it is well possible to detect a treatment effect with Poisson regression where with linear regression we fail to find that treatment effect (e.g., Nussbaum et al., 2010).

Poisson Regression

Several statistical packages have Poisson regression, including *SPSS*, *Stata*, and *Jamovi*. An important condition for obtaining valid Poisson regression estimates is *no overdispersion* (e.g., Agresti, 2002). When the variance is substantially larger than *M*, there is overdispersion; when the variance is similar to *M*, there may be no overdispersion or overdispersion may be minimal. For Experiment 2, this assumption is quite realistic (i.e., in both conditions, the variance is just a bit higher than *M*). However, when dealing with events the occurrence of which is much rarer and/or that come with longer tails (e.g., cases with 10 or more errors), variances substantially exceed the *M*s. Such an overdispersion can be accounted for by applying an overdispersion correction (Frome & Checkoway, 1985; Le, 1998) or by using regression models that are based on a negative binomial distribution (e.g., Cameron & Trivedi, 1998; Nussbaum et al., 2010). The negative binomial distribution is similar to the Poisson distribution, but the probability declines exponentially with the number of counts.

While overdispersion may be quite common in non-experimental studies (e.g., Agresti, 2002), it may be less common or be of a much lower magnitude in experimental studies. A straightforward way of testing for overdispersion is to compute the Pearson χ^2-statistic or the deviance statistic and divide this statistic by the *df*. *SPSS* provides both statistics: deviance = 168.194, and Pearson $\chi^2_{138} = 154.474$. When we divide these numbers by $df = 138$, we find 1.219 and 1.119, respectively. These numbers are still fairly close to 1, so we can proceed with Poisson regression. Using *SPSS*, we find AIC = 407.596 and BIC = 411.538 for Model 0 and AIC = 406.774 and BIC = 414.658 for Model 1. The pseudo R^2-statistic is 0.007 (*Jamovi*). Using the LR test, we find $\chi^2_1 = 2.822$, $p = 0.093$. The coefficient associated with the treatment effect is $b = -0.257$, and the 95% CI of b is [−0.561; 0.043]. The exponent of b, e^b, is 0.773 and is also referred to as *relative risk* ratio (e.g., Nussbaum et al., 2010). Note that a relative risk ratio is not the same as an *OR*; the latter is a ratio of *odds*, whereas the relative risk ratio is a ratio of *probabilities*. The ratio of 0.773 indicates that, on average, the number of errors is about 29.4% (i.e., $1/0.773 \approx 1.294$) higher in the control condition than in the treatment condition.

Correcting for Overdispersion

If we consider the Pearson χ^2-statistic/*df* or deviance statistic/*df* to be not so close to 1, we may also consider a Poisson model with correction for overdispersion and/or a negative binomial model. Both alternatives yield the same point estimates for *b* and e^b but somewhat larger *SE*s. After all, the standard Poisson model assumes no overdispersion. The more the departure from that assumption, the more the *SE*s of the alternatives will differ from that of the standard Poisson model. Negative binomial models can be run in several packages, including *SPSS*, *Stata*, and *Jamovi*. Using *Jamovi*, we find a 95% CI of *b* of [−0.579; 0.060] and a *p*-value of 0.116 for the overdispersion-corrected Poisson model, and we find a 95% CI of *b* of [−0.574; 0.056] and a *p*-value of 0.109 for the negative binomial model. Both intervals are fairly close to that of the standard Poisson model, [−0.561; 0.043].

Experiment 3: Non-linearity

Experiments 1 and 2 indicate two types of departures from normality. Experiment 3 revolves around a different type of departure from what in educational and psychological research is often considered 'normal' or 'typical': a departure from *linearity*. Where two (seemingly) quantitative variables are involved, the relation between these variables is frequently described in terms of a linear one; Pearson's *r* is then reported, perhaps with a 95% CI around it. Although such linear descriptions are attractive because they may often provide a more parsimonious and easier interpretation of a relation of interest than non-linear alternatives, when a relation of interest is inherently non-linear, linear descriptions may more often than not fall short. In any case, *linearity* is an *assumption* underlying *linear* models and should—like other assumptions—be checked. Suppose, some educational psychologists are interested in developing an online learning environment for Bachelor of Science students to practice their probability calculus skills. Many researchers cannot tell the difference between $p(O|H)$ and $p(H|O)$, one of the consequences of which is a large scale and perpetuated misunderstanding of a Frequentist *p*-value as 'the probability that $\boldsymbol{H_0}$ is true'. The *p*-value is the probability of the observed value of the test statistic (e.g., *t* or χ^2) or further away from H_0, if H_0 is true (Chap. 2), so that would be $p(\text{O or more extreme}|\boldsymbol{H_0})$. This conditional probability *cannot* and *should not* be interpreted as the probability of $\boldsymbol{H_0}$ being true. Formulations like *p* being the probability that findings are 'due to chance' are not any better; given random sampling and random allocation, findings are *always* due to chance. Some may respond that if we formulate *p* as the probability that findings are 'due to chance *alone*', the problem is solved. However, with that addition, we are effectively back to interpreting *p* as the probability of H_0 being true. After all, if there is a treatment effect, findings are due to a combination of chance *and* treatment effect, while in the absence of treatment effect findings are due to chance alone.

Starting from an Assumption of Linearity

The educational psychologists who want to develop an online learning environment for Bachelor of Science students to practice their probability calculus skills are well aware of these and other issues and see a learning environment with series of practice tasks for each of a series of different levels of difficulty as one of several possible contributors to a better understanding of statistics. They intend to develop tasks of five levels of difficulty that all use 3-by-3 contingency tables but differ in the question they ask from the student: a marginal probability (*level* 1), a joint probability (*level* 2), a conditional probability (*level* 3), the complement of a joint probability (*level* 4), and the ratio of two complements of joint probabilities (*level* 5). In other words, in each subsequent level of difficulty, *one* element is added. However, can task difficulty—with its five levels—be conceived as a linear predictor of task performance? The educational psychologists decide to investigate this by randomly recruiting $N = 250$ first-year Bachelor of Science students from different universities in the region who did not yet start any probability calculus coursework at their universities and randomly allocating them to either of five conditions: level 1 ($n = 50$), level 2 ($n = 50$), level 3 ($n = 50$), level 4 ($n = 50$), and level 5 ($n = 50$).

Graphical Evidence for Non-linearity

In each condition, students complete 40 tasks of the level (condition) they have been assigned to. Each task is completed by typing a numerator and a denominator derived from the 3-by-3 contingency table in the task. Within level (condition), the tasks differ in nothing else but the content (e.g., different types of fishes in one task, different colours cows in another task) and can be considered parallel versions of the same type (i.e., difficulty level) task. For each task, a correct response yields 1 point while an incorrect response yields 0 points. Hence, a student's total score can range from 0 to 40. All students complete the session, meaning we have $n = 50$ scores in the 0 (*min*) to 40 (*max*) range in each of the five conditions. Figures 8.6 and 8.7 present the histograms and boxplots of the distributions of scores per condition.

We find the following *M*s and *SD*s for the subsequent conditions: $M = 30.131$ and $SD = 3.853$ for level 1, $M = 25.849$ and $SD = 3.629$ for level 2, $M = 24.190$ and $SD = 3.717$ for level 3, $M = 22.886$ and $SD = 4.586$ for level 4, and $M = 23.360$ and $SD = 4.100$ for level 5. Researchers who are focussed on linear relations may now calculate Pearson's r and find a point estimate of -0.516 ($p < 0.001$) with a 95% CI of $[-0.602; -0.419]$. Although r-coefficients of such a magnitude hint at substantial linear correlations, a linear coefficient does not accurately represent the effect of difficulty on performance (i.e., score). Figure 8.8 presents a scatterplot of the findings and the *best linear unbiased estimate* (BLUE; Plackett, 1950) with its *SE* (i.e., the marked area around it).

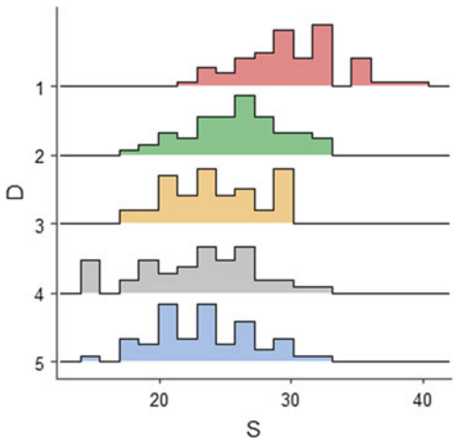


Fig. 8.6 Histogram of the distribution of score (0–40) per condition: D_1–D_5 represent the conditions, here different levels of difficulty (*Jamovi*)

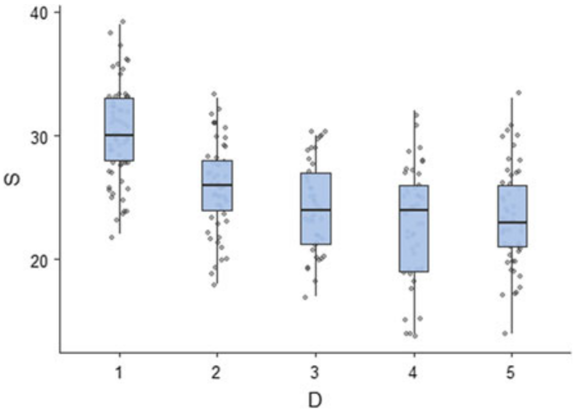


Fig. 8.7 Boxplot of the distribution of score (0–40) per condition: D_1–D_5 represent the conditions, here different levels of difficulty (*Jamovi*)

Figure 8.9 presents the best non-linear alternative with its *SE*. Figure 8.9 appears to hint that a *quadratic* model provides a better description of the difficulty-score relation than a *linear* model. The *EMM* plot in Fig. 8.10 illustrates that even more clearly.

A linear model with difficulty as predictor explains a bit over 26% of the variance in score: $R^2 = 0.266$, and adjusted $R^2 = 0.263$. In ANOVA, where all *M*s are allowed to differ freely, differences in difficulty explain a 32–33% of the variance in score: $R^2 = 0.331$, and adjusted $R^2 = 0.320$ ($\eta^2 = 0.331$, $\omega^2 = 0.319$; *Jamovi*). Note that R^2 never decreases when we go from a simple to a more complex model. However, in the ANOVA-model, we use 3 *df* more than in the linear model. The quadratic model requires only 1 *df* more than the linear model, hence 2 *df* less than the ANOVA model; however, the reduction in R^2 is minimal: $R^2 = 0.329$. In other words, with a quadratic model, we explain almost the same as with the ANOVA model but use fewer *df*. This is in line with Figs. 8.9 and 8.10.

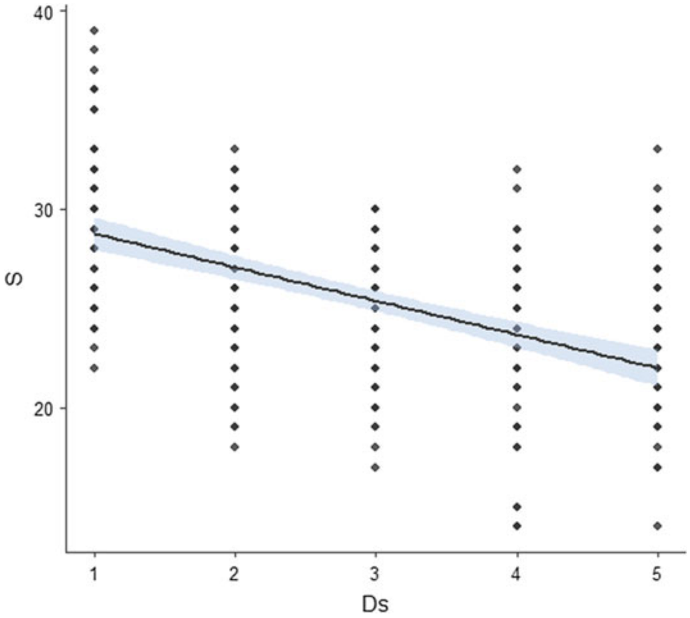


Fig. 8.8 Scatterplot of the relation between difficulty (*Ds*: Ds_1–Ds_5 represent the conditions) and score (*S*) and the best fitting straight line with its *SE* (Jamovi)

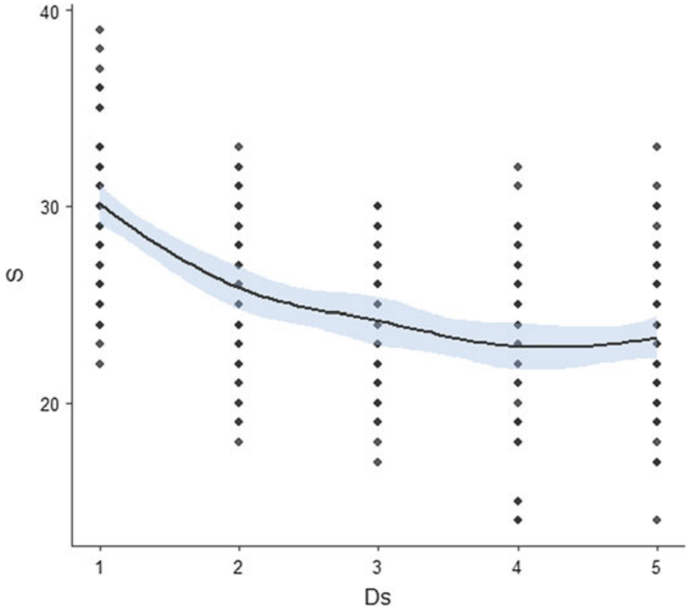


Fig. 8.9 Scatterplot of the relation between difficulty (*Ds*: Ds_1–Ds_5 represent the conditions) and score (*S*) and the best fitting non-straight line with its *SE* (Jamovi)

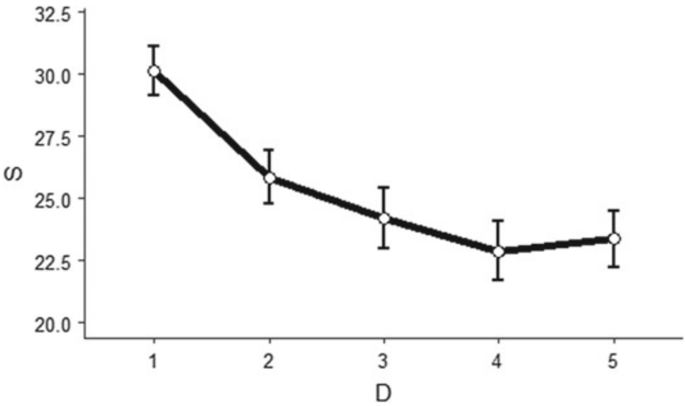


Fig. 8.10 *EMM* plot of the relation between difficulty (D: D_1–D_5 represent the conditions) and score (S): the bars around the points are 95% CIs (*Jamovi*)

Given k observed levels or conditions constituting the independent variable, we can perform polynomial regression analysis and investigate polynomials up to the power $k - 1$ (e.g., Leppink & Pérez-Fuster, 2019). With $k = 2$ all we can 'see' seems linear, and $k = 3$ constitutes the minimum for quadratic polynomials (i.e., to the power $k - 1$ comes down to squared). In the case of $k = 5$, we can test four polynomials: linear, quadratic, cubic, and quartic. In Experiment 3, we find evidence for a quadratic relation between difficulty and score. Firstly, for the quartic polynomial: $B = 0.441$ (the standardised $\beta = 0.092$), $p = 0.458$ (standardised β-values around 0.1, 0.25, and 0.40 indicate 'small', 'medium', and 'large' effects, respectively). Next, for the cubic polynomial, we find: $B = -0.267$ (the standardised $\beta = -0.055$), $p = 0.637$. Although absence of evidence does not equate evidence of absence, the testing outcomes for the cubic and quartic polynomial do not come as a surprise in the light of Figs. 8.9 and 8.10 and in the light of the difference in R^2-value between a quadratic model and the ANOVA model. If the difference in R^2 between the quadratic and the ANOVA model was larger, a cubic or a quartic model might explain substantially more than a quadratic model. However, the R^2 of the cubic and quartic model lie somewhere in between that of the quadratic model and that of the ANOVA model. Given that the R^2 of the quadratic model is almost that of the ANOVA model, going from quadratic to cubic or quartic adds close to nothing. For the quadratic polynomial, we find $B = 2.637$ (standardised $\beta = 0.547$), $p < 0.001$, which is again in line with Figs. 8.9 and 8.10, and with the difference in R^2 between the linear and the quadratic model. For the linear polynomial, we find: $B = -5.219$ (standardised $\beta = -1.083$), $p < 0.001$. Although the linear polynomial is also statistically significant (and large), the linear polynomial only influences the shape of the quadratic curve.

To Conclude: Graphs Before Statistics

This chapter provides three examples of situations where there is a departure from what many researchers view or would like to view as typical: substantial departure from normality (Experiments 1–2) or substantial departure from linearity (Experiment 3). Note that the two can go together. Researchers who tend to start with estimating a linear relation between two variables of interest may inspect a single histogram of the response variable, see and remove outliers, to then estimate a linear relation. However, what may be perceived as an outlier in a linear model may not be an outlier in a more appropriate non-linear alternative (e.g., Leppink & Pérez-Fuster, 2019). Moreover, when the independent variable consists of integer values, such as the difficulty levels in Experiment 3, inspecting histograms (and boxplots) of the distribution of the response variable of interest per level of the independent variable makes more sense than ordering a single histogram of the distribution of the response variable merged over the full range of the independent variable.

When dealing with skewed quantitative non-count outcome variables, researchers can choose from a variety of approaches how to deal with the skewness. Depending on what the graphs indicate, they may decide to ignore the skewness, to use nonparametric methods, to use robust methods, and/or to apply a simple mathematical transformation. When dealing with skewed count outcome variables, these approaches are unlikely to provide useful outcomes; instead, Poisson regression and/or overdispersion-corrected Poisson regression or negative binomial regression should be used. Again, graphs can greatly facilitate decision-making with regard to which type of model to use.

Finally, when dealing with quantitative non-count outcome variables, even if based on theory, previous research or common sense, relations of interest are assumed to be linear, it is important to check linearity assumptions. When deviations from linearity are fairly minimal, the loss in R^2 relative to likely non-linear alternatives (e.g., quadratic, cubic or varying freely like in the ANOVA model) by restricting to a linear model should be fairly minimal. However, when deviations are fairly substantial—such as in Experiment 3 in this chapter—an appropriate non-linear alternative ought to be preferred even if the linear model yields a substantial R^2.

Part III
Types of Comparisons

Common Approaches to Multiple Testing

9

Abstract

To facilitate the introduction of concepts in Parts I and II, the examples discussed in the earlier chapters of this book focus on two-group experiments, with the exception of Experiment 3 in Chap. 8 with ANOVA and polynomial regression analysis on a five-groups treatment factor. When a treatment factor consists of more than two groups but specific hypotheses about group differences are lacking, applying a correction for multiple testing to keep the rate of false alarms limited is recommendable. This chapter discusses seven types of correction for multiple testing based on Frequentist statistics (Tukey, Scheffé, Bonferroni, Dunn, Dunnett, Games-Howell, and Holm) and one type of correction for multiple testing based on Bayesian statistics (Westfall and colleagues) with their advantages and disadvantages. Next, a new approach to omnibus and follow-up testing is proposed, based on the logic of TOST equivalence testing, the Bayesian ROPE, and the TOST-ROPE uniting FOST model.

Introduction

When randomly throwing a fair die, the probability of obtaining '6' is 1 in 6. After all, there are six possible outcomes—1, 2, 3, 4, 5 or 6—and under the assumption of throwing a fair die at random, all outcomes have the same probability of occurring, in any throw. However, when randomly throwing two fair dice, the probability of at least one '6' is (1/6) + (1/6) − [(1/6) * (1/6)] = 11/36. If you find this difficult to understand, draw a tree diagram with six arms and then draw six arms on each of these six arms. These combinations of 6 * 6 = 36 arms represent the possible outcomes of randomly throwing two fair dice. Since each outcome has the same probability of occurring (1 in 6), each combination has the same probability of

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_9

occurring as well, namely 1 in 36. The eleven combinations that include at least one ‘6’ are: 16, 26, 36, 46, 56, 66, 65, 64, 63, 62, and 61. Hence, a probability of 11/36. In Chap. 2, we have a formula for calculating the probability of at least one Type I error in k independent statistical significance tests at level α:

$$\alpha_{\text{total}} = 1 - \left[(1 - \alpha_{\text{per-comparison}})^{k}\right].$$

For $k = 2$, $\alpha_{\text{total}} = 0.0975$; for $k = 3$, $\alpha_{\text{total}} = 0.142625$. The dice example above is based on the same logic. For a moment, let us use 1/6 as $\alpha_{\text{per-comparison}}$ and use $k = 2$:

$$\alpha_{\text{total}} = 1 - \left[(1 - (1/6))^{2}\right] = 11/36 \approx 0.306.$$

Whether we use 1/6, 1/20, or another statistical significance level per comparison, the more tests we perform the higher the probability of at least one Type I error. This is where the concept of *controlling for Type I error rates* comes in, and this has been and continues to be a heavily debated topic for a number of reasons.

Four Types of Error: I, II, S, and M

To start, there are scholars who fear that attempts to keep α_{total} at 0.05 (or lower) by lowering $\alpha_{\text{per-comparison}}$ come at the cost of a reduced Type II error control (e.g., Fiedler, Kutzner, & Krueger, 2012; Perneger, 1998; Rothman, 1990). In many educational, psychological and other research settings, it is already difficult to obtain sample sizes that yield a statistical power for detecting differences of 0.8 or 0.7, even without correcting for multiple testing. Lower $\alpha_{\text{per-comparison}}$ comes with a reduction of statistical power and an increased Type II error rate with it. However, performing many tests not driven by specific hypotheses is like randomly throwing a bunch of dice, and not-large-enough sample sizes often reflect a lack of time (i.e., having to carry out experiments and publish on them fast) rather than an absolute lack of resources over an extended period of time. Lowered statistical power due to applying a correction for multiple testing can be anticipated by a priori agreeing on which correction to apply and to compute the required sample size based on that correction. From this perspective, controlling for Type II error rates is more a matter of carefully designing your experiment.

If we take yet another perspective, Type I and Type II errors—at least as far as it concerns (much of) the social sciences—exist in a ‘parallel universe’ but not in the real world. Going back to Chap. 2 for a moment: “*in social science everything correlates with everything to some extent, due to complex and obscure causal influences*” (Meehl, 1990, p. 125, on the *crud factor*). From this point of view, a difference, relation or effect is rarely if ever *exactly* zero, and consequently, the null hypothesis is rarely of interest or use to researchers (e.g., Perneger, 1998), and the classical concepts of Type I Error (seeing a difference where there is none) and

Type II error (seeing no difference where there is one) are of little use if any. Rather, we should think in terms of Type S and Type M error (e.g., Gelman & Tuerlinckx, 2000): errors in *sign* (S) and errors in *magnitude* (M). Type S error is of the type that we claim a positive effect that is actually negative or vice versa, and type M error is of the type that we state a medium or large size effect while the effect is actually small or vice versa. However, from a TOST equivalence testing, a Bayesian ROPE, or FOST (see Chaps. 2, 5, and 8 for examples) perspective, the concepts of Type I and Type II error can go together with those of Type S and Type M error. Three possible outcomes of FOST (which can be done with TOST and ROPE) are: sufficient evidence against relative equivalence (a), sufficient evidence in favour of relative equivalence (b), or inconclusive (c). If we conclude (b) from our sample(s) while it is (a) in the population sampled from, we are dealing with a Type II error that can also be understood as a Type M error and eventually as a Type S error. Likewise, if we conclude (a) from our sample(s) while it is (b) in the population sampled from, we are dealing with a Type I error that can also be understood as a Type M error and eventually as a Type S error. Finally, if we conclude (c) from our sample(s) while it is either (a) or (b) in the population sampled from, a Type S error may occur but a (substantial) Type M error is probably less likely. A Type S error occurs if we are between relative equivalence (b) and clearly *positive* effect (i.e., a CI or CRI that includes only positive values) (a) while the effect is actually outside the region of relative equivalence on the *negative* side, or vice versa. However, from a Type M perspective, when we conclude (c), we state that we do not yet have sufficient evidence to draw a conclusion and hence still leave both (a) and (b) open.

In this approach, we just give somewhat wider definitions to Type I and II errors, because they now relate to a *region of relative* (or practical) *equivalence* instead of to a ‘no difference’ point state that may rarely if ever happen (at least in much of the social sciences). In FOST, a Type I error is concluding that an effect is *not part of* the region of relative equivalence while in fact it is *part of* the region of relative equivalence, and a Type II error occurs when we conclude that an effect is *part of* the region of relative equivalence when in fact it is *not part of* the region of relative equivalence.

What Are We Controlling?

Whether to control and what to control may partly depend on the school(s) of thought one identifies oneself with. As explained in Chap. 2, the Frequentist, the Likelihoodist, the information-theoretic, and the Bayesian approach are based on different philosophies and questions (e.g., Royall, 1997, 2004). In the Likelihoodist school of thought Royall (2004, p. 129), “*evidence has a different mathematical form than uncertainty. It is likelihood ratios, not probabilities, that represent and measure statistical evidence* […]. *It is the likelihood function, and not any probability distribution, that shows what the data say*.” Contrary to the Frequentist approach, the Likelihoodist and the Bayesian approach do not rely on the idea of infinitely drawing random samples of the same size *N* from the same population of interest. From a Likelihoodist perspective, the interest lies in how to interpret the data at hand as evidence regarding $\boldsymbol{H_0}$ versus $\boldsymbol{H_1}$ (Royall, 1997, 2004). Bayesian inference revolves around updating our beliefs about $\boldsymbol{H_0}$ versus $\boldsymbol{H_1}$ with incoming data. Therefore, to not so few scholars who adhere to the Likelihoodist or Bayesian school of thought, hedging our conclusion based on Test T_{i} because we happen to perform additional tests $T_{\mathrm{i+1,\ 2,\ 3,\ et\ cetera}}$ on other differences of interest may not make much sense. The same holds for scholars who adopt an information-theoretic approach to statistical testing; information criteria such as AIC and BIC are based on the deviance from models to the data and strictly do not require notions of repeated sampling. However, statistics that are not based on the Frequentist school of thought vary from experiment to experiment just like Frequentist statistics do (see for instance Tables 2.1, 2.2, and 2.3 in Chap. 2). Therefore, many Frequentist-oriented scholars would find it odd to just perform all tests possible without any kind of correction for multiple testing. Note that this is not to say that all scholars who advocate for some kind of correction for multiple testing are Frequentists or that all approaches to such corrections are Frequentist.

Many different methods of correcting for multiple testing have been developed. Which method to choose partly depends on what we are testing. One context in which many may argue to apply a correction for multiple testing is when performing a test on every bivariate correlation for a set of *k* variables. Following the formula of the number of possible tests (comparisons) C_p in Chap. 6, given *k* variables, C_p is:

$$C_p = [k * (k-1)]/2.$$

For $k = 3$, $C_p = 3$; $k = 5$, $C_p = 10$; and for $k = 8$, $C_p = 28$. If we consider it possible that a full set of $\boldsymbol{H_0}$**s** tested can be true and we have no specific hypothesis with regard to group differences, a correction to control Type I error rates is desirable. However, it is possible that about some of the correlations we already have specific hypotheses based on theory, previous research, or common sense. In that case, we may want to opt for a slightly less conservative correction for multiple testing; applying the same control for *all* correlations in that case does not make sense and comes with an unnecessary loss of statistical power (e.g., Benjamini, Krieger, & Yekutieli, 2006). The same holds for experiments in which more than two conditions comprising a treatment factor are to be compared. Let us look at an example of the latter in this chapter.

Two Treatments and One Control Condition

Some educationalists have been working on two different types of instructional support in language learning. They want to compare these two types of support to each other and to a control condition where neither of the two types of support are provided in terms of post-test performance immediately after a language learning session. Knowing that providing support does not always facilitate learning but can in some cases hinder learning as well, the educationalists decide to randomly allocate a random sample of $N = 153$ French language learners who are native speakers in English to either of the three conditions: treatment A (support type A, $n = 53$), treatment B (support type B, $n = 53$), and control (neither A nor B; $n = 53$). These numbers are based on a desired statistical power of 0.80 for a medium size difference ($f = 0.25$) with a one-way ANOVA testing at $\alpha = 0.05$ (*GPower*).

Participants in the three conditions in the experiment study the same content, just in a different condition-specific approach with regard to instructional support (A, B, or none). After the study stage, all participants complete a post-test that consists of translating 15 sentences from English to French. Each correct sentence yields 1 point, resulting in a post-test score somewhere from 0 (no correct sentences) to 15 (all sentences correct). Figure 9.1 presents the histograms of the post-test score distribution for each of the three conditions ($GR = 0$: control, $GR = 1$: treatment A, $GR = 2$: treatment B).

*M*s and *SD*s of post-test performance are as follows: $M = 8.170$ and $SD = 2.128$ in the control condition, $M = 9.302$ and $SD = 1.917$ in treatment A, and $M = 8.962$ and $SD = 1.951$ in treatment B. These differences correspond with $R^2 = 0.054$ and adjusted $R^2 = 0.042$ ($\eta^2 = 0.054$ and $\omega^2 = 0.042$). One-way ANOVA yields a statistically significant outcome: $F_{2,\ 156} = 4.468$, $p = 0.013$. Like with the two-samples *t*-test, the common ANOVA model assumes population *SD*s to be equal. In the case of a substantial departure from this assumption, there are two alternatives that do not assume equal *SD*s: Brown-Forsythe's *F*-test (Brown & Forsythe, 1974) and Welch's *F*-test (Welch, 1951). In the experiment at hand, the largest *SD* is 2.128 and the smallest *SD* is 1.917. The resulting ratio is: $2.128/1.917 \approx 1.110$. This is smaller than the ratio in Chaps. 2 and 5, where we see the two variants of the two-samples *t*-test yield near identical results. *SPSS* provides

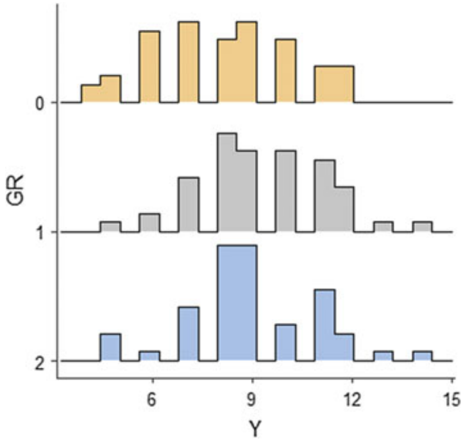


Fig. 9.1 Histograms of the distribution of post-test score in the control condition ($GR = 0$), treatment A ($GR = 1$), and treatment B ($GR = 2$) (*Jamovi*)

all three tests—the default ANOVA F-test and its two alternatives—and we see very similar outcomes: Brown-Forsythe's $F_{2,\ 154.645} = 4.468$, $p = 0.013$; Welch's $F_{2,\ 103.797} = 4.240$, $p = 0.017$. In short, we may as well proceed treating the *SD*s as (approximately) equal.

The −2LL of the null model (i.e., Model 0, which represents $\boldsymbol{H_0}$: no differences) is 677.615, and that of the ANOVA model is 668.759. In the null model, one *M* is used for all conditions; in the ANOVA model, each condition has its own *M*. Therefore, given three conditions, the difference between the two models in $df = 2$. The difference in −2LL is approximately χ^2-distributed, and the χ^2-distribution is that of the difference in *df* between the model, hence: χ^2_2. In our case, we find $\chi^2_2 = 8.856$, $p = 0.012$. For Model 0, we find AIC = 681.615 and BIC = 687.752 (*Mplus*). For the ANOVA model, we find AIC = 676.759 and BIC = 689.034. Running a Bayesian one-way ANOVA with a default prior (Rouder, Engelhard, McCabe, & Morey, 2016; Rouder, Morey, Speckman, & Province, 2012; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017; Wetzels, Grasman, & Wagenmakers, 2012) in *JASP*, we find $\text{BF}_{10} = 2.789$ (error = 0.009%). In other words, based on AIC and *p*-value, we may prefer the ANOVA model, while based on BIC we may prefer the null model, and BF_{10} indicates some preference towards the ANOVA model but the evidence is negligible.

Follow-Up or Post Hoc Comparisons

Researchers who give more weight to BIC than to other criteria may not dig further than the omnibus test. Others may argue that explaining about 5% of the post-test variance is not bad and the other criteria ($p < 0.05$, AIC, BF_{10}) hint at some difference. However, if we proceed, the question that arises is *how* we should proceed. There are many ways to proceed, but they generally have in common comparisons of sets of two conditions. In the case of three conditions, that comes down to a maximum of three comparisons.

Approach (a): Statistical Significance Testing Without Correction

Some may argue that there is no need to correct for multiple testing in post hoc comparisons—we can just test each two conditions at the same 5% as the omnibus test—for one of the following reasons. To start, by testing at $\alpha = 0.05$ in the omnibus test (i.e., one-way ANOVA, Brown-Forsythe, and/or Welch) we already have an overall Type I error probability of 5%, and that if the omnibus test yields a statistically significant difference probably at least one *M* differs from the rest. More stringent testing at this stage by lowering alpha then just results in a loss of statistical power (i.e., an increased probability of at least one Type II error). Building forth on this, some may argue that if the omnibus test indicates that probably at least one *M* differs from the rest, assuming all $\boldsymbol{H_0}$**s** to be true in the post hoc stage makes little sense. If we have four conditions and one *M* differs from the rest, for the three comparisons of two conditions that involve that different *M* $\boldsymbol{H_0}$ is incorrect, meaning that $\boldsymbol{H_0}$ can be true in only three other comparisons. In the case of three conditions, when one *M* differs from the rest, in two of the three comparisons—namely the two that involve the *M* that differs from the rest—$\boldsymbol{H_0}$ is incorrect, meaning that $\boldsymbol{H_0}$ can be true in only one other comparison. Thus, effectively, we are back to a single test at $\alpha = 0.05$, and there is no need to lower that alpha.

If we use Brown-Forsythe's or Welch's *F* as overall test, the *SE* in each of the three post hoc comparisons can be different; if we use the default one-way ANOVA *F*-test because deviations from homogeneity (i.e., heterogeneity in *SD*s) are fairly small at best, we can use *one SE* for all three comparisons (in our case: 0.389). This yields the following outcomes: control versus treatment A, $p = 0.004$; control versus treatment B, $p = 0.043$, and treatment A versus treatment B, $p = 0.384$.

Approach (b): Information Criteria for All Competing Models

A different approach to post hoc comparisons which is also based on no correction is to compare all competing models directly in terms information criteria, such as AIC and BIC. AIC and BIC can be seen as two different criteria to model comparison, with BIC being somewhat more conservative than AIC. The null model and the ANOVA model represent two of five possible models:

Model 0: null model
Model 1: control differs from A and B
Model 2: A differs from control and B
Model 3: B differs from control and A
Model 4: all *M*s are different (ANOVA model).

We already know AIC and BIC of Model 0 and Model 4. For Model 1, we find AIC = 675.535 and BIC = 684.742. For Model 2, we find AIC = 678.940 and BIC = 688.147. For Model 3, we find AIC = 683.178 and BIC = 692.385. Both AIC and BIC are lowest for Model 1. In other words, the model that states that the control *M* is different from the treatment *M*s but that the treatment *M*s do not differ from one another, appears to be preferred. AIC might in some cases prefer a difference between two *M*s where $p > 0.05$, while BIC may in some cases prefer no difference between two *M*s although $p < 0.05$. In this sense, this approach may be combined with Approach (a), that is: at least in the case of three conditions. In cases where AIC and BIC disagree, AIC may be given more weight in the case of partial availability of hypotheses (i.e., for one of the three comparisons we may already expect a difference based on theory or previous research) or where the cost of a Type II error may be high, while BIC provides a more stringent threshold than $\alpha = 0.05$ for comparisons where no hypotheses are available or where the cost of a Type I error may be high.

That said, a challenge of this information criteria approach is that with four or more conditions the number of possible models increases much faster than the number of sets of two conditions to be compared in Approach (a). A possible solution to this problem is to not compare all possible models but to select candidate models based on theory, previous research or common sense. However, when specific hypotheses are absent, that may be more easily said than done.

Approach (c): Statistical Significance Testing Involving Corrections

Quite a variety of corrections for multiple testing involving an adjustment of α can be found in the literature. This section discusses some common ones and some that are perhaps less common to many readers but may be useful in some situations.

In Chap. 2, the Bonferroni correction for multiple testing is introduced: given k comparisons, we perform each comparison not at α but at α/k. Hence, if the omnibus test is carried out at $\alpha = 0.05$ and the post hoc analysis consists of three tests, each test is carried out at $0.05/3 \approx 0.0167$. In the experiment at hand, this means that, after a Bonferroni correction, the difference between control condition and treatment B ($p = 0.043$) is no longer statistically significant. Many software packages apply Bonferroni correction not through lowering α but by multiplying the uncorrected p-values with k. This results in the following Bonferroni-corrected outcomes: control versus treatment A, $p = 0.012$; control versus treatment B, $p = 0.129$, and treatment A versus treatment B, $p > 0.999$ (software may indicate ‘$p = 1$’ or ‘1.000’; e.g., *JASP*, *SPSS*).

Several alternatives to Bonferroni correction have been developed, which like Bonferroni provide some correction but that correction differs to some extent from the one applied by Bonferroni. One intuitive approach comes from Holm. We first order the uncorrected p-values from low to high: control versus treatment A, $p = 0.004$; control versus treatment B, $p = 0.043$, and treatment A versus treatment B, $p = 0.384$. The correction for multiple testing equals k for the lowest p-value, $k - 1$ for the next p-value, and $k - 2$ for the third (i.e., highest) p-value. Hence, correction factors of 3, 2, and 1, respectively. This then yields the following Holm-corrected p-values (e.g., *JASP*): control versus treatment A, $p = 0.012$; control versus treatment B, $p = 0.086$, and treatment A versus treatment B, $p = 0.384$.

When researchers are interested only in comparisons of any treatment condition versus the control condition, Dunnett’s approach is based exactly on that logic and the correction for multiple testing is therefore a bit less conservative than that of Bonferroni or Holm. In the experiment at hand, we find: control versus treatment A, $p = 0.008$; control versus treatment B, $p = 0.079$.

Two other commonly encountered correction methods that apply somewhat different corrections but often yield results quite similar to those after Bonferroni correction or Holm correction come from Tukey and Scheffé. Tukey’s correction is also used in the Games-Howell’s approach, which allows the *SE*s to vary per comparison (i.e., a logical follow-up on statistically significant outcome from Brown-Forsythe's or Welch's *F*-test, which are recommended as alternatives to the common one-way ANOVA *F*-test in the case of substantial heterogeneity in *SD*s). Bonferroni and Holm corrections are also common in Dunn's nonparametric comparison approach. Dunn's approach is a logical follow-up if the data are distributed such that parametric tests are not considered approach. In Chap. 8, Mann-Whitney's test is discussed as nonparametric alternative to the two-samples *t*-test. A nonparametric alternative to one-way ANOVA is found in the Kruskal-Wallis test (Kruskal & Wallis, 1952). Dunn's post hoc approach constitutes a logical follow-up on a statistically significant outcome from the Kruskal-Wallis test.

Approach (d): Fixing the Prior Odds

A Bayesian approach to multiple testing comes from Westfall et al. (1997) and is implemented in software packages like *JASP*. In this approach, the posterior odds are corrected for multiple testing by fixing the prior odds to 1 (for an explanation of prior and posterior odds, see Chap. 2). When the prior odds are fixed to 1, the probability of all $\boldsymbol{H_0}$s being correct is the same as the probability of at least one of the $\boldsymbol{H_0}$s being incorrect: both probabilities are 0.5. For the experiment at hand, *JASP* returns the following posterior odds: control versus treatment A: 4.490; control versus treatment B: 0.707; and treatment A versus treatment B: 0.174. In other words, only for control versus treatment A the posterior odds are in favour of a difference. Fixing the prior odds to 1 is in line with the default for the Bayesian two-samples *t*-test, Bayesian one-way ANOVA, and Bayesian linear regression such as implemented in software like *JASP*: prior to data collection, the probabilities of $\boldsymbol{H_0}$ (no treatment effect) and $\boldsymbol{H_1}$ (treatment effect) are treated as equal (i.e., 0.5 each). In the prior odds fixing post hoc comparisons approach, this logic is applied to *all* $\boldsymbol{H_0}$s.

An Alternative Approach to Multiple Testing: FOST-OF

When three conditions are involved, at most three post hoc comparisons are involved, and in that case each of the aforementioned four approaches can be defended from one perspective or another. However, as the number of conditions increases, the number of possible comparisons increases faster. Not applying any correction for multiple comparisons in an experiment with five conditions in which *all* ten possible comparisons are made in the post hoc stage is hard to defend. With that number of conditions, the information criteria approach is also hard to implement. However, the corrections from the other two (types of) approaches may in such cases come with a substantial or even dramatic loss of statistical power. In other words, unless we make informed (i.e., hypothesis-driven) choices, any of the aforementioned approaches may be problematic.

A core problem underlying the different methods discussed under umbrella term Approach (c) is that all or most (depending on which method used) of the $\boldsymbol{H_0}$**s** are being assumed correct. In a social science world, this is highly unlikely to begin with. The same goes for Approach (d), where the prior probability of all $\boldsymbol{H_0}$**s** being correct is put at 0.5; this is a very high probability for a highly unlikely real-world scenario. Moreover, even if a 'no difference' $\boldsymbol{H_0}$ holds in a particular case, Approach (a) and Approach (c) do not provide researchers with evidence in favour of that $\boldsymbol{H_0}$. Another problem with the methods discussed thus far is that they pretend that an experiment at hand kind of provides the 'final verdict' for a hypothesis or model and that is virtually never the case. TOST, ROPE, and the TOST-ROPE uniting FOST model resonate much better with a social science reality in two ways. Firstly, differences are rarely if ever exactly zero but many differences may be too small to practically matter (i.e., the idea of relative or practical equivalence). Secondly, a single experiment (or single study otherwise) virtually never provides conclusive evidence for a difference, relation or effect of interest. Therefore, I propose a new approach to omnibus and follow-up (OF) testing that is based on TOST, ROPE, and the TOST-ROPE uniting FOST model (i.e., FOST-OF).

FOST-OF Step 1: Omnibus (O)

As in the traditional approach, we start with an Omnibus test. In the experiment at hand, this comes down to a one-way ANOVA. We have already seen some common statistics of proportion of explained variance. ANOVA is a special case of (piecewise) linear regression model. In linear regression models, R^2 and adjusted R^2 are generally reported as measures for overall model fit. What is called R^2 in linear regression jargon is commonly referred to as η^2 in the case of ANOVA, and ω^2 provides an outcome very similar to adjusted R^2. Generally speaking, ω^2 tends to be less biased than η^2 (Howell, 2010, 2017), but the difference between η^2 and ω^2 decreases with increasing sample sizes. Both η^2 and ω^2 can be calculated from the ANOVA output provided by the software programme used. Moreover, both can be used to do statistical power and required sample size calculations for future experiments. For instance, η^2-values of around 0.01, 0.06, and 0.14 are generally interpreted as 'small', 'medium', and 'large' effects, respectively, in statistical power software (e.g., *GPower*), and ω^2-values can be interpreted and used in a similar way. Just like we can define a range of Cohen's d values that represent the region of relative or practical equivalence in a given context (e.g., $-0.3 < d < 0.3$), we can do the same for η^2. In a two-group experiment, where one-way ANOVA and the two-samples t-test assuming equal SDs yield the same p-value (i.e., $df_{\text{groups}} = 1$, hence $F = t^2$), Cohen's d values of around 0.3, 0.4, and 0.5 roughly correspond with η^2-values around 0.02, 0.04, and 0.06.

When more than two groups (conditions) are involved, a problem with η^2- and ω^2-values is that they tell us nothing about *how* the different groups differ. An η^2-value of 0.06 may arise from modest differences between all Ms as well as from one

M somewhat substantially deviating from the other *M*s. However, when η^2-values are small (i.e., 0–0.02), it appears we are dealing with small differences. We may agree that, in a particular context, η^2-values in the [0; 0.02] or perhaps [0, 0.03] region indicate differences that practically speaking are not very useful. Although many software packages by default only provide point estimates of η^2, software packages like *Stata* make it very easy to calculate 90 or 95% CIs of η^2 (i.e., run ANOVA via the *anova* command, then type in the command window either *estat esize* or *estat esize, level(90)* for the 95% or 90% CI, respectively). In line with the TOST and FOST logic, I prefer a 90% CI of η^2, and in the experiment at hand, we find: [0.007; 0.114]. In a much larger experiment, or in a meta-analysis of a series of experiments, the interval would be much smaller. If in given context, the region of relative equivalence is [0, 0.03] and a meta-analysis yields a 90% CI that falls entirely within that region, we declare sufficient evidence *in favour of* relative equivalence. If a 90% CI of η^2 has no overlap with [0, 0.03], we declare sufficient evidence *against* relative equivalence. Intervals like the one at hand leave us inconclusive, which—as we have seen several times earlier in this book—is not uncommon for single experiments.

FOST-OF Step 2: Follow-Up (F)

If Step 1 yields a 90% CI of η^2 that lies completely within the region of practical equivalence, there is little reason to expect meaningful differences when comparing each set of two conditions separately. However, it is still possible that for one or some of the sets of conditions we find evidence for relative equivalence while we remain inconclusive for one or some other sets. Simultaneously, even if Step 1 yields a 90% CI of η^2 that lies completely outside the region of practical equivalence, it is *still possible* that we can establish sufficient evidence against relative equivalence for one or some of the sets of conditions but remain inconclusive or even find evidence in favour of relative equivalence for one or some other sets of conditions. In other words, in FOST-OF, Step 1 (Omnibus) is *always* followed by Step 2 (Follow-Up), *regardless* of the outcome of Step 1. The explanation for this is simple: contrary to the traditional focus of multiple testing approaches on *differences* between groups, in FOST-OF there are three equally interesting outcomes: *for relative equivalence*, *against relative equivalence*, or *inconclusive*.

In the comparison of pairs of conditions, we are back to Cohen's *d* as a useful measure of effect size. Suppose that in the given context, the [−0.3; 0.3] interval constitutes a sensible *d*-region of relative or practical equivalence. What we can do next is to calculate the 90 and 95% CI of *d* for each set of two conditions. The 90% CI constitutes the default in TOST and FOST when there is no correction for multiple testing. The 95% CI then corresponds with a factor 2 correction for multiple testing (i.e., from $2\alpha = 0.10$ to $2\alpha = 0.05$) and will be somewhat wider than the 95% CRI using a realistic prior (see also Chap. 2). No one would argue for a factor 10 correction or going from a 90% to a 99% CI, and given that in practice

Table 9.1 Cohen's *d* point estimates along with 90 and 95% CIs (LB, UB) for each of the three pairs of conditions

Comparison	Cohen's *d*	90% LB	90% UB	95% LB	95% UB
A–control	0.559	0.232	0.883	0.169	0.946
B–control	0.388	0.065	0.710	0.003	0.772
A–B	0.176	−0.145	0.495	−0.206	0.557

we often do have at least a partial set of hypotheses with regard to the group differences of interest even a factor 2 correction may be on the conservative side (see also Chap. 10). Although there are more comparisons to be made when there are four or more conditions, a complete absence of hypotheses is in such cases quite unrealistic. And again, corrections such that we would need a 99% instead of a 90% CI are based on the unrealistic assumption of all or most $\boldsymbol{H_0}$**s** being correct *and* unrealistically pretend that a single experiment can provide the 'final verdict' on a particular hypothesis or model. Table 9.1 therefore presents Cohen's *d* point estimates along with 90 and 95% CIs for each of the three pairs of conditions.

Contrary to Approach (a) and Approach (c) in the previous section, the estimates in Table 9.1 are based on comparison-specific *SE*s (i.e., assuming more or less equal *SD*s per comparison) not on one constant *SE* across comparisons. The rationale behind that is that even though the experiment at hand comprises three conditions, each comparison of two conditions in itself can be of interest and be part of separate two-group experiments. Moreover, *SD*s, just like *M*s, are rarely exactly equal. Although it is efficient to treat the *SD*s as equal in an Omnibus ANOVA when the *SD*s do not deviate substantially (as in the experiment at hand), there is no need to do so in follow-up comparisons that may be interesting as part of an experiment including more conditions but may also be interesting in isolation.

By way of comparison, let us now have a look at the 95% CRIs (cf. Bayesian ROPE) obtained in *JASP* using the default prior. For treatment A versus control, we find: [0.133; 0.896]. For treatment B versus control, we find: [-0.008; 0.724]. For treatment A versus treatment B, we find: [−0.199; 0.523]. As is to be expected, these intervals are somewhat wider than the 90% CIs and somewhat smaller than the 95% CIs. Whether we use 90% CIs, 95% CIs or 95% CRIs, the conclusion we draw is the same: inconclusive with regard to relative or practical equivalence for all three comparisons.

Where the approaches discussed earlier in this chapter stimulate a dichotomous thinking in terms of which *M*s may or may not differ, FOST-OF indicates that the best conclusion we can draw from this experiment is that we remain inconclusive. Although two of the three Cohen's *d* point estimates are outside the [−0.3; 0.3] interval and the other point estimate lies within that interval, for all three comparisons we obtain CIs that partially overlap with the [−0.3; 0.3] interval. Hence, for all three pairs, we come to the same conclusion: we have neither sufficient evidence *against* nor sufficient evidence *in favour of* relative equivalence. In my view, this reading of findings is much more informative than any set of *p*-values (corrected or uncorrected), information criteria or posterior odds.

To Conclude: PASTE-FOST-OF (PFO) as a New Framework for Thinking About Evidence

Chapter 2 introduces a pragmatic approach to statistical testing and estimation (PASTE) which comes down to combining Frequentist, information-theoretic, and Bayesian statistics to come to informed decisions: *p*-values, information criteria, and BFs are recognised as potentially useful statistics in this approach when combined appropriately—along with CIs and optionally CRIs—in the light of the questions and design of the experiment (cf. the QDA heuristic in Chap. 1). In Chaps. 2, 5, and 8, FOST—an approach uniting Frequentist TOST and Bayesian ROPE—is presented as a framework of thinking about evidence, with three possible (temporary) decisions: sufficient evidence *in favour of* relative or practical equivalence, sufficient evidence *against* relative or practical equivalence, or *inconclusive* (i.e., we find ourselves *somewhere in between* or *neither in favour nor against* relative equivalence). These decisions are rarely if ever reached in single experiments (or single studies otherwise), are subject to change with new data coming in, and are generally easier to reach in (large-sample or) meta-analytic studies. In this chapter, FOST-OF is presented as a *generalisation* or *natural extension* of FOST for experiments (or other studies) that involve multiple comparisons, such as the example experiment in this chapter. Where traditional approaches stimulate a dichotomous thinking in terms of differences versus absence of differences, FOST-OF states that we should *always* do *both* omnibus (O) *and* follow-up (F) testing and estimation regardless of whether some statistical criterion states 'significance' or some other kind of dichotomous decision. The rationale behind PASTE, FOST and its generalisation FOST-OF—which together constitute my working approach for the remaining chapters in this book: PASTE-FOST-OF (PFO)—is that differences, relations, and effects of interest are—at least for much of the social sciences—rarely if ever *exactly* zero but that many differences may be too small to really matter in a given practical context (e.g., a not sufficiently substantial improvement of educational practice with a more resource-intensive method, or an insufficiently substantial improvement in mental health with a more costly intervention).

10 Directed Hypotheses and Planned Comparisons

Abstract

In Chap. 9, different approaches to multiple testing are discussed. Most of these ideas are based on the ideas of (the possibility of) all null hypotheses being tested being true (no matter how large that series) and not having specific hypotheses that would justify focusing on only a few instead of on all comparisons. When a treatment factor consists of more than two conditions but—based on theory, previous research or common sense—we do have specific hypotheses with regard to which groups differ and eventually in what direction they differ, we may not need to compare all conditions with each other but may gain statistical power to detect treatment effects of interest by performing tests in line with the hypotheses we have (i.e., planned comparisons). In this chapter, different types of planned comparisons are discussed.

Introduction

In Chap. 2, preregistration of a priori hypotheses, in the form of for instance registered reports, is discussed as a powerful way of preregistering and defending one-sided over two-sided tests and/or defending (an appropriate method of) sequential testing. When done properly, one-sided testing and sequential testing can substantially increase statistical power for a given sample size and result in a substantial reduction of the sample size required to achieve a desired statistical power. Although in Chap. 2, sequential testing is presented in a two-sided testing fashion—which ought to be the default when a priori we have no idea what to expect—one-sided sequential testing is defendable when we do have a one-sided hypothesis about a difference of interest; the prescribed statistical significance levels then apply to one side of the difference range, knowing that differences in the other

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_10

direction in that case—even if they are large—should not be interpreted as statistically significant since they go against the prespecified hypothesis.

Next to one-sided and sequential testing, another way to increase our statistical power for a given sample size, and simultaneously, decrease the sample size required for a given statistical power, is found in the use of *planned comparisons*. For instance, when the independent variable is a dosage of medication and participants in the different conditions receive 0, 5, 10, and 15 mg, respectively, and we expect a linear relation between dosage and say performance in a driving simulator, one test may do. In another context, when in an experiment with three conditions we expect that two treatment conditions A and B lead to better driving performance than a control condition and additionally condition B will do better than condition A, two one-sided tests—one for the difference between control and treatment A, and another for the difference between treatments A and B—can do. These and other planned comparisons, and how they follow from a specific set of hypotheses, are discussed in this chapter. What all planned comparisons have in common, apart from requiring (substantially) fewer tests, is that preregistration—such as in the form of registered reports, in funded research proposals, or study proposals that were approved by an Ethical Review Board—provide powerful ways to register prespecified hypotheses and justify planned comparisons. Just like sequential testing, two-sided testing should be default but one-sided testing of planned comparisons can be defended if the hypotheses these comparisons are based on state a difference in a specific direction.

Finally, sequential testing with planned comparisons *is* possible; hence, whenever one-sided planned comparisons are available, *one-sided sequential testing of these planned comparisons* provides the best way to achieve a statistical power of 0.80 with a considerably smaller sample size. Lakens (2014) discusses how with sequential testing we may achieve a statistical power of 0.80 for a Cohen's d of 0.5 with a sample size of about $n = 51$ instead of $n = 68$ per condition or just below 80% of the sample size we need if we do not consider sequential testing. We also know that in the case of no sequential testing, opting for one-sided testing instead of two-sided testing comes with the same reduction of $n = 51$ instead of $n = 68$ per condition (e.g., *GPower*; see also Chap. 2). When we square the numbers—hence: $(51/64)^2$—we obtain 0.635 or 63.5%. This is the *proportion* of the total sample size needed in a default two-sided non-sequential test that we would need in a one-sided sequential-testing procedure. Although the *exact* proportion depends on the type of sequential testing procedure chosen (e.g., a four-tests procedure in Lakens, 2014) and which approach to correcting the statistical significance level we use (see also Chap. 2), the reduction in required sample size is *substantial* to say the least; depending on the specifics just mentioned, we may need a sample size somewhere in between 60 and 70% of the sample size we would need normally. If we multiply $n = 64$ per condition with 0.635, we obtain about 40.641 meaning $n = 41$ per condition (about 64% of 64). That is quite a reduction in sample size. In some cases, we may need to continue until the full planned sample (e.g., $n = 51$ per condition in a one-sided testing situation), whereas in some other cases we may be able to stop after an interim test, as in the following experiment.

Experiment 1: Linear

A first type of directed hypothesis and planned comparison is found in the *linear* one. Suppose, some researchers from a psychopharmacology department want to study the influence of a particular drug on cognitive performance. They decide to set up an experiment with four conditions: 0, 20, 40, and 60 mg. In each of the situations, the dose is put in a glass of orange juice, such that from the taste of the drink participants cannot tell which condition they are part of. Half an hour after the glass of orange juice, participants individually complete a cognitive task on a computer which entails reacting to different types of stimuli from different sides of the screen by clicking a stimulus- and direction-dependent button. This test results in a score of 0 (worst performance) to 20 (best performance). The researchers have good reasons to expect a (fairly) negative linear relation between drug dose and task performance: a *higher* drug dose predicts *worse* performance. This hypothesis is formulated in the study proposal approved by their local Ethical Review Board.

A random sample of $N = 120$ adults who have experience with the drug but have no known physical or mental health complaints are randomly allocated to the four conditions ($n = 30$ per condition). Not so few readers may automatically think about one-way ANOVA as a way to analyse the data and may note that a sample size of $N = 120$ does not yield enough statistical power for a medium size effect. Indeed, a one-way ANOVA with four groups of $n = 30$ each yields a statistical power of about 0.61 for a medium size difference (i.e., $\eta^2 = 0.06$; e.g., *GPower*) testing at $\alpha = 0.05$. To achieve a power of 0.80, we would need $N = 180$ participants (i.e., $n = 45$ per condition). However, contrary to the example experiment in Chap 9, one-way ANOVA does not constitute an appropriate method for Experiment 1 in this chapter. The core question driving one-way ANOVA is: *is there any difference in Ms across conditions*? The researchers in Experiment 1 are not interested in *just any difference*; they have a specific hypothesis: a *linear* relation between drug dose and task performance. Given *k* conditions, a linear contrast test uses $k-1$ number of *df* less than one-way ANOVA, because in the linear contrast test the difference in *M*s between conditions can be summarised in *a single slope*. Higher drug doses decreasing performance is translated as a *negative* slope, whereas higher drug doses increasing performance would mean a *positive* slope. The researchers hypothesise a negative slope. Assuming a medium size standardised regression coefficient (β), which in the case of a single independent variable equals Pearson's *r*, of 0.30, for a two-sided test of this linear contrast at $\alpha = 0.05$, the researchers would achieve a statistical power of 0.80 with a sample size of $n = 21$ per condition; in a one-sided test at $\alpha = 0.05$, that power would be achieved with $n = 16$ per condition (i.e., about 53.3% of $n = 30$; *GPower*). Using a one-sided test at $\alpha = 0.05$, which is defendable in this case, $N = 120$ ($n = 30$ per condition) yields a statistical power of about 0.96 for $\beta = 0.30$ and yields a statistical power of 0.80 for $\beta \approx 0.22$.

Testing Linearity (a): One One-Sided Test

Note that the power and required sample size calculations for Experiment 1 just mentioned are based on *no* sequential testing. Figures 10.1 and 10.2 present the boxplots of the distribution of task performance per condition, and the scatterplot of the relation between drug dosage (*Ds*) and task performance (*Y*), respectively.

The line in Fig. 10.2 is the best fitting non-linear relation between the two variables, and the margin around it is the *SE*. To conclude, we have little reason *not* to assume a linear relation between drug dosage and task performance. We find a standardised linear coefficient β (=r) of −0.249, with a one-sided (i.e., negative correlation) p-value of 0.003 and a 90% CI of [−0.386; −0.102]. This corresponds

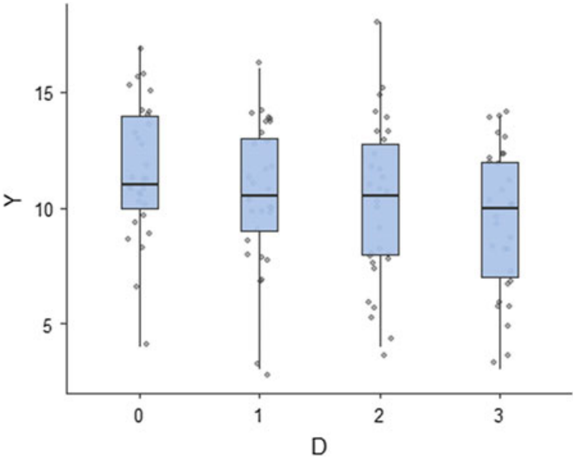


Fig. 10.1 Boxplots of the distribution of task performance per condition: D_0–D_3 represent the conditions (*Jamovi*)

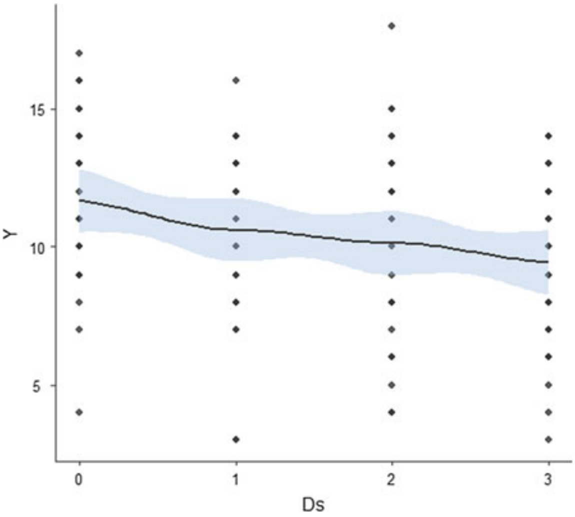


Fig. 10.2 Scatterplot of the relation between drug dosage (*Ds*) and task performance (*Y*): Ds_0–Ds_3 represent the conditions (*Jamovi*)

with a 90% CI of R^2 of about [0.010; 0.149]. Cohen's d values are about twice the size of β-values, so Cohen's d-values of 0.20, 0.30, and 0.50 roughly correspond with β-values of 0.10, 0.15, and 0.25. In other words, a region of practical equivalence for d of [−0.30; 0.30] roughly corresponds with a region of practical equivalence for β of [−0.15; 0.15] and with a region of practical equivalence for R^2 of about [0; 0.023]. Thus, we have sufficient evidence to reject positive correlations (i.e., the upper bound of the 90% CI is −0.102), but if in the context at hand [−0.15; 0.15] constitutes a meaningful region of practical equivalence, we would not yet have sufficient evidence to reject practical equivalence.

Note that we could have used a Bayesian approach as well. Using *JASP*, we find a two-sided BF_{10} for the slope of interest of 4.727 and a one-sided BF_{10} in the expected direction of 9.423. The 95% CRI of β is [−0.406; −0.076]. Although the interval still overlaps with the region of practical equivalence, it does not include positive values.

Testing Linearity (b): Sequential Testing

Suppose, the researchers had planned in advance to apply sequential testing with two tests—one after N = 60 (n = 15 per condition) and one after N = 120 (the full sample)—and they use Pocock's recommended alpha values of 0.0294 for each of the two tests (see Chap. 2). Figure 10.3 presents the scatterplot of the relation between drug dosage (Ds) and task performance (Y) for the interim analysis (N = 60 of the 120).

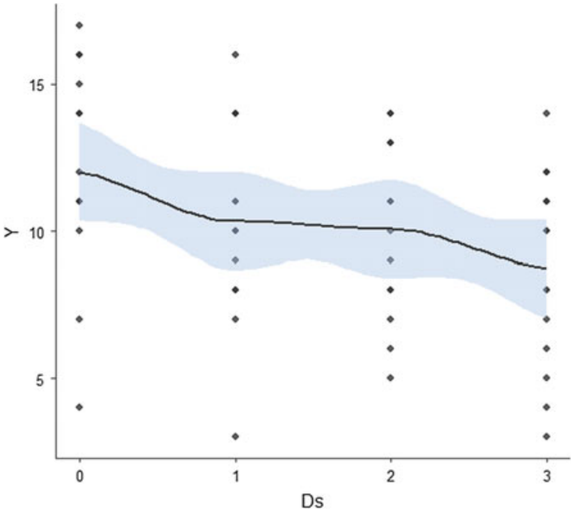


Fig. 10.3 Scatterplot of the relation between drug dosage (Ds) and task performance (Y) for the 'halfway' (i.e., N = 60 of the 120) interim analysis: Ds_0–Ds_3 represent the conditions (*Jamovi*)

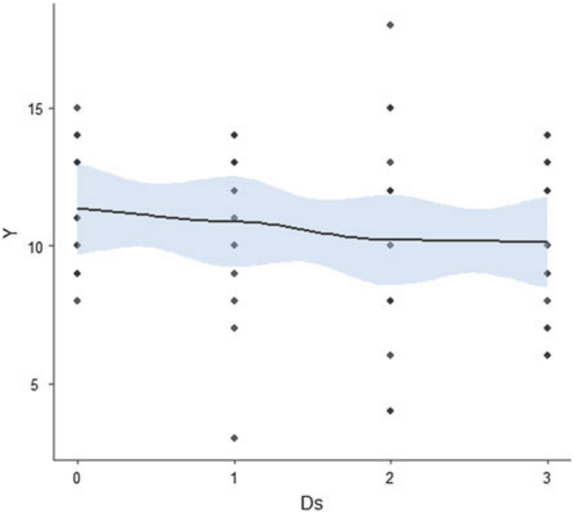


Fig. 10.4 Scatterplot of the relation between drug dosage (*Ds*) and task performance (*Y*) for the 'halfway' (i.e., $N = 60$ of the 120) interim analysis had the other sixty participants appeared first: Ds_0–Ds_3 represent the conditions (*Jamovi*)

Again, no convincing evidence for a non-linear relation between dosage and test performance. Based on the first 60 participants, we find $\beta = -0.339$ with a 95% CI of [−0.546; −0.093] and a one-sided p-value of 0.004, which is statistically significant at $\alpha = 0.0294$. Given one-sided testing at $\alpha = 0.0294$, the 95% CI is the default nearest to the $1 - (2 * 0.0294) = 0.9412$ or 94.12% CI we could use, on the conservative side. Some software packages may provide the exact 94.12% CI, but most packages provide 90, 95, and 99% CIs. That said, the difference in width of a 94.12% and a 95% CI is barely noticeable. Although the interval of [−0.546; −0.093] still overlaps with the region of practical equivalence, it includes negative values only, meaning we have sufficient evidence to reject positive slope values and we can stop our experiment; no need to continue until $N = 120$, so we can save the other 60 participants for another experiment.

Now, suppose that the 60 participants that will now be saved for a future experiment are actually the ones who participated first instead of the ones for which we just presented the findings. Figure 10.4 presents the scatterplot of the relation between drug dosage (*Ds*) and task performance (*Y*) for the interim analysis ($N = 60$ of the 120).

Again, no convincing evidence for a non-linear relation between dosage and test performance. Based on the first 60 participants, we find $\beta = -0.154$ with a 95% CI of [−0.393; 0.104]. The one-sided p-value is now 0.120, which is not statistically significant at $\alpha = 0.0294$. Hence, we continue data collection until the full $N = 120$, the one-sided p-value of 0.003 found at the end (i.e., see the section of no sequential testing) is statistically significant at $\alpha = 0.0294$. The 95% CI (i.e., default nearest to 94.12% on the conservative side) of β is [−0.410; −0.073].

Experiment 2: Helmert

In not so few cases where researchers run an experiment that includes a control condition and two treatment conditions, expectations with regard to the direction of differences *are* available, but they may not be formulated in terms of hypotheses because two-sided testing is commonly considered the default for hypothesis testing. Like in the type of scenario in Experiment 1, researchers then proceed with one-way ANOVA and, in the case of a statistically significant outcome, with one of the follow-up approaches discussed in Chap. 9. However, in the light of the hypotheses at hand, this is probably not the most sensible approach.

Let us revisit the experiment discussed in Chap. 9. Suppose, the researchers prior to that experiment hypothesised that both treatments—A and B—would outperform the control condition but had no expectations with regard to any difference between treatments A and B. In this case, two tests are needed: one for the difference between the control condition and the two treatments together, and one for the difference between the two treatments. This type of contrasting of (groups of) conditions is also known as *Helmert* coding (e.g., Field, 2018). Given our directed hypothesis of a *positive* difference between the treatments on the one hand and the control condition on the other hand, the first test ought to be *one-sided*. The second test, which is about the difference between treatments and is one the researchers had no hypothesis about, is to be done *two-sided*.

The model with these two contrasts is a multiple linear regression model which uses the same number of *df* and has the same R^2 as the default one-way ANOVA but uses these *df* differently. For the first contrast, the two treatments versus control, we find a non-standardised regression coefficient B of 0.962, which is the M_d between the two parts of the contrast. The corresponding standardised regression coefficient $\beta = 0.223$, and the 90% CI of β is [0.094; 0.351]. The one-sided p-value is about 0.002. For the second contrast, treatment A versus treatment B, we find a non-standardised regression coefficient B of 0.340, which is the M_d between the two treatments. The sign is positive, because the M of treatment A minus the M of treatment B is positive. The corresponding standardised regression coefficient $\beta = 0.068$, the 95% CI of β extends from −0.086 to 0.222, and the two-sided p-value is 0.384.

Experiment 3: Ordinal Hypothesis, Quantitative Outcome

Experiment 1 provides an example of a linear relation between a treatment factor and an outcome variable. However, assuming such a linear relation requires *equidistance* between the levels of the treatment factor; we in fact treat the independent variable as a variable of interval or ratio level of measurement. In Experiment 1, where the doses are 0, 20, 40, and 60 mg, a ratio variable (i.e., 0 is a natural 0 and hence 40 mg is twice as much as 20 mg), this makes sense. However, what if different conditions may be orderable but there is no equidistance? For example, consider an experiment where mathematics education researchers have developed two new methods for helping secondary school learners to improve their linear algebra skills. Both methods offer a type of support that is not supported in the way linear algebra is presented in textbooks and is conventionally taught in classrooms. Moreover, one new method (B) offers a bit more support than the other new method (A). In other words, we can order conventional (control), method A (treatment A), and method B (treatment B) in terms of the *degree* of support they provide but assuming *equidistance* in support between the methods may not make sense. Suppose, the mathematics education researchers have solid grounds to assume that, for novices in linear algebra, more support results in better learning outcomes. They decide to randomly allocate a random sample of $N = 159$ secondary school learners who are still novices in linear algebra to the three conditions ($n = 53$ each): control, treatment A, and treatment B. In each condition, participants complete the same series of linear algebra assignments but in their condition-specific approach. Immediately after that practice session, they complete a post-test of fifteen items that can be considered equally difficult. Each correctly responded item yields 1 point, hence the post-test score of a participant can range from 0 (all items responded incorrectly) to 15 (all items responded correctly).

While the hypothesis in Experiment 1 is a linear one, the hypothesis in Experiment 3 is an *ordinal* one. In other words, using a linear contrast like in Experiment 1 does not work here. However, the ordinal hypothesis can be translated into *two ordered one-sided tests*:

$$\boldsymbol{H_{0.1}}: \mu_{\mathrm{A}} \leq \mu_{\mathrm{Control}}, \text{ and}$$
$$\boldsymbol{H_{0.2}}: \mu_{\mathrm{B}} \leq \mu_{\mathrm{A}}.$$

These can also be written in one statement as:

$$\boldsymbol{H_0}: \mu_{\mathrm{B}} \leq \mu_{\mathrm{A}} \leq \mu_{\mathrm{Control}}.$$

The ordinal *alternative* hypothesis is:

$$\boldsymbol{H_1}: \mu_{\mathrm{Control}} < \mu_{\mathrm{A}} < \mu_{\mathrm{B}}.$$

Suppose, we find the following *M*s and *SD*s. In the control condition, $M = 10.604$, $SD = 1.801$. In treatment A, $M = 11.453$, $SD = 2.062$. In treatment B, $M = 12.113$, $SD = 2.136$. Indeed, there seems to be an order in *M*s as expected. The model with the two contrasts of interest—(1) treatment A versus control, and (2) treatment B versus treatment A—is a multiple linear regression model which uses the same number of *df* and has the same R^2 as the default one-way ANOVA (here: 0.088) but uses these *df* differently. For the first contrast, we find: $B = 0.849$ (*M* treatment A minus *M* control), one-tailed $p = 0.015$, $\beta = 0.192$, and a 90% CI of β of [0.046; 0.339]. For the second contrast, we find: $B = 0.660$, one-tailed

$p = 0.046$, $\beta = 0.150$, and a 90% CI of β of [0.004; 0.296]. Had we done a one-way ANOVA followed up with Bonferroni-corrected post hoc comparisons—which is what quite some researchers do in this kind of case—the two-tailed Bonferroni-corrected p-values of these contrasts would have been 0.092 (first contrast) and 0.276 (second contrast); *six* times larger than needed (i.e., two-tailed instead of one-tailed, and Bonferroni correction instead of no correction).

Experiment 4: Ordinal Hypothesis, Dichotomous Outcome

Let us repeat the exercise of Experiment 3 with a dichotomous outcome variable. Suppose, we design an experiment to test the same kind of ordinal hypothesis as in Experiment 3 but the outcome variable of interest in our experiment is pass/fail. We register our ordinal hypothesis in our study proposal that will be approved by the local Ethical Review Board and will be undergo first-stage peer review with a registered reports journal. The outcome of that first-stage peer review is positive, so we can start collecting data, add the Results and Discussion section to our manuscript and submit that to the same journal. Provided that we have adhered to the plan outlined in the study proposal, our manuscript will get accepted for publication regardless of the findings (Center for Open Science, 2018).

Suppose, the pass rates are as follows: 28.3% in the control condition (15 out of 53 participants), 47.2% in treatment A (25 out of 53 participants), and 67.9% in treatment B (36 out of 53 participants). We can now run a binary logistic regression model (see also Chap. 5) with the same two contrasts as in Experiment 3. $R^2_{\text{McF}} = 0.078$. For the first contrast, we find a b of 0.816 with a 90% CI of [0.141; 1.492] and a one-tailed p-value of 0.023. For the second contrast, we find a b of 0.864 with a 90% CI of [0.201; 1.526] and a one-tailed p-value of 0.016. Again, in the typical Bonferroni-correction approach that we see in the research literature, the p-values would be *six* times higher.

Note that contrary to previous chapters, in none of the experiments in this chapter we report LR tests or information criteria such as AIC or BIC. The reason for this is that these criteria work are based on two-sided testing; positive and negative differences are treated as the same: as *differences*. Consequently, although there are p-values and BFs for two-sided testing for each of two possible one-sided directions, there are no such equivalents for AIC and BIC. For the LR test that is done on a single comparison that is in the expected direction, a solution is to divide the resulting p-value by two or to test the resulting p-value at 2α (i.e., a one-sided test for a difference in a given direction at level α for that critical value corresponds with a test in which no distinction in direction is made at 2α).

Experiment 5: Ordinal Hypothesis, Ordinal Outcome

Finally, suppose that we had the same ordinal hypothesis and that the outcome variable in our experiment was neither quantitative (Experiment 3) nor dichotomous (Experiment 4) but *ordinal* (see also Chap. 7): poor performance (0), acceptable performance (1), great performance (2). Performance in this final experiment is as follows. In the control condition, we have 26 cases of poor, 16 cases of acceptable, and 11 cases of great performance. In treatment A, we have 17 cases of poor, 16 cases of acceptable, and 20 cases of great performance. In treatment B, we have 9 cases of poor, 16 cases of acceptable, and 28 cases of great performance. *SPSS* returns deviance (−2LL) values of 24.803 for the *proportional odds model* and 24.534 for the *no proportional odds* alternative. The resulting LR test is: $\chi^2_2 = 0.269$, $p = 0.874$. This outcome is very similar to the outcome in Experiment 1 in Chap. 7; we have no reason to go beyond proportional odds. $R^2_{\text{McF}} = 0.046$, and the two contrasts yield the following outcomes (*Jamovi*). For the first contrast, $b = 0.776$, one-tailed $p = 0.017$ (with a LR test, dividing the resulting p-value by two, we find $p = 0.016$), and the 90% CI is [0.179; 1.383]. For the second contrast, $b = 0.692$, one-tailed $p = 0.030$ (with a LR test, dividing the resulting p-value by two, we find $p = 0.029$), and the 90% CI is [0.092; 1.300]. As in Experiments 3–4, the p-values would be six times the ones found here in the default two-sided testing with Bonferroni correction habit.

To Reiterate: The Many Uses of 90% Confidence Intervals

Several earlier chapters in this book (Chaps. 2 and 5–9) together provide quite a few examples from different types of experiments of how 90% CIs can be of use even if two-sided testing is applied. After all, statistically non-significant p-values from ‘no difference’ null hypothesis tests cannot be interpreted as evidence in favour of that null hypothesis. Besides, p-values and BFs involving ‘no difference’ null hypotheses do not provide information with regard to whether an effect of interest is of practical importance or it is more likely to be somewhere in a region of practical equivalence that represents values that from a practical perspective are not really interesting. The same 90% CI that can be used in TOST equivalence testing to provide evidence in favour of relative equivalence can also be used to provide evidence against relative equivalence, just like the 95% CRI has a similar function in the Bayesian ROPE. Since rejecting practical equivalence requires two one-sided tests as well (see Chap. 2), as stated in FOST, the same 90% CI can be used for the decision with regard to whether or not we have sufficient evidence against or in favour of practical equivalence. In this chapter, we remember that the 90% CI also remains useful to understand the outcomes of one-sided tests of ‘no difference’ null hypotheses. As seen in Chap. 9 and also in Experiment 2 in this chapter, the 95%

CI provides a useful backup alternative when we want to apply a reasonable correction for multiple testing. Therefore, where we consider $\alpha = 0.05$ as an acceptable statistical significance level, we should report 90% CI by default, and where considered appropriate, the 95% CI in addition to the 90% CI.

Two-Way and Three-Way Factorial Designs

11

Abstract

The previous chapters in this book focus on one-factor experiments, that is: there is only one independent variable aka treatment factor. When two or more factors are involved, different types of effects can be distinguished: main effects, interaction effects, and simple effects. Although experiments with four or more factors are rather uncommon in educational and psychological research, experiments with three factors (three-way design) and especially experiments with two factors (two-way design) are common in these fields. This chapter presents some important guidelines for the testing, estimation, interpretation, and reporting of main effects, interaction effects, and simple effects. For the sake of simplicity of the introduction, covariates are not yet included in this chapter; they are introduced in Chap. 12 (no repeated measurements) and Chap. 15 (repeated measurements). However, the main, interaction, and simple effects distinction and guidelines discussed in this chapter are also of use when dealing with covariates, unless we deal with baseline measurements (i.e., prior to treatment) in randomised controlled experiments (see Chap. 15). Two example experiments are discussed in this chapter: first one with a two-way design, then one with a three-way design.

Introduction

Suppose, we are interested in the effect of providing instructional support to novices who are practicing with a systematic approach to solving a particular type of problem and how that effect of instructional support may depend on whether or not these learners are provided with feedback right after task practice. There is some evidence that, for the type of problem-solving approach at hand, novices tend to learn a bit more from practicing completion tasks (i.e., support) in which they have

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_11

to complete the missing steps than from practicing complete tasks (i.e., having to do *all* steps by themselves, no support). However, not much is known about how the effect of providing this type of guidance depends on whether students receive feedback about their performance right after a task or only at the end of a practice session that involves a larger series of tasks.

We have developed an online learning environment in which high school learners can practice their skills of solving heredity problems where the genotypes of some family members are known but learners have to find the possible genotypes for the other family members. Every problem is to be solved through a five-step approach. In the *support* condition, the first three of these steps are already worked out and learners have to complete the final two steps in order to solve the problem. In the *no support* condition, learners have to complete all five steps autonomously, without support. In the *feedback* condition, learners receive feedback on their performance after each completed problem. In the *no feedback* condition, learners receive only general feedback at the end of the practice session. For all learners, a practice session consists of ten problems. After this practice session, all learners complete a post-test of 20 items, each of which yields a maximum of 5 points (0 = all steps incorrect, 5 = all steps correct), and hence the total score can vary from 0 (all incorrect) to 100 (all correct). For the sake of simplicity for the introduction of the concepts of main, interaction, and simple effects, we are not going to delve into psychometric questions concerning the post-test score; the latter are discussed in Chaps. 14, 15 and 16 of this book.

Since heredity is a topic taught across high schools in the country where we work, we have no trouble drawing a random sample of $N = 240$ high school learners for our experiment. We randomly allocate them to either of four possible combinations ($n = 60$ each): support *no* immediate feedback *no*, support *no* immediate feedback *yes*, support *yes* immediate feedback *no*, and support *yes* immediate feedback *yes*. In other words, we are dealing with a two-way design, in which support (no, yes) and immediate feedback (no, yes) are the two factors, with two levels each (i.e., no vs. yes). These numbers yield sufficient statistical power not only for medium but also for somewhat smaller effect sizes (e.g., *GPower*).

Different Types of Effects: Main, Interaction, and Simple

Figures 11.1 and 11.2 present the histograms of the distribution of post-test performance (0–100) and the *M*s with 95% CIs in the four conditions in our experiment, respectively.

The *M*s and *SD*s are as follows. In the condition where participants received support but no immediate feedback, $M = 48.150$, $SD = 14.568$. In the condition where participants received neither support nor immediate feedback, $M = 43.000$, $SD = 16.059$. In the condition where participants received both support and immediate feedback, $M = 54.083$, $SD = 16.152$. In the condition where participants received no support but immediate feedback, $M = 63.800$, $SD = 12.880$. Figure 11.2

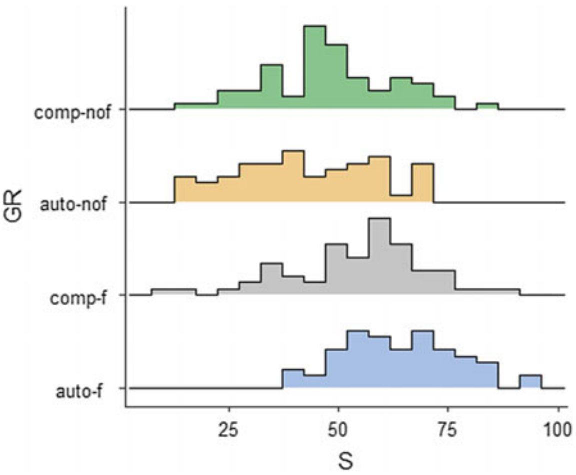


Fig. 11.1 Histograms of the distribution of post-test performance score (*S*: 0–100) in each of the four conditions: support without immediate feedback (*comp-nof*), neither support nor immediate feedback (*auto-nof*), support and immediate feedback (*comp-f*), and no support but immediate feedback (*auto-f*) (*Jamovi*)

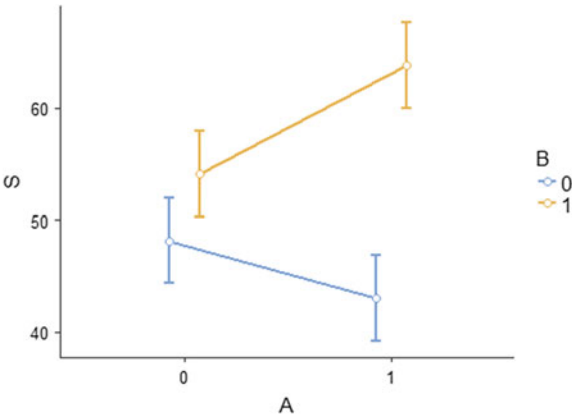


Fig. 11.2 *M*s and 95% CIs around these *M*s for each of the four conditions (*A* = 0: completion, support and hence *not* autonomous; *A* = 1: autonomous, no support; *B* = 0: no immediate feedback; *B* = 1: immediate feedback); *S* represents the score (*Jamovi*)

presents the *M*s and 95% CIs around these *M*s for each of the four conditions (*A* = 0: completion, support; *A* = 1: autonomous, no support; *B* = 0: no immediate feedback; *B* = 1: immediate feedback).

The pattern that we see here hints at what, depending on the field in which one is active, is called *effect modification* aka *moderation* aka an *interaction effect* (e.g., Field, 2018): the effect of support on post-test outcome appears to depend on whether or not learners receive immediate feedback. In our experiment, among learners who do not receive immediate feedback ($B = 0$ in Fig. 11.2), the condition in which participants have to perform the task themselves ($A = 1$ in Fig. 11.2) on average performs worse than the condition in which participants receive support ($A = 0$ in Fig. 11.2, the learners who do not have to do everything by themselves). Simultaneously, among learners who do receive immediate feedback ($B = 1$ in Fig. 11.2), the condition in which participants have to perform the task themselves on average performs better than the condition in which participants receive support. If the lines in Fig. 11.2 were (more or less) parallel, we could speak of *main effects* of support and feedback on post-test performance: the effect of support on post-test performance would then be (approximately) the same for learners who receive immediate feedback as for learners who do not receive immediate feedback, and the effect of feedback on post-test performance would then be (more or less) the same for learners who receive support as for learners who do not receive support. In such a case, looking at *simple effects* would not be needed. The simple effects of support on post-test performance are the effects of support on post-test performance (1) for learners who receive immediate feedback and (2) for learners who do *not* receive immediate feedback. Likewise, the simple effects of immediate feedback on post-test performance are the effects of immediate feedback on post-test performance (1) for learners who receive support and (2) for learners who do *not* receive support. The study of simple effects only makes sense if we find a (substantial) interaction effect, because only in that case the simple effects (of feedback per level of support, and of support per level of feedback) differ (substantially).

Competing Models

In Chap. 9, we see one-way ANOVA as a special case of linear regression analysis. As in Chap. 9 and 10 in this book, we can compute R^2 for the model as a whole, but we can now compute effect size estimates for interaction, main, and—if needed—simple effects as well. Moreover, when treatment factors include more than two levels (e.g., Chap. 9) but we do not have specific hypotheses with regard to effects of interest (as in Chap. 10), a follow-up approach (cf. Chap. 9) may be needed for at least one main or simple effect. Two-way ANOVA, a special case of multiple linear regression generally constitutes a sound method of analysing the kind of data found in the experiment in this chapter. In the absence of specific hypotheses with regard to the direction of a main or interaction effect, two-sided testing constitutes the default, and there five possible models:

Model 0: null model;
Model 1: only a main effect of support (A);
Model 2: only a main effect of immediate feedback (B);
Model 3: two main effects (A and B); and
Model 4: two main effects (A and B) plus an interaction effect (A *by* B).

The interaction effect is a *combined* effect of the two factors and can only be examined in a model that also includes the main effects of the two factors. Therefore, there is *no such a model as* ‘Model 5: interaction but no main effects’, and Model 4 is also referred to as full (factorial) model (i.e., all effects are present). Given that R^2 cannot decrease when we add an effect to our model (R^2 adjusted can though), R^2 is lowest for Model 0 (i.e., 0) and highest for Model 4. However, if the increase in R^2 by adding an effect is minimal, the question arises whether we need to include that effect. In the traditional ‘no difference’ NHST approach, we start with Model 4 and compute the p-value of the interaction effect. If that p-value is not statistically significant, we continue with Model 3 and compute the p-values of the main effects. If one or both main effects are not statistically significant, we may follow up with Models 1 and 2 and compute the p-value of each of the two main effects. A more efficient way of comparing all these models is to compare all models *simultaneously* in terms of information criteria such as AIC and BIC and/or in terms of BFs (*JASP*; Rouder, Engelhard, McCabe, & Morey, 2016; Rouder, Morey, Speckman, & Province, 2012; Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017; Wetzels, Grasman, & Wagenmakers, 2012). The model with the lowest AIC and BIC and the highest BF is generally to be preferred, with larger differences indicating a clearer difference between the preferred model and the other candidate models. Table 11.1 presents the R^2, AIC, BIC, and BF for each of the five models (Models 0–4).

The best model clearly is Model 4; the gain in R^2 from Model 3 to Model 4 is quite substantial, and AIC, BIC, and BF indicate that Model 4 is to be preferred over the other models. The logic of the BF is similar to that in earlier chapters on the t-test and one-way ANOVA, only that the number of competing models is larger now. In the case of a one-way ANOVA, the null model and the alternative model constitute the two competing models and are given equal prior probability each: 0.5. In the case of two-way ANOVA, all competing models are given equal prior probability as well: 0.2. Generally speaking, given m number of competing models, the prior probability of each model equals $1/m$. Whether this is the best choice is open to debate, but the rationale behind this approach is that if we have no evidence to prefer a particular model we may as well treat them as equally likely until we observe data with which some models become more likely while other models become less likely. We usually use BF = 1 for the null model or, in some cases, for

Table 11.1 R^2, AIC, BIC, and BF for each of the five competing models for the quantitative outcome variable (*Mplus* for the R^2, AIC and BIC, *JASP* for the BF using default priors)

Model	R^2	AIC	BIC	BF
0: null	0.000	2037.676	2044.637	1.000
1: main A	0.005	2038.557	2048.999	2.390e−1
2: main B	0.159	1998.011	2008.453	6.060e+7
3: main A & B	0.164	1998.680	2012.602	1.553e+7
4: full factorial	0.213	1986.096	2003.499	2.256e+9

the best model. When we use BF = 1 for the null model, all models with BF < 1 are considered *worse* than the null model whereas all models with BF > 1 are considered *better* than the null model. When we use BF = 1 for the best model (here: Model 4), all other BFs are smaller than 1, and the inverse of each of the BFs than indicates the BF in favour of the best model over each of the respective alternatives. For instance, in Table 11.1, Model 4 is the best model; if we specify BF = 1 for that model, the inverse of the BF of Model 3 indicates the strength of evidence of Model 4 over Model 3. In that case, the inverse of the BF of Model 3 would be a bit over 145, indicating (very) strong evidence in favour of Model 4 over Model 3. Given that the only difference between these two models is the interaction effect, the resulting BF for the difference between these models can be used as an indicator of evidential strength of the interaction effect. In terms of model preference, the BF is usually situated somewhere in between AIC and BIC (see also Chap. 2).

The η^2 of the interaction effect is 0.049 (partial $\eta^2 = 0.059$, $\omega^2 = 0.046$, $p < 0.001$), which constitutes a small-to-medium effect (values around 0.01, 0.06, and 0.14 are generally interpreted as 'small', 'medium', and 'large' effects; e.g., *GPower*). The η^2 of the two main effects are 0.005 and 0.159, respectively (for main effect A, partial $\eta^2 = 0.006$, $\omega^2 = 0.001$, $p = 0.239$; for main effect B, partial $\eta^2 = 0.168$, $\omega^2 = 0.156$, $p < 0.001$).

Interaction Follow Up: Simple Effects Analysis

When Model 4 is to be preferred, we may want to follow up with simple effects analyses. This follow up would not make sense if another model was to be preferred, because the effect of one factor on the outcome variable of interest would likely be (approximately) the same across levels of the other factor. When Model 4 is the best model, analysing simple effects *may* be useful. I stress *may* be useful, because tests of simple effects generally have a substantially lower statistical power than tests for main or interaction effects. Moreover, Fig. 11.2 already indicates what the interaction effect may look like. If apart from estimating the magnitude of the interaction effect we have a specific interest in simple effects, we *can* test and estimate one or two types of interest depending on our interest: (1) the effect of support per level of the immediate feedback factor and/or (2) the effect of immediate feedback per level of the support factor. This way, we can obtain statistical testing outcomes and CIs or CRIs for each simple effect. However, the intervals we obtain will probably be quite wide and perhaps not so informative. For instance, for the effect of support among learners who receive no immediate feedback, we obtain a 95% CI of Cohen's d of [−0.025; 0.696] and a 90% CI of [0.033; 0.638]. Note that the sample size of the experiment in this chapter is considerably larger than what we often see in educational and psychological research; in the samples of sizes that are more common in these fields, the intervals will be considerably wider. Even with intervals of a width like in this experiment, it is difficult to establish evidence in favour of a (meaningful) difference. Even if an interval does not include zero, the point estimate will have to be well in the medium-to-large range (d-values well above 0.6 or well below −0.6) for the interval not to overlap with [−0.3; 0.3] or a similarly reasonable region of practical equivalence.

Since we always report the *M*s and *SD*s in our manuscripts, meta-analyses over a series of experiments can obtain much more accurate estimates of simple effects. Based on Fig. 11.2 and Table 11.1, we can draw the temporary tentative conclusion that the effect of feedback is larger for learners who receive no support ($A = 1$ in Fig. 11.2) than for learners who do receive support ($A = 0$ in Fig. 11.2) and that practicing without support is easier in the presence than in the absence of immediate feedback. Replication experiments and eventually a meta-analysis can help to investigate this phenomenon further and to obtain sufficiently accurate simple effects that are hard to obtain in single experiments.

A Note on Coding

Knowing that two-way ANOVA can be seen as a special case of multiple linear regression, not so few researchers directly run a multiple linear regression with the software package they use and code the factors manually. Commonly, the two levels of a dichotomous treatment factor are then coded '0' and '1' (i.e., dummy coding), and when dealing with multicategory treatment factors multiple 0/1 (dummy) variables are created. However, in ANOVA, not dummy coding but *contrast* coding is applied. Statistical software packages like *Jamovi* show the user how this coding is applied, and allows users to switch to different kinds of coding (e.g., polynomial, which is sometimes useful for multicategory treatment factors as seen in Chaps. 8 and 10) if desired. With contrast coding, what is listed as main effect are main effects; with dummy coding, we are actually interpreting *simple effects* while we may think we are interpreting main effects. Consider the following regression equation for the predicted post-test score S_p using Model 4:

$$S_p = B_0 + [B_1 * \text{Factor A}] + [B_2 * \text{Factor B}] + [B_3 * \text{Interaction}].$$

In this formula, B_0 is the intercept, and B_1, B_2, and B_3 are slopes. For condition $A = 0$ & $B = 0$, $S_p = B_0$. In other words, using dummy coding, the intercept is the *M* of condition $A = 0$ & $B = 0$. In two-way ANOVA, the intercept is the *overall M* of all conditions together (e.g., Field, 2018; Howell, 2017). Next, for condition $A = 1$ & $B = 0$, $S_p = B_0 + B_1$. In other words, what one may from the output interpret B_1 as 'main effect of *A*' is in fact the *simple effect* of *A* for $B = 0$. Likewise, for condition $A = 0$ & $B = 1$, $S_p = B_0 + B_2$; this is not the main effect of *B* but the *simple effect* of *B* for $A = 0$. Finally, for condition $A = 1$ & $B = 1$, $S_p = B_0 + B_1 + B_2 + B_3$.

From the aforementioned calculations, it follows that the difference between condition $A = 1$ & $B = 1$ and condition $A = 1$ & $B = 0$ is: $\Delta S_p = [B_0 + B_1 + B_2 + B_3] - [B_0 + B_1] = [B_2 + B_3]$. Likewise, the difference between $A = 1$ & $B = 1$ and condition $A = 0$ & $B = 1$ is: $[B_0 + B_1 + B_2 + B_3] - [B_0 + B_2] = [B_1 + B_3]$. Finally, the difference between $A = 1$ & $B = 1$ and condition $A = 0$ & $B = 0$ is: $\Delta S_p = [B_0 + B_1 + B_2 + B_3] - B_0 = [B_1 + B_2 + B_3]$. In this, B_1 and B_2 are simple effects, and B_3

expresses the extent to which the lines in Fig. 11.2 are not parallel (i.e., the *interaction effect*). Therefore, whether we do contrast coding or dummy coding, the p-value we would obtain for the interaction effect is going to be the same; however, the meaning and interpretation of the other coefficients are different. However, in Model 3, which does *not* hold the interaction effect, the regression equation is as follows:

$$S_p = B_0 + [B_1 * \text{Factor A}] + [B_2 * \text{Factor B}].$$

In this model, we *can* interpret B_1 and B_2 as main effects; in the absence of the interaction term, the simple effect of A on S_p is the same across levels of B and the simple effect of B on S_p is the same across levels of A. Note, however, that using Model 3 would only make sense if from the criteria in Table 11.1 (and eventually the p-value of the interaction effect, which for the data at hand is <0.001) Model 3 was to be preferred over Model 4 (which is not the case for the data at hand).

When there is substantial interaction, the main effects are often difficult to interpret. Although they are still needed in the model to estimate the interaction effect, they may not have much meaning by themselves because they are based on the assumption that the simple effect of A on S_p is the same across levels of B and the simple effect of B on S_p is the same across levels of A. Of course, researchers can still report the main effects from a *contrast coded* regression model (i.e., two-way ANOVA), but they cannot be understood without the interaction effect. In the experiment in this chapter, for example, the main effect of A is small and not statistically significant; this is because the main effect of A comes down to the average of the difference in A1 versus A0 for B0 (i.e., the lower line in Fig. 11.2; a negative slope) and the difference in A1 versus A0 for B1 (the upper line in Fig. 11.2; a positive slope): that average is close to zero. However, the main effect of B is the difference between the average of the two points connected by the upper line and the average of the two points connected by the lower line; this difference is much larger and statistically significant.

Categorical Outcome

The note on the difference between dummy and contrast coding is important, because not so few researchers automatically apply dummy coding to predictors when they want to perform a regression analysis with a categorical outcome. However, the same contrast coding that is used in two-way ANOVA for quantitative outcome variables can also be used for binary logistic and ordinal logistic regression analysis, and the same holds for the types of contrasts discussed in Chap. 10. The same goes for the model comparison strategy just discussed for quantitative outcome variables. Consider, for example, that we had a dichotomous instead of a quantitative outcome variable. When we have scores, dichotomising them rarely if ever makes sense; it just results in an unnecessary loss of information and statistical power. However, suppose that in our experiment we did not have a

Table 11.2 R^2_{McF}, AIC, and BIC for each of the five competing models (*JASP*) for the dichotomous outcome variable

Model	R^2_{McF}	AIC	BIC
0: null	0.000	328.013	331.493
1: main A	0.005	328.296	335.257
2: main B	0.095	299.018	305.979
3: main A & B	0.101	299.052	309.494
4: full factorial	0.120	294.803	308.726

quantitative score but only a qualitative pass/fail judgement, and the findings were as follows.

In the condition where participants received support but no feedback, 26 of the 60 participants pass. In the condition where participants received neither support nor feedback, 23 of the 60 participants pass. In the condition where participants received support and feedback, 39 of the 60 participants pass. In the condition where participants received support but no feedback, 52 of the 60 participants pass. Table 11.2 presents R^2_{McF} along with AIC and BIC for each of the five competing models for the data at hand.

For the interaction effect, we find an *OR* of 4.306 with a 95% CI of [1.337; 13.862] and a 90% CI of [1.614; 11.486]. The *p*-value of the interaction effect is 0.014. In other words, researchers who prefer BIC (or a statistical significance test at a stricter significance level, such as $\alpha = 0.01$ or $\alpha = 0.005$; for main effect B, $p < 0.001$) will prefer Model 2, whereas researchers who prefer AIC or a statistical significance test at $\alpha = 0.05$ will prefer Model 4. An advantage of presenting the outcomes of an experiment as in Tables 11.1 and 11.2 is that readers can decide for themselves what model they would prefer.

Three Factors

Two-way designs are very common in educational and psychological research and, to some extent, three-way designs are as well. However, two-way and three-way designs are not always treated as they should be treated; not so few researchers treat them as one-way designs. This is unfortunate, because treating a two-way or three-way design as a one-way design disables us to formally test interaction effects and comes with a substantial or even severe loss of statistical power (e.g., Leppink, O'Sullivan, & Winston, 2017). One recent example from experimental educational research is found in studies that focus on the effects of worked examples of learning outcomes as measured through immediate post-testing. Consider four possible conditions: problem-problem, problem-example, example-problem, and example-example.

In the problem-problem condition, participants try to solve two problems A and B that follow the same structure and are of (more or less) the same difficulty. In the problem-example condition, participants first try to solve problem A and then study a worked example of problem B. In the example-problem condition, participants first study a worked example of problem A and next try to solve problem B. Finally, in the example-example condition, participants study worked examples of both problems, meaning they solve nothing autonomously. Right after study, participants individually complete the same post-test consisting of ten problems of the same difficulty yielding a total post-test score ranging from 0 (all incorrect) to 10 (all correct; 1 point for each correctly solved problem). Researchers who analyse the data as one-way typically start with a one-way ANOVA and follow up with post-hoc comparisons with Bonferroni correction. As seen in Chap. 9, this constitutes an inefficient way of analysing data when specific hypotheses are available. In this example, comparisons ought not to be made with separate sets of two groups (as in Chap. 9) but through *combinations* of groups (i.e., contrasts).

In two-way ANOVA, three types of contrasts are made: *main effect of first task*, *main effect of second task*, and *interaction effect*. The main effect of *first* task comes down to comparing: [the average of problem-problem and problem-example] versus [the average of example-problem and example-example]. The main effect of *second* task comes down to the following comparison: [the average of problem-problem and example-problem] versus [the average of problem-example and example-example]. Finally, the interaction effect is found in the third possible contrast: [the average of problem-problem and example-example] versus [the average of problem-example and example-problem]. In other words, each contrast involves a comparison of *M*s of two groups of 2*n* each (given *n* per condition). In the post-hoc follow-up on one-way ANOVA, however, each comparison of a set of two conditions is one of the *M*s of two groups of *n* each. Next, a Bonferroni correction for multiple testing is added, and the statistical power is reduced even further; all in all, the *SE* for comparison in the Bonferroni-corrected post-hoc testing is more than 1.7 times the *SE* in two-way ANOVA). The difference between the correct two-way ANOVA and the incorrect one-way ANOVA is explained in more detail by Leppink et al. (2017).

Some researchers may argue that if the question is which *strategy* results in best performance, motivation or whatever outcome variable is of interest, example-example, example-problem, problem-example, and problem-problem can be perceived as four different strategies and should therefore be treated as one-way. The answer to this is *no*, because any pattern of differences between these four strategies can be captured through one or two main effects and/or through the interaction in a two-way analysis. For instance, if one strategy stands out from the other strategies and the latter do not really differ from one another, this should be reflected in the interaction effect. The only difference is that a one-way ANOVA will come at the cost of a loss of statistical power and precision. For instance, given four conditions of $n = 32$ each, a two-way ANOVA has a statistical power of about 0.80 for a medium size difference ($f = 0.25$; *GPower*) for each of the aforementioned contrasts that constitute the interaction and main effects, respectively, whereas one-way ANOVA has a power of only about 0.64, and that power goes down further when researchers apply a correction for multiple testing (i.e., a power of about 0.47 when applying Bonferroni correction).

Some other researchers may argue that the one-way approach is more easily generalised when the sequence has more than two tasks, but again the answer is *no*: a multi-factor ANOVA is then still more appropriate. Let us look at this for a sequence of three tasks.

Different Sequences

A total of $N = 240$ participants is randomly assigned to eight conditions that differ in their sequences of problems (P) and examples (E) studied for the consecutive tasks A, B, and C (the order in these tasks being the same across conditions): PPP ($n = 30$), PPE ($n = 30$), PEP ($n = 30$), PEE ($n = 30$), EPP ($n = 30$), EPE ($n = 30$), EEP ($n = 30$), and EEE ($n = 30$). With these eight sequences, we can estimate the following effects: the main effect of *A* (first task), the main effect of *B* (second task), the main effect of C (third task), the A-by-B interaction, the A-by-C interaction, the B-by-C interaction, and the A-by-B-by-C interaction. The latter is the *three-way* interaction. Post-test performance (0–10) in the various conditions is follows: PPP: $M = 3.067$, $SD = 0.944$; PPE: $M = 4.000$, $SD = 0.983$; PEP: $M = 3.800$, $SD = 0.887$; PEE: $M = 5.967$, $SD = 1.033$; EPP: $M = 5.167$, $SD = 1.085$; EPE: $M = 6.033$, $SD = 0.999$; EEP: $M = 7.200$, $SD = 0.847$; and EEE: $M = 6.067$, $SD = 1.143$.

Figure 11.3 presents the *M*s of each of the eight conditions with their 95% CIs

Contrary to two-way ANOVA, three-way ANOVA is best understood with *two* plots. In the case of a two-way interaction, the effect of *A* on an outcome variable of interest depends on the level of B. In the case of a three-way interaction, the aforementioned A-by-B two-way interaction depends on the level of C. Figure 11.3

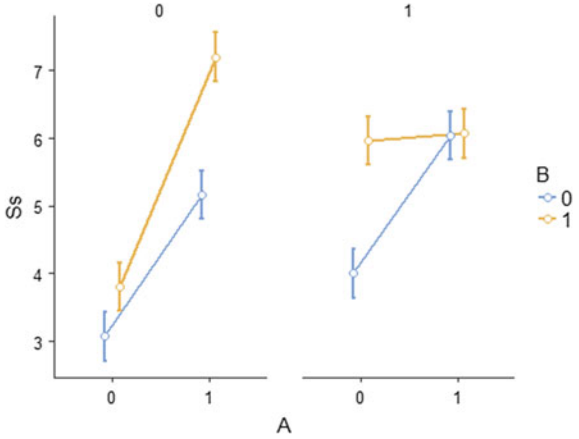


Fig. 11.3 *M*s and 95% CIs around these *M*s for each of the eight conditions ($A = 0$: first task = P; $A = 1$: first task = E; $B = 0$: second task = P, lower line in both plots; $B = 1$: second task = E, upper line in both plots; $C = 0$: third task = P, left plot; $C = 1$: third task = E, right plot); *Ss* represents the score (*Jamovi*)

hints at such a three-way interaction effect. In the left plot (i.e., C = 0: first task = P), we see that the expected difference in post-test performance for a learner whose second task was an example versus a problem is larger for learners who started with an example (i.e., $A = 1$: first task = E) than for learners who started with a problem (i.e., $A = 0$: first task = P). However, in the right plot (i.e., C = 1: first task = E), we see that the pattern is the other way around: there is hardly any difference between the second task being a problem or an example for learners who started with an example, but there is a big difference in favour of the second task being an example for learners who started with a problem. Besides, the *M*s for $A = 1$ (i.e., first task being an example) are consistently higher than the *M*s for $A = 0$ (i.e., first task being a problem), so at least at first glance, the findings appear to indicate that starting with an example tends to yield better performance than starting with a problem.

Different Effects and Models

Researchers who choose a one-way ANOVA with Bonferroni-corrected post-hoc comparisons follow up will find themselves in trouble: with eight conditions, the post-hoc stage involves a total of 28 comparisons! A three-way ANOVA, on the contrary, enables researchers to simultaneously estimate the aforementioned three main and four interaction effects (here: $F_{1,\ 232}$ for each effect). Table 11.3 presents the F-values and p-values along with η^2-, partial η^2-, and ω^2-values for each of the effects.

In other words, we appear to be dealing with a medium size ($\eta^2 = 0.060$) three-way interaction effect. The largest effect is the main effect of A and is easy to understand from Fig. 11.3 (i.e., $A = 1$ consistently having higher *M*s than $A = 0$). Table 11.4 presents the R^2 of each possible model along with AIC and BIC (A, B, and C represent main effects, while AB, AC, BC, and ABC represent interaction effects).

Both AIC and BIC clearly prefer Model 18, the full factorial three-way model. Some readers may wonder if this collection of nineteen models indeed constitutes

Table 11.3 F-values and p-values along with η^2-, partial η^2-, and ω^2-values for each of the effects in three-way ANOVA (*JASP*)

Effect	$F_{1,\ 232}$	p-value	η^2	partial η^2	ω^2
Main A	220.949	<0.001	0.337	0.488	0.335
Main B	86.157	<0.001	0.131	0.271	0.130
Main C	30.441	<0.001	0.046	0.116	0.045
A-by-B	1.521	0.219	0.002	0.007	0.001
A-by-C	42.980	<0.001	0.066	0.156	0.064
B-by-C	2.229	0.137	0.003	0.010	0.002
A-by-B-by-C	39.643	<0.001	0.060	0.146	0.059

Table 11.4 R^2, AIC, and BIC for each of the nineteen competing models in three-way ANOVA (*Jamovi*, *Mplus*)

Model	R^2	AIC	BIC
0: null	0.000	923.716	930.677
1: A	0.337	827.134	837.576
2: B	0.131	891.920	902.361
3: A, B	0.468	776.156	790.078
4: A, B, AB	0.471	777.107	794.510
5: C	0.046	914.311	924.753
6: A, C	0.383	811.721	825.644
7: B, C	0.178	880.742	894.664
8: A, B, C	0.515	756.240	773.643
9: A, B, AB, C	0.517	757.091	777.975
10: A, C, AC	0.449	786.763	804.166
11: A, B, C, AC	0.580	723.435	744.318
12: A, B, AB, C, AC	0.582	724.105	748.470
13: B, C, BC	0.181	884.748	899.151
14: A, B, C, BC	0.518	756.554	777.438
15: A, B, AB, C, BC	0.520	757.397	781.761
16: A, B, C, AC, BC	0.584	723.484	747.849
17: A, B, AB, C, AC, BC	0.586	724.144	751.989
18: A, B, AB, C, AC, BC, ABC	0.646	688.284	719.610

the full set of possible models. Indeed, other models are not possible. For instance, we cannot have a model with AC (i.e., the A-by-C interaction) without having A and C in the model as well. Likewise, to estimate ABC (i.e., the three-way interaction), we need all underlying main and two-way interaction effects in the model as well.

A Note on Effect Size

As mentioned in Chap. 9, η^2 and ω^2 are common measures of effect size in ANOVA, with the latter being slightly less biased but with a difference slowly going to zero with increasing sample sizes (Howell, 2010, 2017). The η^2 of an effect is easily computed by hand by dividing the sum of squares (*SS*) of that effect by the total *SS* (of all effects in the model plus residual). In the three-way experiment, for instance, the total *SS* (SS_{T}) is the sum of the following eight *SS*:

$$SS_{\mathrm{T}} = SS_{\mathrm{A}} + SS_{\mathrm{B}} + SS_{\mathrm{C}} + SS_{\mathrm{AB}} + SS_{\mathrm{AC}} + SS_{\mathrm{BC}} + SS_{\mathrm{ABC}} + SS_{\mathrm{Residual}}.$$

In our three-way experiment, we find the following *SS*: SS_{A} = 218.504; SS_{B} = 85.204; SS_{C} = 30.104; SS_{AB} = 1.504; SS_{AC} = 42.504; SS_{BC} = 2.204; SS_{ABC} = 39.204; SS_{Residual} = 229.433; SS_{T} = 648.663 (the difference in the third decimal between SS_{T} and the total of the other *SS* is due to round-off error). The η^2 of an effect, η^2_{E}, is:

$$\eta^2_{\mathrm{E}} = SS_{\mathrm{E}}/SS_{\mathrm{T}}.$$

Hence, for the main effect of *A* we find η^2 = 0.337, and for the three-way interaction we find η^2 = 0.060. Contrary to η^2 of an effect, the partial η^2 of an effect is based on *that part of* SS_{T} that cannot be explained by other effects in the model. Consider the difference between η^2 and partial η^2 as cutting a pie. Suppose, we cut a pie into eight pieces for each of the different effects (i.e., seven in this example) and the residual (i.e., what cannot be explained by any of the effects in the model). These pieces are not equally big, since some effects explain more than other effects, and the residual piece is biggest because about 35.4% (i.e., $SS_{\mathrm{Residual}}/SS_{\mathrm{T}}$) remains unexplained. Now, η^2 (i.e., η^2_{E} in the formula) expresses what is the proportion of the surface of the original (uncut) pie that is explained by a given effect. However, partial η^2 works like this. Suppose, we are interested in the partial η^2 of the three-way interaction. We take out from the cut pie *all* pieces of the main effects and two-way interactions, so only the piece of A and the piece of residual are left on the plate:

$$\eta^2_{\mathrm{EP}} = SS_{\mathrm{E}}/\left[SS_{\mathrm{E}} + SS_{\mathrm{Residual}}\right].$$

For the three-way interaction, we find 39.204/229.433 ≈ 0.146. The difference between the standard η^2 (i.e., η^2_{E}) and the partial η^2 (i.e., η^2_{EP}) is important for several reasons. Firstly, the interpretation is different; while η^2_{E} responds to the question what proportion of SS_{T} can be explained by effect E, η^2_{EP} responds to the question what proportion of the *rest of* SS_{T}, after cutting out all other effects in the model. In other words, the latter is about the proportion of that part of SS_{T} that cannot be explained by other effects in the model that *can* be explained by effect E. Statistical software packages do not always correctly label η^2; for instance, although in the *SPSS* version used in this book, 'Partial Eta Squared' correctly refers to η^2_{EP}, in some previous versions it stated 'Eta Squared' where it actually provided η^2_{EP}, and mistaking η^2_{EP} for η^2_{E} may have led not so few *SPSS* users to overestimate their effects of interest (e.g., Levine & Hullett, 2002). Luckily, statistical packages such as *Jamovi* and *JASP* include *both* variants of η^2 (as well as ω^2) with their correct labels, so that this kind of overestimation is easily avoided. Finally, for statistical power and required sample size calculations, we use partial η^2 and partial ω^2.

The Bridge

In line with the QDA heuristic introduced in Chap. 1, two-way designs call for two-way analysis, three-way designs call for three-way analysis, et cetera. Analysing two- or multi-way data as if we were dealing with one-way data comes at cost of no longer being able to formally test interaction effects and with a substantial or even severe loss of statistical power. When dealing with two- or multi-way data, there is no question that can be studied with one-way analysis that cannot be studied with the much more appropriate alternative. Some readers may wonder if we should not apply a correction for multiple testing since we are testing several effects at the same time. The answer to this is: *no*, because the rationale behind a correction is to keep the *family-wise* error rate at a given level. For the main effect of a factor, the conditions that constitute that factor constitute one *family*. The same holds for other main effects as well as for interaction effects. In short, each effect can be conceived as one family. The only case in which researchers might want to consider some correction for multiple testing is if a factor has more than two levels and therefore a follow-up analysis on a statistically significant main or interaction effect entails a comparison of more than two conditions. In such a case, the same logic as in Chap. 9 applies, unless the availability of one or more specific hypotheses justifies one of the approaches discussed in Chap. 10, but the latter might also have implications for how to test the interaction effect (i.e., any kind of interaction pattern or an expected specific type of interaction pattern, see also Chap. 12, Experiment 6).

Factor-Covariate Combinations

12

Abstract

The examples in the previous chapters all revolve around the effects of one (Chaps. 2 and 5–10) or two or three (Chap. 11) categorical treatment factors on a categorical or quantitative outcome variable. However, in not so few cases, we have at least one other, non-treatment variable measured along with our primary outcome variable of interest. How to deal with this additional variable depends on whether it is measured before or after the start of treatment and how it relates to the primary outcome variable of interest. When this additional quantitative variable is measured together with the primary outcome variable, we may in some cases treat this additional variable as a second outcome variable that is correlated with our primary outcome variable. When this additional variable is measured after the start of treatment but before our primary outcome variable, the additional variable may mediate treatment effects of interest. Finally, when the additional variable is measured before the start of the treatment, we may under conditions outlined in this chapter include the additional variable as a covariate. Both covariates and mediators may moderate a treatment effect of interest.

Introduction

How to deal with observed variables that are neither treatment factors nor outcome variables of direct interest first of all depends on theory. Including variables to our statistical models in a purely data-driven, a-theoretical manner may contribute to an increased likelihood of findings not being replicable more than anything else, especially when samples are fairly small, since findings with regard to any variable fluctuate from sample to sample. Therefore, such a data-driven approach to variable inclusion, which has also been called *covariate fishing* (e.g., Gruijters, 2016) is

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_12

generally not recommended. Besides, even if adding a variable to our model can be defended based on theory or common sense, how to treat this variable depends on when this variable is measured and what is its expected relation to our primary outcome variable of interest. In this chapter, we focus on different types of such ‘additional variables’, based on when they are measured and how they relate to the primary outcome variable of interest, and how they can be treated in our models: as an *additional outcome* variable that is correlated with our primary outcome variable of interest (i.e., *common response*: two correlated outcome variables commonly respond to a treatment variable; Experiment 1); as an additional predictor variable aka *covariate* (Experiments 2, 3, and 6); as a *mediator* (Experiment 4); and as a *moderator* (Experiment 5).

Experiment 1: Correlated Outcome Variables

In some cases, it makes sense to add an additional variable as a second outcome variable that is correlated with the primary outcome variable of interest. For example, suppose that in a two-group experiment ($N = 130$, $n = 65$ per condition) on the effect of a particular type of instruction in simulation training on the effort invested in a task we ask participants to self-rate both the *effort* they invested in the (condition-specific) task just completed and the experienced *difficulty* of that task, say both on a continuous VAS ranging from 0 (min) to 100 (max). Although we are primarily interested in the effect of condition on effort, difficulty and effort are conceptually related and are therefore likely correlated: effort is likely to increase with difficulty. Figure 12.1 presents the scatterplot of the relation between difficulty (D) and effort (E) for each of the two conditions ($X = 0$: control, $X = 1$: treatment).

In both conditions, the relation between difficulty and effort can be reasonably summarised in linear terms. In the control condition, we find a Pearson’s r of 0.621, and in the treatment condition we find $r = 0.636$. In other words, in both conditions, around 40% of the variance in effort can be explained by difficulty, and vice versa. This correlation is not so high that we may treat difficulty and effort as perhaps measuring the same thing (i.e., $r > 0.80$) but is not in a range where we might call difficulty and effort being more or less independent (e.g., $0 < r < 0.15$) either.

Figure 12.2 provides so-called *quantile plot* to assess *multivariate normality* of the residuals (e.g., Field, 2018), the variant of normally distributed residuals when more than a single outcome variable is involved.

The more the dots deviate from the straight line, the more they deviate from multivariate normality. In this case, the deviations are not so strong. As mentioned in earlier chapters in this book, I am generally not a big fan of statistical significance tests as a way to check assumptions, but Shapiro–Wilk’s multivariate normality test is not statistically significant at any meaningful significance level (i.e., 1, 5, or 10%), $W = 0.990$, $p = 0.431$.

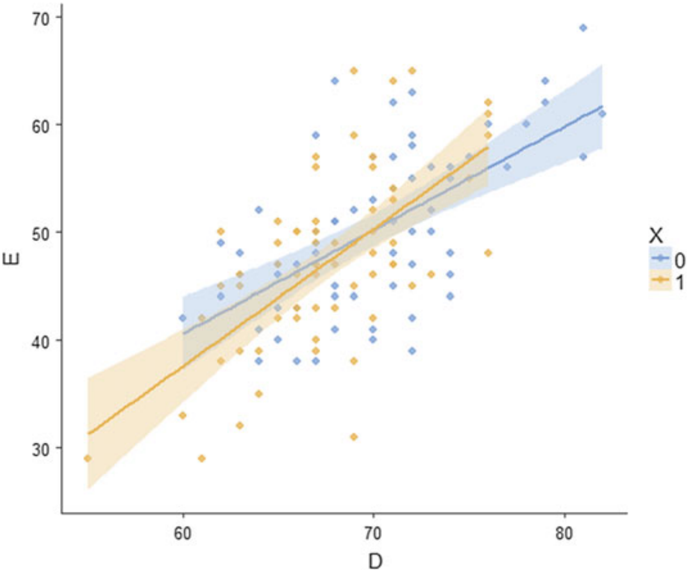
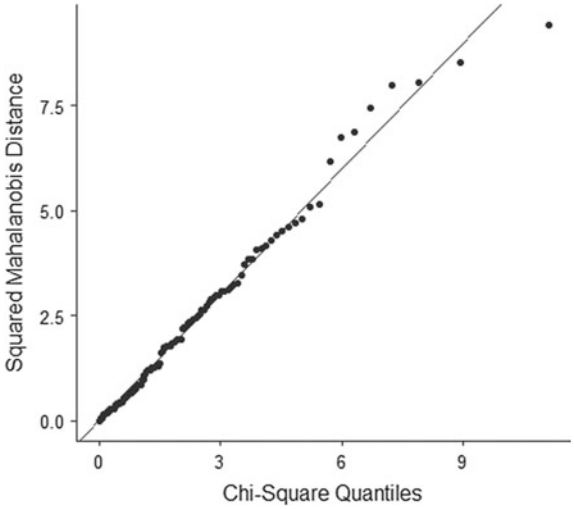


Fig. 12.1 Scatterplot of the relation between difficulty (D) and effort (E) for the control (X = 0) and treatment (X = 1) condition in Experiment 1 (*Jamovi*)

Fig. 12.2 Quantile plot to assess multivariate normality of the residuals in Experiment 1 (*Jamovi*)



Apart from multivariate normality, we can inspect the homogeneity of the variance-covariance matrix across conditions: the matrix of the variances of the different outcome variables and their covariances. In the case of two outcome variables, the variance-covariance matrix consists of two variances (i.e., the diagonal of the matrix) and one covariance (i.e., off-diagonal, there is only one pair of variables to be correlated). In our case, the variances (and in square root form: *SD*) are 23.947 ($SD = 4.894$) for difficulty and 57.248 ($SD = 4.127$) for effort in the control condition and 17.035 ($SD = 7.566$) for difficulty and 67.994 ($SD = 8.246$) for effort in the treatment condition. The covariance is 23.947 in the control condition and 21.631 in the treatment condition. Box's homogeneity of covariance matrices test (Box, 1949) is not statistically significant at any meaningful significance level (i.e., 1, 5, or 10%): $\chi^2_3 = 3.709$, $p = 0.295$ (*Jamovi*).

Under these conditions, we can reasonably treat difficulty and effort as two correlated outcome variables in a multivariate linear regression model that is also called *multivariate* ANOVA (MANOVA; e.g., Field, 2018). Examples of statistical packages that can do MANOVA include *SPSS*, *Jamovi*, and *Stata*, with the latter also including multivariate follow-up tests which may be useful for treatment factors that have more than two levels. For the data at hand, we find $F_{2,\ 127} = 4.569$, $p = 0.012$. *SPSS* also provides a partial η^2-estimate along with this multivariate test, and in the case of a single treatment factor, η^2 and partial η^2 in are the same. In our case, we find $\eta^2 = 0.067$, which corresponds with a medium size effect.

With MANOVA, we need one multivariate instead of several separate (i.e., univariate) ANOVAs and generally gain some statistical power relative to ANOVA. The rationale behind MANOVA and a single η^2-estimate makes sense if the variables are conceptually related, are reasonably correlated, yield residuals that do not deviate too much from multivariate normality, and constitute a variance-covariance matrix that is more or less the same across conditions. If correlations between outcome variables are much smaller (e.g., smaller than 0.30) MANOVA as an alternative to separate ANOVAs becomes more difficult to defend and the outcomes will be more difficult to understand. Simultaneously, if correlations between outcome variables are high (i.e., over 0.80), we may as well compute one composite score for the set of outcome variables together.

Experiment 2: Covariate in a Linear Model

In cases where the additional variable is not measured along with the primary outcome variable of interest but *before* the start of the experiment (i.e., before the start of the treatment), the MANOVA approach does not make sense. After all, given random assignment, variables measured before treatment are not affected by treatment. In such a case, if we have good reasons to include a variable measured prior to treatment in our models, we can add that variable as a *covariate*. If this variable is a categorical variable, this covariate can be included in the form of an additional factor; if it is a quantitative variable, it can be added as a quantitative covariate. In the simplest case, we do not have to worry about treatment-by-covariate interactions. However, we always need to check that assumption.

Consider the following example. A total of $N = 200$ randomly sampled Bachelor of Science students are randomly allocated to an experiment in which two different methods of learning inferential statistics are compared in terms of post-test score

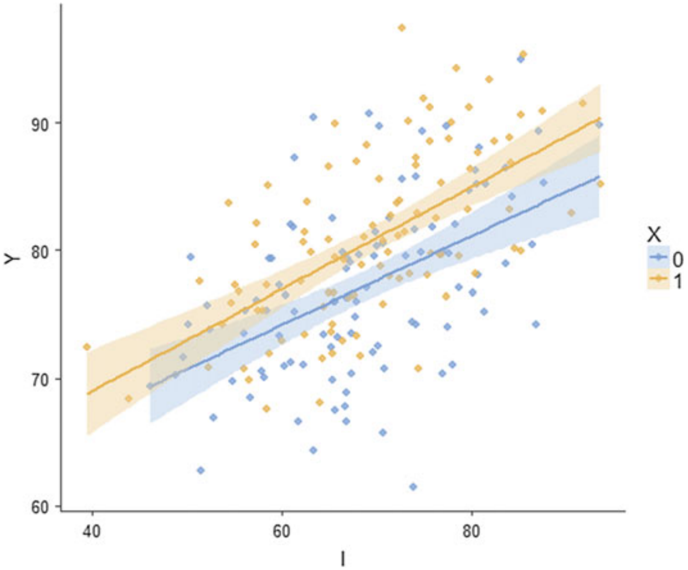


Fig. 12.3 Scatterplot of the relation between interest (I) and post-test score (Y) for the control (X = 0) and treatment (X = 1) condition in Experiment 2 (*Jamovi*)

(0–120). Prior to the start of the experiment, all learners are asked to self-rate their interest in inferential statistics on a 0–100 VAS. Figure 12.3 presents the scatterplot of the relation between interest (I) and post-test score (Y) for the control (X = 0) and treatment (X = 1) condition.

Interest is on average 68.431 (SD = 10.178) in the control condition and 69.646 (SD = 10.370) in the treatment condition, and post-test performance is on average 77.153 (SD = 6.981) in the control condition and 80.874 (SD = 6.723) in the treatment condition. As expected, average interest is very similar for the two conditions. In terms of post-test performance, the difference is a bit over 0.5 SD. Post-test score and interest correlate considerably: r = 0.506 in the control condition, and r = 0.615 in the treatment condition.

The regression lines in Fig. 12.3 are more or less parallel, indicating that the factor-by-covariate interaction is more or less zero. The magnitude of the interaction effect has consequences for the statistical modelling of treatment effects (e.g., Leppink, 2018a, b). A nice advantage of this kind of situation is that the difference between the regression lines, which constitutes the effect of the treatment factor, is more or less the same across the range of the covariate. As in the case of two- or multi-factor ANOVAs, speaking of main effects tends to make little sense when there is substantial interaction. Whether or not we need a model with interaction effect can be examined through a comparison of competing models as explained in Chap. 11. Given one treatment factor and one covariate, the five competing models are:

Table 12.1 R^2, AIC, and BIC of the five competing models in Experiment 2 (*Mplus*)

Model	R^2	AIC	BIC
0: null	0.000	1353.809	1360.406
1: T	0.069	1341.446	1351.341
2: C	0.307	1282.362	1292.257
3: T and C	0.361	1268.363	1281.556
4: T, C, and T-by-C	0.362	1269.924	1286.415

Model 0: null model;
Model 1: only treatment effect (T);
Model 2: only an effect of the covariate (C);
Model 3: both treatment and covariate effect (T and C); and
Model 4: T and C as well as the T-*by*-C interaction effect.

Model 0 uses the observed *M* post-test score as the predicted post-score for any individual, Model 1 comes down to a two-samples *t*-test, Model 2 is a simple linear regression, Model 3 (i.e., main-effects only) is also known as ANCOVA, and Model 4 is also referred to as moderated regression. Table 12.1 presents the R^2 along with AIC and BIC for each of these five models.

Both AIC and BIC prefer Model 3 and from the differences in R^2 it is easy to understand why. Model 3 (ANCOVA) provides outcomes of both main effects. For the treatment effect, we find a standardised β of 0.231 with a 95% CI of [0.119; 0.344] and a 90% CI of [0.137; 0.325]. For the effect of the covariate, we find a standardised β of 0.541 with a 95% CI of [0.428; 0.653] and a 90% CI of [0.446; 0.635] (*Jamovi*).

Experiment 3: Covariate in a Categorical Outcome Model

The approach used in Experiment 2 is also applicable to models for dichotomous, multicategory nominal and ordinal outcome variables. For example, in Chap. 6, we deal with an experiment where the outcome variable of interest is a *choice*. Suppose that some researchers do a follow-up experiment on the experiment presented in Chap. 6, with slightly larger numbers: they randomly assign $N = 400$ citizens of age group 18–75 years old to either of two conditions. In the control condition ($n = 200$), participants see a 10 min YouTube video on a specific contemporary question in EU politics. In the treatment condition ($n = 200$), participants see a 10 min video that covers the same content as the video in the control condition but presents that content in a slightly different way. In both conditions, immediately after the video, participants are asked to choose which of four different words describes best how they feel about the EU after the video: *indifferent*, *embarrassed*, *surprised* or *disappointed*.

Table 12.2 R^2_{McF}, AIC, and BIC of the five competing models in Experiment 3 (*Jamovi*)

Model	R^2_{McF}	AIC	BIC
0: null	0.000	1024.186	1036.160
1: T	0.010	1019.829	1043.778
2: C	0.020	1009.643	1033.592
3: T and C	0.030	1005.529	1041.452
4: T, C, and T-by-C	0.032	1009.176	1057.073

However, in the follow-up study, participants are asked prior to the experiment to self-rate their mood on a VAS ranging from −5 (very bad) to +5 (very good). The researchers want to add this variable as a covariate to their model, because they expect that mood can explain some of the choice behaviour. The findings are as follows. The average mood is 0.046 ($SD = 1.092$) in the control condition and 0.080 ($SD = 0.996$) in the treatment condition, very similar as expected since it is not a function of treatment but entirely the result of random assignment. In the control condition, 62 participants (31%) are indifferent, 22 participants (11%) are embarrassed, 24 participants (12%) are surprised, and 92 (46%) are disappointed. In the treatment condition, 41 participants (20.5%) are indifferent, 32 participants (16%) are embarrassed, 40 participants (20%) are surprised, and 87 participants (43.5%) are disappointed. Table 12.2 presents R^2_{McF} along with AIC and BIC for each of the five competing models for the data at hand.

AIC prefers Model 3 while BIC prefers Model 2. In other words, neither of the two criteria prefers Model 4. Indeed, the gain in R^2_{McF} from Model 3 to Model 4 is minimal. Model 3, the equivalent of ANCOVA for multicategory nominal outcome variables, provides outcomes (*b*s) of both main effects. For the treatment effect, we find the following *b*s. For embarrassed versus indifferent, $b = 0.796$, $p = 0.022$, and 95% CI = [0.113; 1.478]. For surprised versus indifferent, $b = 0.923$, $p = 0.005$, and 95% CI = [0.275; 1.571]. For disappointed versus indifferent, $b = 0.359$, $p = 0.161$, and 95% CI = [−0.143; 0.860]. For the covariate, we find the following outcomes. For embarrassed versus indifferent, $b = 0.611$, $p < 0.001$, and 95% CI = [0.272; 0.949]. For surprised versus indifferent, $b = 0.433$, $p = 0.007$, and 95% CI = [0.118; 0.748]. For disappointed versus indifferent, $b = 0.485$, $p < 0.001$, and 95% CI = [0.241; 0.730]. In line with previous examples, we could also provide the 90% CIs but they are not needed in this example; the goal of this example is to illustrate that models with covariates are not limited to quantitative outcome variables but work for categorical outcome variables as well. Had Model 4 been the preferred one, we would have calculated three additional *b*s: one for the factor-by-covariate interaction for embarrassed versus indifferent, one for that interaction for surprised versus indifferent, and one for that interaction for disappointed versus indifferent. In Model 3, we do provide a *b* for each of embarrassed versus indifferent, surprised versus indifferent, and disappointed versus indifferent, but we treat each of these *b*s as equal across conditions.

Experiment 4: Mediation

In Experiment 1, the additional variable is measured along with the primary outcome variable of interest, and in Experiments 2–3 it is measured prior to the start of the treatment. In other cases, the additional variable is measured after the start of the treatment but before the outcome variable of interest. In this kind of case, it is an outcome variable of the treatment effect and a predictor variable of the outcome variable of interest. In other words, the additional variable is an *intermediate* variable in a causal chain (e.g., Hayes, 2018; Leppink, 2015b, 2017). If the treatment results in differences on this intermediate variable and the latter influences the outcome variable of interest, the intermediate variable *mediates* at least part of the treatment. A form of path analysis may then be needed. Path analysis can provide three treatment effects—direct, indirect, and total—that are related as follows:

$$\textit{total effect} = \textit{direct effect} + \textit{indirect effect}.$$

The *total effect* comes down to a two-samples *t*-test or one-way ANOVA for the effect of treatment on the outcome variable of interest. The *direct effect* is what could be obtained through an ANCOVA in which the additional variable is treated as covariate. Not so few researchers think that what they report from the ANCOVA is the treatment effect 'corrected for' the third variable instead of the direct effect. This is like treating a mediator as a confounder, and we erase part of our treatment effect as if it did not exist (e.g., Leppink, 2017a). Finally, the *indirect effect* is that part of the total effect of treatment that can be explained by the mediator, that is: the part of the treatment effect that is mediated by the additional variable in the model.

Since the treatment effect of interest is the *total effect*, if researchers' interest lies exclusively in the treatment effect and are not really interested in the mediator as one of the *possible* mechanisms of that treatment effect (i.e., *possible* because mediation is necessary but not sufficient evidence of mechanism; e.g., Tryon, 2018), a one-way analysis will do. However, whenever researchers *are* interested in the role of the mediator, path analysis constitutes a more meaningful approach than just an ANCOVA. That said, the indirect effect can be calculated once the total effect from a one-way analysis and the direct effect obtained through ANCOVA are known. In other words, an appropriate combination of ANOVA and ANCOVA can yield the same estimates of total, direct, and indirect treatment effect as the path analysis. However, it is good to keep in mind that ANCOVA does not provide the total treatment effect or some kind of 'confounding-corrected' treatment effect.

Let us look at an example. A total of N = 150 high school students are randomly allocated to control (n = 75) and treatment (n = 75) condition in an experiment on training grammar of a foreign language in an online learning environment. After a condition-specific practice session, participants individually rate on a VAS from 0 (min) to 100 (max) the effort invested in the practice session. Five minutes later, they are presented a post-test that consists of 100 sentences that they have to complete. Each correctly completed sentence generates 1 point, hence a participant's post-test score is somewhere between 0 (all incorrect) and 100 (all correct).

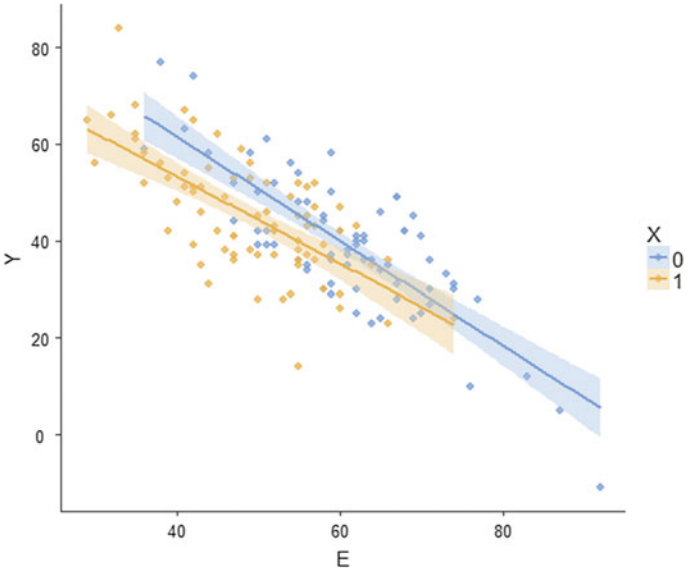


Fig. 12.4 Scatterplot of the relation between effort (E) and post-test score (Y) for the control ($X = 0$) and treatment ($X = 1$) condition in Experiment 4 (*Jamovi*)

The average effort is 60.880 ($SD = 10.520$) in the control condition and 48.947 ($SD = 9.228$) in the treatment condition, and the average post-test score is 39.000 ($SD = 14.069$) in the control condition and 45.267 ($SD = 12.240$) in the treatment condition. Figure 12.4 presents the scatterplot of the relation between effort (E) and post-test score (Y) for the control ($X = 0$) and treatment ($X = 1$) condition. Pearson's r for the linear relation between effort and score is $r = -0.805$ in the control condition and -0.676 in the treatment condition.

Contrary to what we see in Experiments 2 and 3, the conditions do not differ only in average post-test score but in average effort as well. For post-test score, we find: $t_{148} = 2.910$, $p = 0.004$, Cohen's $d = 0.475$, 95% CI of $d = [0.150; 0.799]$, and 90% CI of $d = [0.202; 0.747]$. For effort, we find: $t_{148} = -7.385$, $p < 0.001$, $d = -1.206$, 95% CI of $d = [-1.552; -0.856]$, and 90% CI of $d = [-1.496; -0.912]$. In other words, a medium size treatment effect on post-test score and a (very) large treatment effect on effort. Table 12.3 presents R^2 along with AIC and BIC for each of the five competing models for the data at hand.

Table 12.3 R^2, AIC, and BIC of the five competing models in Experiment 4 (*Mplus*); in Models 2–4, M stands for mediator

Model	R^2	AIC	BIC
0: null	0.000	1209.773	1215.794
1: T	0.054	1203.425	1212.457
2: M	0.553	1090.854	1099.886
3: T and M	0.586	1081.631	1093.673
4: T, M, and T-by-M	0.590	1082.088	1097.141

AIC and BIC indicate a preference for Model 3, which includes the treatment factor and mediator but not their interaction effect. In other words, we do not need to treat the mediator as a moderator of the treatment effect (that would be Model 4). Now, the *total treatment effect* can be found through Model 1: $R^2 = 0.054$. Model 1 comes down to the two-samples *t*-test for post-test score just reported. Model 3 presents how much of the variance in post-test score treatment and effort (but *not* their interaction, that would be Model 4) can explain together. Going from Model 2 to Model 3, the difference in R^2 is considerably smaller than the 0.054 from Model 1 (i.e., going from Model 0 to Model 1). This is because treatment has a strong effect on effort.

Several software packages can be used to calculate the percentage of mediation of the treatment effect, including *Jamovi*, which allows one to use bootstrapping as well. For the data at hand, *Jamovi* estimates that nearly 68% of the total treatment effect is mediated by effort. Moreover, in the path model, the direct effect of treatment is *negative* not positive: $B = -5.654$, $p < 0.001$, 95% CI = [−8.960; −2.345], 90% CI = [−8.423; −2.884]. The rationale behind this is that the treatment *lowers* effort and that is a good thing because higher effort predicts lower post-test performance. Consequently, the direct effect of treatment, which is what researchers would obtain with an ANCOVA (Model 3), is negative even though the total effect is positive. It appears that the effect of the treatment to quite a large extent is found in a reduction of effort during practice. However, researchers who only report ANCOVA thinking they are controlling for a ‘confounder’ called effort would report a negative ‘treatment’ effect, while what they were to report is actually only that part of the treatment effect that cannot be explained in terms of effort reduction, and the total treatment effect—which is ultimately of interest—is positive.

Experiment 5: Moderation

In some cases, the additional variable *moderates* the treatment effect. Suppose, the findings of Experiment 2 would have been as presented in Fig. 12.5 and Table 12.4.

Both AIC and BIC indicate a preference for Model 4, and Fig. 12.5 indicates that the treatment effect is small for low levels of interest but increases with increasing interest. Intuitively, this makes sense; students who are not interested in a subject at all tend to be less motivated to invest in learning that subject than students who have a genuine interest in the subject. In this case, the methods used in Experiments 1–4 fall short because they fail to account for this interaction effect. Instead, we need an approach similar to factorial ANOVA (see Chap. 11) to evaluate the treatment effect at different levels of the covariate: a *picked-points analysis* (Huitema, 2011) aka *pick-a-point analysis* (Hayes, 2018). A common approach, which is implemented in several software packages (e.g., Hayes, 2017; *Jamovi*), is to evaluate the treatment effect at three points: at the *average* of the covariate (cf. ANCOVA), at *one SD below the average* of the covariate, and at *one*

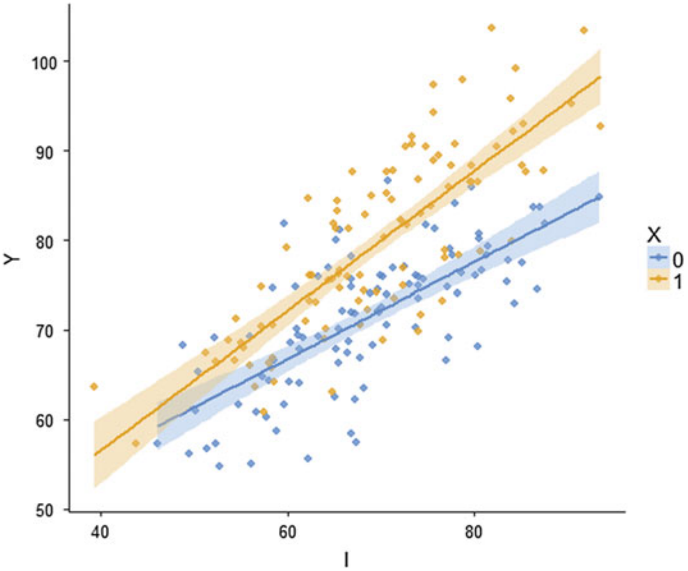


Fig. 12.5 Scatterplot of the relation between interest (*I*) and post-test score (*Y*) for the control ($X = 0$) and treatment ($X = 1$) condition in Experiment 5, an alternative scenario of Experiment 2 (*Jamovi*)

Table 12.4 R^2, AIC, and BIC of the five competing models in Experiment 5 (*Mplus*); in Models 2–4, M stands for moderator

Model	R^2	AIC	BIC
0: null	0.000	1488.960	1495.556
1: T	0.177	1452.036	1461.931
2: M	0.500	1352.258	1362.153
3: T and M	0.644	1286.390	1299.584
4: T, M, and T-by-M	0.659	1279.635	1296.127

SD above the average of the covariate. Figure 12.6 presents the simple slope plot from *Jamovi* resulting from that three-point comparison.

At the average of the covariate, we find: $B = 7.532$, $p < 0.001$, 95% CI = [5.893; 9.171], and 90% CI = [6.157; 8.907]. At average −1*SD* of the covariate, we find: $B = 5.080$, $p < 0.001$, 95% CI = [2.773; 7.387], and 90% CI = [3.144; 7.016]. At average +1*SD* of the covariate, we find: $B = 9.984$, $p < 0.001$, 95% CI = [7.677; 12.291], and 90% CI = [8.048; 11.920]. As can already be seen in Fig. 12.5, the treatment effect is positive at all three points. This may make some readers wonder why bother about the interaction term and not just do ANCOVA. The answer to that is twofold. To start, we are rarely just interested in the question if there is 'a difference' but are usually interested in the magnitude of a difference as well, and in the presence of interaction that magnitude depends on the level of the covariate. Besides, although in this specific case the treatment effect is positive at all

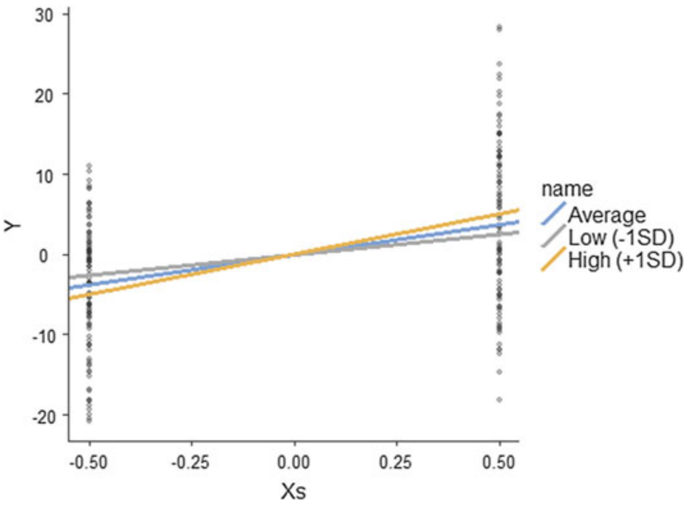


Fig. 12.6 Simple slope plot from picked-points aka pick-a-point analysis: the treatment effect at the average of the covariate as well as at average $-1SD$ and at average $+1SD$, respectively (*Jamovi*)

three points, in other cases it may be negative at one or two points and positive at the other one or two points (e.g., Leppink, 2018b). Therefore, just like a two-way ANOVA model with interaction term is a more appropriate choice than a two-way ANOVA model without interaction term, moderated regression is to be preferred over ANCOVA when the treatment effect of interest depends on the level of the covariate. The non-interaction assumption is an assumption underlying ANCOVA that should be checked even if an interaction is not of primary interest. Speaking of a main treatment effect, hence to pretend that a treatment effect of interest is constant across the range of the covariate, while in fact that treatment effect differs substantially across the range of the covariate, generally does not make sense.

Some have proposed ANCOVRES (e.g., Kane, 2013) as an alternative to ANCOVA when there is interaction or when small samples leave researchers with a fairly low statistical power to detect such an interaction effect. Contrary to ANCOVA, the residuals of the regression of outcome variable on covariate are then not pooled over conditions, but condition-specific residuals are used. Consequently, ANCOVRES does not rely on the non-interaction assumption, and this may result in ANCOVRES having slightly more statistical power than ANCOVA to detect a treatment effect at the average of the covariate. The average of the covariate constitutes the default comparison in both ANCOVA and ANCOVRES. This is a problem, because the interaction remains ignored. Besides, especially when samples are somewhat smaller, using condition-specific regression slopes can be quite tricky. As seen in Chap. 2, smaller samples come with larger deviations between sample and population and with larger fluctuation of estimates from sample to sample. Departures from normally distributed residuals may then also have more severe consequences for the validity of model outcomes, and after we correct the *df* of the residual term for estimating condition-specific slopes instead of a condition-overall slope (Keppel, 1991; Maxwell, Delaney, & Manheimer, 1985; Winer, Brown, & Michels, 1991) the difference in statistical power between ANCOVRES and ANCOVA may well be zero. The only case where ANCOVRES may really be more useful than ANCOVA is when there is a large difference between conditions in variance of the outcome variable and/or in the pattern of variance around the regression line. However, this kind of situation is far more common in non-experimental than in experimental studies.

Experiment 6: Non-linearity

In not so few cases, a linear relation between the additional variable—covariate, mediator, or moderator—and the outcome variable of interest provides a reasonable approach because the scatterplot hints at no substantial deviation from linearity and/or a limited sample size leave us fairly poorly equipped to detect non-linearity. However, there certainly are cases where non-linear models are to be preferred over linear ones (e.g., Leppink & Pérez-Fuster, 2019). How to study and interpret that non-linearity depends not only on the sample size but on the nature of the additional variable as well. Consider the following example. In a two-group experiment with N = 200 Bachelor of Science in Psychology students (n = 100 per condition), two different approaches of learning probability calculus are compared in terms of immediate post-test performance (0–20 points, with 20 being the maximum score). Prior to the experiment, participants are asked to self-rate their prior knowledge of probability calculus on an integer scale from 1 (min) to 5 (max).

In the control condition, prior knowledge is rated as '1' by 7 participants, as '2' by 25 participants, as '3' by 36 participants, as '4' by 19 participants, and as '5' by 12 participants. In the treatment condition, prior knowledge is rated as '1' by 11 participants, as '2' by 24 participants, as '3' by 37 participants, as '4' by 20 participants, and as '5' by 8 participants. Researchers who prefer to treat prior knowledge as a variable of interval level of measurement calculate the *M*s and *SD*s for the two conditions: in the control condition, we find M = 3.060 and SD = 1.118; in the treatment condition, we find M = 2.900 and SD = 1.096. Let us agree for the sake of the example that this approach is justifiable. However, others might argue that prior knowledge is a variable of ordinal level of measurement. If we follow the latter approach, prior knowledge can be added as a second factor in a two-way ANOVA. If we consider prior knowledge of interval level of measurement, we can treat it either as a second factor in two-way ANOVA—after all, it is *still* an integer variable—or as a quantitative covariate. Either way allows us to do the model comparison as in Experiments 2–5 (Models 0–4). Let us take a closer look at both options.

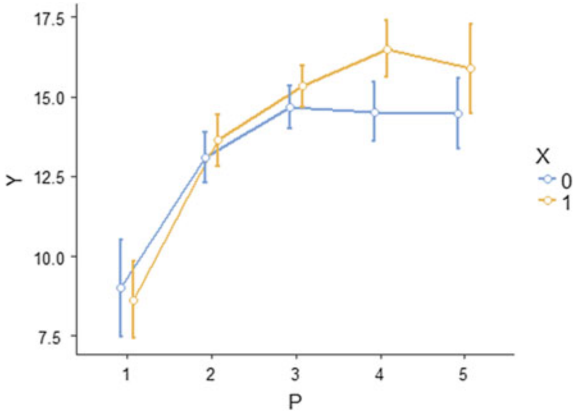


Fig. 12.7 *EMM* plot of a two-way ANOVA treating condition ($X = 0$: control, $X = 1$: treatment) and prior knowledge (P: 1, 2, 3, 4, 5) as factors that may have an interaction effect on post-test score (Y) in Experiment 6 (*Jamovi*)

Figure 12.7 presents the *EMM* plot of a two-way ANOVA treating condition ($X = 0$: control, $X = 1$: treatment) and prior knowledge (P: 1, 2, 3, 4, 5) as factors that may have an interaction effect on post-test score (Y) and Fig. 12.8 demonstrates a quantile plot of the residuals.

The quantile plot indicates no substantial deviations from normality. For the interaction effect, we find $F_{4,\ 190} = 1.347$, $p = 0.254$, $\eta^2 = 0.015$, partial $\eta^2 = 0.028$, and $\omega^2 = 0.004$. In other words, a small interaction effect that is not statistically significant at any meaningful significance level (i.e., 1%, 5%, or 10%). Table 12.5 presents R^2, AIC, and BIC of each of the usual Models 0–4.

Fig. 12.8 Quantile plot of the residuals of two-way ANOVA treating condition ($X = 0$: control, $X = 1$: treatment) and prior knowledge (P: 1, 2, 3, 4, 5) as factors that may have an interaction effect on post-test score (Y) in Experiment 6 (*Jamovi*)

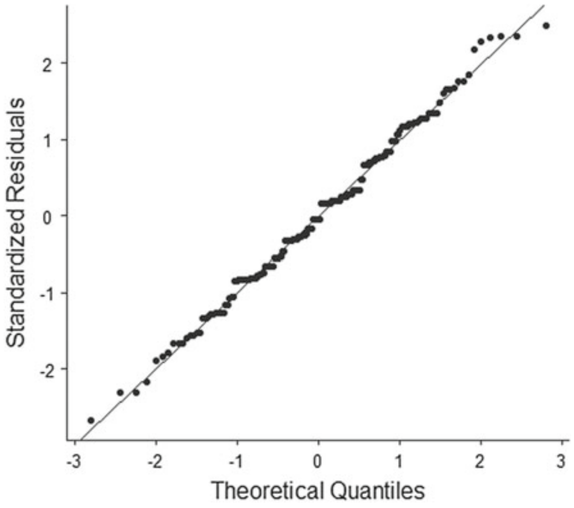


Table 12.5 R^2, AIC, and BIC of the five competing models in Experiment 6, treating treatment (T) and prior knowledge (P) as categorical variables (*Mplus*)

Model	R^2	AIC	BIC
0: null	0.000	978.265	984.862
1: T	0.013	977.567	987.462
2: P	0.454	865.136	884.926
3: T and P	0.479	857.811	880.899
4: T, P, and T-by-P	0.494	860.216	896.498

AIC and BIC both prefer Model 3. Now, if we prefer to view prior knowledge as a variable of interval level of measurement, we can also use Model 3 to test polynomial contrasts for prior knowledge, given five levels of prior knowledge (*Jamovi*): linear ($B = 4.749, p < 0.001$), quadratic ($B = -3.020, p < 0.001$), cubic ($B = 0.658$, $p = 0.057$), and quartic ($B = -0.213$, $p = 0.447$); and for the treatment effect, we find: $F_{1,\ 190} = 6.498$, $p = 0.003$, $\eta^2 = 0.025$, partial $\eta^2 = 0.046$, and $\omega^2 = 0.022$. Some statistical packages, such as *SPSS*, can help us obtain AIC and BIC for the models that treat the relation between prior knowledge and post-test score as linear, quadratic, cubic, and quartic. For the linear model, we find: AIC = 910.072 and BIC = 923.265. For the quadratic model, we find: AIC = 858.304 and BIC = 874.796. For the cubic model, we find AIC = 856.410 and BIC = 876.200. For the quartic model, we find AIC = 857.811 and BIC = 880.899. The latter are the same as the ones we see for Model 3 in Table 12.5, because the quartic model uses the same *df* = 4 for the effect of prior knowledge as Model 3 in Table 12.5. The other three models use fewer *df*: the cubic model 3 *df*, the quadratic model 2 *df*, and the linear model only 1 *df*.

AIC prefers a cubic model, whereas BIC prefers a quadratic model. Although different researchers may use different criteria (*p*-values, AIC, BIC) to decide whether a quadratic or a cubic model is to be preferred, all will agree that to treat the relation between prior knowledge and post-test score as linear (cf. the linear model) is *not* a good idea.

If there was substantial treatment-by-prior-knowledge interaction, the interaction effect could also take the form of a polynomial and same kind of comparison of linear, quadratic, cubic, and quartic could be applied to that interaction effect. Regardless of whether the relation between prior knowledge and post-test score was linear or non-linear, as in linear models (e.g., Experiments 2, 4, and 5) the interaction effect would be linear if there was a constant increase or a constant decrease in distance between the condition-specific regression lines with increasing prior knowledge.

The Different Roles of ‘Third’ Variables

Additional or so-called ‘third’ variables, even though we are often not really interested in them, may help us gain a better understanding of a treatment effect of interest. Mediators may help us to gain an understanding of one of the potential mechanisms of a treatment effect (i.e., how a treatment effect works, though see Chap. 1 for a note of caution, hence the addition ‘potential’ in one of the potential mechanisms), moderators may help us think in terms of different effects of a treatment under different conditions, and even if an additional variable is neither a mediator nor a moderator it may constitute a good covariate because it helps to increase the statistical power for a treatment effect of interest. Note that the examples in this chapter all use a single additional variable next to a single treatment factor. When sample sizes are large enough, it is possible to deal with combinations of more than one treatment factor and more than one mediator, moderator or covariate. The steps discussed in this chapter hold for these extensions as well, but extra steps will have to be taken. For instance, in an experiment with two treatment factors and one covariate or in an experiment with one treatment factor and two covariates, the non-interaction assumption needs to be checked for several combinations of variables (cf. the three-way example in Chap. 11).

Early on in Chap. 1, I state that we cannot just treat quasi-experimental data as if we were dealing with data from a randomised controlled experiment, and that doing so comes at the risk of inappropriate conclusions and recommendations for future research and practice in a field. If our groups to be compared are pre-existing instead of randomised groups, they may already differ in several key variables, observed and unobserved, to degrees that are very unlikely in true experiments. Adding covariates to ‘correct for confounding’ does not guarantee a solution to this problem (see also Chap. 15). It is quite common to add one or more covariates in a quasi-experiment in an attempt to control for differences in key variables prior to the start of treatment. This is problematic for a number of reasons.

Firstly, substantial differences between groups may lie in key variables that have not been measured and that would, due to appropriate randomisation, probably differ (much) less in a true experiment. In a true experiment, differences in key unobserved variables A, B, C, D, E, and F will, prior to the experiment, vary around zero from somewhat or slightly in favour of the control condition for some variables (e.g., A, D, and E) to somewhat or slightly in favour of the treatment condition for the other variables and can be expected to cancel each other out across the range of unobserved variables. In a quasi-experiment, it is possible that the difference on most or all of variables A–F is in favour of *the same condition* prior to the treatment. This is problematic partly because we cannot really correct for variables we have not observed.

Even in the highly unlikely case that substantial differences existed only in observed variables—say covariates G and H—we have problems. Predictor variables in the regression models discussed in this and previous chapters should preferably be uncorrelated or otherwise correlate as little as possible. In a factorial design where every cell has the same number of participants (e.g., $N = 160$ randomly allocated to the four combinations or cells in a 2-by-2 design yielding $n = 40$ for each cell), the correlation between factors is zero. In a randomised controlled experiment which includes a covariate measured prior to the start of treatment, randomisation will usually result in minor differences on the covariate between groups, resulting in a correlation fairly close to zero. In quasi-experiments, larger differences may be more likely to occur, and the larger the differences the more problematic. Adding the covariate in an attempt to 'correct for confounding' will not solve the problem; that step will not magically provide us with the 'pure' treatment effect. All that happens when we add a covariate, is that the M_d between groups is evaluated at the average of the covariate (and perhaps at some other values of the covariate if the covariate moderates that M_d). When pre-existing groups (subpopulations) by nature differ substantially on the covariate at hand, a comparison at the average of that covariate is a *counterfactual* that may have little meaning if any. Such a comparison is then based on an incorrect assumption of regression towards a *common* mean on the covariate while in fact there is regression towards *different* means on that covariate (see also Chap. 15). Besides, even if there is substantial difference on the covariate in the sample but not in the population—which may well happen for example when dealing with conveniently available pre-existing groups (i.e., neither random sampling nor random allocation)—more correlation between predictor variables in a sample comes with higher *SE*s and thus reduced statistical power and precision for group differences of interest. As a rule of thumb, the more we have to correct from a sample to a population, the more our *SE*s are inflated and more power and precision we lose.

Simultaneously, while variables measured after the start of treatment and before the outcome variable of interest may serve as mediators in randomised controlled experiments, we may not be able to tell to what extent differences in that potential mediators already existed prior to the start of the treatment in the case of a quasi-experiment. Baseline measurements, regardless of what statistical approach we use to test and estimate them, may in quite a few educational and psychological research settings not provide a reasonable solution to the problem. When the outcome variable of interest is about learning, for example, we will likely face an *assessment and learning paradox* (Leppink, 2018c): the very *measurement* of knowledge or skill at a given point in time can *itself influence learning*. In other words, by adding a pre-test on say probability calculus when the post-test will measure probability calculus as well, we are most probably adding an effect to our model. To investigate the effect of that addition, we would need to consider pre-test itself as an additional factor (i.e., pre-test: yes vs. no). Again, this is more feasible in true experiments than in quasi-experiments.

Interpreting effect size estimates and other outcomes of a meta-analysis over a series of non- or quasi-experimental studies as if the underlying studies were true experiments is problematic for the same reasons outlined in the previous. For instance, in a quite widely used and cited meta-analysis by Grynszpan, Weiss, Perez-Diaz, & Gal (2014), three types of pre-post designs were identified depending on the control condition used: (1) studies on an intervention's efficacy using control groups of participants with Autism Spectrum Disorder who did not receive the intervention; (2) studies on learning characteristics of people with Autism Spectrum Disorder that relied on control groups of participants not diagnosed with Autism Spectrum Disorder; and (3) studies on the feasibility of an intervention that were devoid of a control group. Several of the studies called 'experiment(s)' did not apply randomisation, yet that constitutes one of the key features to speak of an experiment in the first place. Several studies included applied some form of matching by age, gender, symptom severity, or cognitive abilities. Why and how matching for these covariates could create groups that on characteristics not matched for and not randomised either were comparable, remains unclear. Possibly in the majority of studies included the evaluator was not blind to group allocation and that may have had an influence on the outcomes in those studies. Correlations of variables like age with outcomes of interest are interpreted in terms of causal relations (e.g., age influencing or not influencing the outcome). Such conclusions cannot be drawn even in well-designed randomised controlled experiments, because age is not a factor that we can randomise like we can do with treatment factors. If people with Autism Spectrum Disorder in different age groups tend to receive somewhat different treatments—for instance: in technology-based interventions for Autism Spectrum Disorder, robots may be used more often with kids while applications (apps) may be used more frequently with adults—differential type of treatment may fully explain the apparent 'causal' relation between age and an outcome variable of interest (e.g., learning, behaviour change). Apart from the fact that most of the methodological and statistical choices made in the article reporting on the meta-analysis are not or poorly explained, and the main aim of the meta-analysis appeared to reach '$p < 0.05$' for the mean effect size estimate more than anything else (which would surely not be the case if some correction for publication bias was applied), with such a heterogenous collection of studies most of which do not meet some vital criteria to be called experiments the whole idea of a meta-analysis becomes questionable. Most of the conclusions drawn by the authors of this meta-analysis article simply cannot be drawn.

In line with the statement in Chap. 1, the reasons outlined here to refrain from treating data from a quasi-experiment as if that data came from a true experiment should *not* be interpreted as a recommendation against doing quasi-experimental research. Quasi-experiments can be very useful and may in not so few settings even make more sense or at least be more feasible than true experiments; we should just be aware of the dangers of using statistical methods that go well in experiments and interpreting the outcomes the same way as if we were doing an experiment. Even in the context of a meta-analysis, we cannot just include true experiments and other types of studies in the same meta-analysis and treat all studies as if they were experiments.

Part IV
Multilevel Designs

Interaction Between Participants

13

Abstract

For the sake of introduction of all the concepts thus far, the example experiments in previous chapters are about individual participation of participants and a few other conditions (e.g., no twins, family members or natural groups otherwise) that likely do not invalidate the assumption of individual participants counting as independent observations. However, as in natural settings, increasing numbers of experiments have to deal with some kind of interaction between participants prior or during the experiment (e.g., twins, family members, students from the same learning groups, employees or patients from the same centre). In such cases, independence of observations is an unrealistic assumption. Whether a dependence results from natural ties or from an interaction component explicitly incorporated in an experiment, we need to account for that dependence in our statistical models. Three different types of situations are discussed in this chapter: dyads (e.g., couples), small-size groups (e.g., project teams), and larger groups or social networks. In each of these types of situations, individuals are treated as actors nested within higher-level actors or units (e.g., pairs, teams, centres or cliques). Also discussed in this chapter is how failing to account for this kind of data structure (dependence) can result in substantial distortions of our perspective on a treatment effect of interest.

Introduction

As discussed in Chap. 1, SUTVA (*stable unit treatment value assumption*) is an important assumption in causal inference, and under this assumption, a treatment applied to one participant does not affect the outcome of another participant. In experiments where participants undergo treatment and are measured individually, SUTVA may well hold. However, when a component of interaction is introduced,

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_13

things get tricky. In this case, the treatment applied to one participant may influence the outcome of another participant through the interaction between participants. This is problematic especially if the interaction involves participants from *different* conditions. When the interaction occurs *within* but not *between* conditions, there may be no problem, especially if interaction is natural in the kind of behaviour that is studied. For example, for researchers who are interested in comparisons involving different types of group learning, it appears natural to apply condition-specific treatment to *groups* of participants. In some cases, these groups may be a random sample of existing groups from a population of a much larger number of existing groups. In other cases, individual participants may be sampled randomly, then randomly allocated to learning groups, and these randomly composed groups are then allocated randomly to treatment conditions. We can then account for the interaction and interpret a treatment effect of interest without fearing conditions influencing each other.

Experiment 1: Dyads

In the simplest case, the interaction between participants lies in working in dyads (i.e., pairs, teams of two participants) to perform a particular kind of task. In some experiments, the outcome variables of interest may be measured *while* the interaction takes place, whereas in other experiments the outcome variables may be measured *after* the interaction. In the context of learning dyads or learning groups, an example of measuring an outcome variable *while* the interaction takes place is found in a short questionnaire on effort or motivation during the group learning session, and an example of an outcome variable measured *after* the interaction is found in a post-test or examination after the group learning session. Even if the post-test or other kind of outcome variable of interest is measured *after* the interaction, learners from the *same* group will likely yield somewhat more common responses on an outcome variable of interest than randomly drawn learners from *different* groups. This tendency can be estimated through the *ICC*. If pairs or groups of participants did not interact, the *ICC* might well be 0, as in experiments where everyone received individual treatment. However, where pairs or groups of participants interact, likely $ICC > 0$, and that has to be accounted for in our statistical models and required sample size calculations.

Required Sample Size

Chapter 1 includes a formula to compute the design effect.

$$\text{design effect} = \sqrt{(1 + [(n - 1) * ICC])}.$$

In this formula, n is the sample size per cluster. In the case of dyads, $n = 2$, so the design effect formula comes down to:

$$\text{design effect} = \sqrt{(1 + ICC)}.$$

For *ICC* = 0, the design effect is 1; for *ICC* = 0.10, the design effect is about 1.049; for *ICC* = 0.20, the design effect is about 1.095. We also recall from Chap. 1 that the *square* of the design effect can be interpreted as the factor by which we would need to multiply our total sample size (M_N) to achieve the same precision as if *ICC* = 0. For *n* = 2 (i.e., dyads), this comes down to:

$$M_N = 1 + ICC.$$

In other words, the *ICC* can be interpreted directly as a proportion of *increase* in total sample size *N* given the departure from *ICC* = 0. Thus, *ICC* = 0.10 would correspond with a 10% increase, and *ICC* = 0.20 with a 20% increase. Now, at first it may seem that larger cluster sizes *n* come with a much higher M_N. This is because given total sample size *N* and cluster sizes *n*, fewer groups will be randomised when *n* is larger. However, larger *n* may also come with lower *ICC* values. Succinctly put, in learning groups of 10–20 students, *ICC* values in the 0.05–0.20 range may be common, whereas intensive 1-on-1 interaction in dyads may well create an *ICC* in the 0.30–0.60 range. Martin, Bobis, Anderson, Way and Vellar, (2011) provide a very good example of this kind of variation in the context of psycho-educational phenomena.

If we consider that *ICC* = 0.50 is a plausible estimate in a given context with dyads, the above formula indicates that we would have to increase our sample size by 50%. Note that in the very extreme case of *ICC* = 1, the maximum possible ICC —which we would obtain if there were differences *between* pairs but no differences *within* pairs—we would need to increase our sample size by 100%: that would come down to doubling the sample size. How does this make sense? Well, if each pair yields two identical scores, the number of independent observations is reduced to the number of pairs, that is: we have *one* independent observation for each pair while *ICC* = 0 would generate *two* independent observations for each pair. Hence, when there are differences between pairs (which is usually the case) but within pairs outcomes are always the same (i.e., *ICC* = 1), we have to double the number of pairs to have the same statistical precision and power as if we were dealing with individuals. In a two-group experiment with $n = 64$ participants per condition, we would then need 64 pairs per condition, hence a total sample size of 256 instead of 128.

Consider the following example of an experiment involving interacting dyads. Suppose, researchers are interested in a comparison between two types of learning a foreign language in pairs. After all, language learning involves among others practice with peers. Initially, this practice takes the form of exchanging small sentences in turns. Next, bit by bit, sentences are supposed to become more elaborate, and eventually we get to the level of having short and ultimately longer dialogues. Let us call these two types of learning the *control* and the *treatment* condition. The researchers have no idea which of the conditions will yield better outcomes on a post-test (score: an integer ranging from a minimum of 0 to a maximum of 25), so they decide to opt for two-sided testing. They expect that having participants learn in pairs will likely create an *ICC* in of about 0.40. Knowing that in the case of *ICC* = 0, a two-group experiment with *N* = 128 or *n* = 64 per condition would yield a statistical power of 0.80 for a difference of *d* = 0.50 testing two-sided with a good-old-fashioned Student's *t*-test at $\alpha = 0.05$, the authors reason that to achieve the same statistical power with *ICC* = 0.40, they will need *N* = 128 * 1.40 = 179.2. They have logistic means for 180 participants so they decide to randomly recruit *N* = 180 participants, randomly assign them to *K* = 90 pairs, and then randomly allocate the *K* = 90 pairs to the two treatment conditions such that each condition has *k* = 45 pairs.

Mixed-Effects Modelling (1): Estimating Random Effects

Figures 13.1 and 13.2 present the histogram and boxplot of the distribution of post-test score (*S*) for the control (*X* = 0) and treatment (*X* = 1) condition, respectively.

In the control condition, post-test score ranges from 10 to 21, *M* = 15.700, and *SD* = 2.654. In the treatment condition, post-test score ranges from 10 to 22, *M* = 14.789, and *SD* = 2.758. The *ICC* can now be estimated through a *mixed-effects* model. The models used in earlier chapters in this book are so-called *fixed-effects* models, because the treatment effects of interest are fixed effects and we have not estimated any random effects other than the residual term. In the experiment at hand, *pair* and *individual* constitute two *hierarchical levels*, with pair constituting the upper level (level 2) and individual the lower level (level 1). At the level of individual, we cannot estimate random effects. However, at the level of pair, we *can* estimate a random effect. In a fixed-effects regression model for data from a two-group experiment, the intercept and slope relate the *M*s of the control and treatment condition. These fixed-effects terms can still be estimated in mixed-effects models. However, now each pair has its own intercept around which the individual

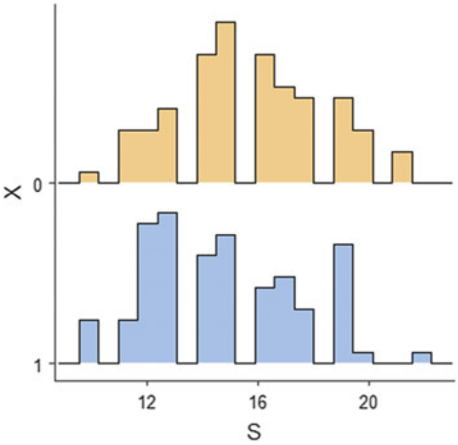
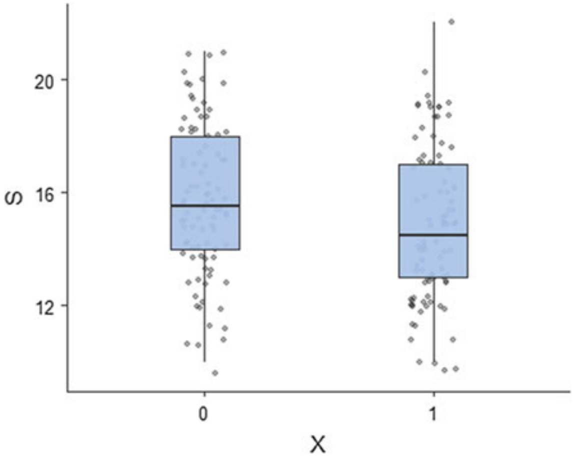


Fig. 13.1 Histogram of the distribution of post-test score (*S*) for the control (*X* = 0) and treatment (*X* = 1) condition (*Jamovi*)

Fig. 13.2 Boxplot of the distribution of post-test score (*S*) for the control (*X* = 0) and treatment (*X* = 1) condition (Jamovi)



scores vary. In the extreme case that there is no variation within pairs, all variance is variance *between* pairs and hence *ICC* = 1. Given a particular variance between pairs, the more variation *within* pairs, the more *ICC* goes down. In the extreme case that the variance between pairs is zero but that within pairs is not, *ICC* = 0.

In experiments like the one at hand, pair-level intercepts are called *random effects* because we assume the pairs to result from random sampling and want to generalise to a population of possible pairs. This is an important distinction with *fixed effects* like a treatment effect. A treatment factor comes with predefined conditions; generalisation takes place from the sample of these predefined conditions to a population of these predefined conditions (e.g., Leppink, 2015a). The set of prespecified conditions is *not* some random sample of possible conditions to which we generalise; we are exclusively interested in *this set* of prespecified conditions. In line with earlier chapters in this book, fixed effects such as treatment effects can be estimated using FIML. In some statistical packages, this is simply called 'maximum likelihood' (ML). However, for random effects such as the pair-level intercepts variance, FIML results in somewhat *underestimated* effects, and REML is therefore recommended as an alternative to FIML for random effects (e.g., Tan, 2010; Verbeke & Molenberghs, 2000). For the kind of models suitable for the experiment at hand, many statistical packages can do the job, including Stata, SPSS, Mplus, and Jamovi.

In sum, we are dealing with a *mixed-effects* model, because the treatment effect of interest is a *fixed effect*, and the pair-level RI variance is a *random effect*. We are not adding the latter because of a genuine interest—our interest lies in the treatment effect—but we need to include that random effect to account for the correlational (i.e., dependence of observations) structure, unless it is so small that *ICC* is very close to zero (i.e., *ICC* < 0.01). Using REML in Jamovi, we find *ICC* = 0.570 when we assume the means of the conditions to be different and *ICC* = 0.579 when we assume the means of the conditions to be the same. Unless the difference in means is really zero or so close to zero that we may as well call it zero, treating the means as different in the estimation of *ICC* is more appropriate. The reason for that is the following. In a model that includes fixed effects and a RI term, the resulting *ICC* is:

$$ICC = V_{\mathrm{RI}}/(V_{\mathrm{FIXED}} + V_{\mathrm{RI}} + V_{\mathrm{RES}}).$$

If no fixed effects were estimated or V_{FIXED} was 0, V_{FIXED} would be dropped from the last equation. Not including fixed effects while they should be included is likely to result in an overestimation of *ICC*, because what should be seen as V_{FIXED} is then attributed partly to V_{RI} and partly to V_{RES} and likely such that the resulting *ICC* not accounting for V_{FIXED} is slightly higher than the *ICC* in which V_{FIXED} is appropriately accounted for. This does not hold only for treatment conditions in a randomised controlled experiment; there are good examples of how computing a Cronbach's alpha over a group of experts and a group of novice participants together provides a heavily inflated outcome compared to when that Cronbach's alpha is estimated for the two groups separately (e.g., a rather extreme example from Cook, 2015: well over 0.9 when treating experts and novices as *one* group, equivalent to not accounting for V_{FIXED}, instead of hardly over 0.6 when including V_{FIXED} hence correctly treating experts and novices as distinguishable fixed-effect groups).

Analogous to the LR test for fixed effects, which uses FIML, we can perform a LR test for random effects using REML. This way, we can test whether a random effect of interest differs significantly from 0 (i.e., if a random effect was 0, we would not need it). For $ICC = 0.570$, the LR test using REML yields: $\chi^2{}_1 = 35.191$, $p < 0.001$. The χ^2-test is one of $df = 1$, because the RI variance is a *single* estimate, hence we do not need more *df*. With regard to the statistical significance level, some may suggest to compare the *p*-value with significance level α as in a two-sided test, whereas others may argue to divide the *p*-value by two as in a one-sided test, because given one more complex and one simpler model the direction of the difference between the two models in what they can explain is always known. Either way may be defended, as long as we justify our choice.

Most statistical packages also allow researchers to compare a model with and without a random effect in terms of AIC, BIC or information criteria alike. With fixed effects, competing models can be compared in terms of AIC and BIC using FIML (as in the earlier chapters in this book). With random effects, REML is more appropriate than FIML and this comes with a problem at least for BIC. The penalty in BIC to protect against overfitting combines the number of parameters in a model and the sample size. In fixed-effects models, where a sample size of *N* can be interpreted as *N* number of independent observations, this is fine. However, in mixed-effects models, what *N* represents is open to discussion and the correction applied by BIC may then be too severe (e.g., Delattre, Lavielle, & Poursat, 2014; Fitzmaurice, Laird, & Ware, 2004; Hedeker & Gibbons, 2006). When random effects are large, like the *ICC* estimating the RI variance in the experiment at hand, BIC will favour a model with that random effect over a model without that random effect. However, when effects are more modest yet still possibly substantial, BIC may prefer a model without random effect (i.e., assuming $ICC = 0$) while actually a model *with* random effect makes more sense even though the *ICC* seems fairly 'small' (e.g., 0.05). Furthermore, while given a particular random effects structure, different software packages provide the same AIC and BIC values for the fixed effects (i.e., using FIML), under REML the outcomes of AIC and BIC may differ somewhat across packages (e.g., Delattre et al., 2014) and some packages do not even provide BIC under REML (e.g., *Jamovi*).

Decisions on the random effects structure in an experiment should best be based on a combination of common sense and LR testing under REML, with *p*-values from the LR test to be divided by two or to be tested at 2α (cf. one-sided testing explained in previous chapters). Common sense means staying as closely as possible to the features of the experimental design. In the experiment at hand, this comes down to a simple decision of including or excluding a RI variance. The *ICC* is large and the LR test under REML is statistically significant; in this case, including the RI variance makes sense.

Mixed-Effects Modelling (2): Estimating Fixed Effects

The decision on the inclusion of the RI variance has importance for testing and estimating the treatment effect of interest. If we were to estimate that treatment effect without including the RI term (i.e., a fixed-effects model), we would find for B (i.e., M_d): $B = -0.911$, $t_{178} = -2.258$, $p = 0.025$, and a 95% CI of the B of [−1.707; −0.115]. In a mixed-effects model which includes the RI term and uses FIML for the estimation of fixed effects, we find (*Jamovi*, *SPSS*): $t_{90} = -0.911$, $p = 0.073$, and a 95% CI of M_d of [−1.894; 0.072]. The point estimate is the same, because $n = 2$ across clusters (i.e., pairs) and no data is missing. However, the CI is wider and consequently the *p*-value is higher as well. In the fixed-effects model, we overestimate the effective sample size by incorrectly assuming $ICC = 0$; the higher the *ICC*, the more the reduction in effective sample size relative to the total sample size (*N*).

In our mixed-effects model, under FIML, we find AIC = 842.199 and BIC = 851.778 for Model 0 (i.e., $\boldsymbol{H_0}$: no treatment effect) and AIC = 840.958 and BIC = 853.730 for Model 1 (i.e., $\boldsymbol{H_1}$: treatment effect). Performing a LR test under FIML for the treatment effect, we find: $\chi^2{}_1 = 836.199$ (i.e., -2LL of null model)—832.958 (i.e., -2LL of alternative model) = 3.241, $p = 0.072$.

Some statistical packages, such as *Jamovi*, also provide a marginal and conditional R^2 (henceforth: $R^2{}_{\mathrm{M}}$ and $R^2{}_{\mathrm{C}}$) for each model. While $R^2{}_{\mathrm{C}}$ combines the fixed and random effects, $R^2{}_{\mathrm{M}}$ is about the fixed effects and can therefore be interpreted in a similar fashion as R^2 in fixed-effects models (Nakagawa, Johnson, & Schielzeth, 2017; Nakagawa & Schielzeth, 2013). In the computation of both $R^2{}_{\mathrm{M}}$ and $R^2{}_{\mathrm{C}}$, the denominator is the sum of fixed-effects variance (V_{FIXED}), random-effects variance (V_{RANDOM}), and remaining (i.e., lowest-level) residual variance (V_{RES}). However, the numerator is the sum of V_{FIXED} and V_{RANDOM} in the case of $R^2{}_{\mathrm{C}}$ but only

V_{FIXED} in the case of $R^2{}_{\text{M}}$ (e.g., Nakagawa et al., 2017). In other words, the formulae of $R^2{}_{\text{M}}$ and $R^2{}_{\text{C}}$ are:

$$R^2{}_{\text{M}} = V_{\text{FIXED}}/(V_{\text{FIXED}} + V_{\text{RANDOM}} + V_{\text{RES}}), \text{and}$$
$$R^2{}_{\text{C}} = (V_{\text{FIXED}} + V_{\text{RANDOM}})/(V_{\text{FIXED}} + V_{\text{RANDOM}} + V_{\text{RES}}).$$

Note that if there are no random effects to be included, V_{RANDOM} is 0, and hence $R^2{}_{\text{C}} = R^2{}_{\text{M}} = R^2$ of a fixed-effects model. Note also that the difference between $R^2{}_{\text{C}}$ and $R^2{}_{\text{M}}$ can be interpreted as $R^2{}_{\text{R}}$:

$$R^2{}_{\text{R}} = R^2{}_{\text{C}} - R^2{}_{\text{M}} = V_{\text{RANDOM}}/(V_{\text{FIXED}} + V_{\text{RANDOM}} + V_{\text{RES}}).$$

With this, we are effectively back to the formula of *ICC*. In an experiment, where the interest lies in fixed effects and where random effects are only included to account for the data structure so that appropriate testing and estimation outcomes can be acquired for the fixed effects of interest, $R^2{}_{\text{R}}$ and $R^2{}_{\text{C}}$ are usually not really of interest, but $R^2{}_{\text{M}}$ is. In the experiment at hand, we find $R^2{}_{\text{M}} = 0.028$. Moreover, some may wonder if we can still use effect size estimates such as Cohen's *d* or standardised *β*, like in fixed-effects models. This is where opinions among scholars diverge. Some argue that in mixed-effects models, the lowest level variance is "*the amount of variation in the outcome measure attributable to the individual observation after appropriate controls have been made*" (Schagen & Elliot, 2004, p. 13) and therefore state that "*such calculations are considered appropriate because they explicitly model the extent and impact of clustering in the data*" (Schagen & Elliot, 2004, p. 13). However, doing so will result in inflated effect size estimates, and very much so when *ICC* values are substantial. In the extreme case that V_{RES} goes to 0 because the random effect(s) included in a model can fully explain the residuals, effect size estimates go to infinity.

In the experiment at hand, V_{RES} is 3.167. If we use this to estimate Cohen's *d*, we obtain $d = -0.512$. However, when we use V_{RES} from a fixed-effects model, we find $d = -0.337$. This is quite a difference, and $d = -0.337$ is more in line with Fig. 13.1 (i.e., with a difference of 0.5 *SD* or more, we would see a somewhat clearer difference in location of the histograms). Besides, the approach proposed by Schagen and Elliot (2004) is not consistent with the formulae of $R^2{}_{\text{R}}$, $R^2{}_{\text{C}}$, and $R^2{}_{\text{M}}$; we do not leave out V_{RANDOM} of the nominators of these three R^2-statistics, so why would we do that when estimating effect sizes?

Finally, we have already seen that the fixed-effects and mixed-effects model, given constant *n* per cluster (i.e., 2) and no missing data, yield the same point estimate of M_d. In that light, it makes much more sense to adjust the 95% CI (or the 90% one when we are using a 90% CI) for clustering by computing the bounds of *d* from the *ICC*-adjusted bounds of M_d (i.e., −1.894 and 0.072). The fixed-effects (i.e., *ICC*-uncorrected) model yields a 95% CI of *d* of [−0.630; −0.042]. In the mixed-effects (i.e., *ICC*-corrected) model, we then find a 95% CI of [−0.700;

0.027]. This can also be done when n is not constant across clusters (i.e., due to starting with unequal n and/or due to missing data), just that in that case not just the interval but also the point estimate will differ from what we would obtain with a fixed-effects model.

Experiment 2: Groups

If instead of pairs, we have larger groups of learners, the factor by which we would need to multiply our total sample size (M_N) comes down to:

$$M_N = 1 + [(n - 1) * ICC].$$

If we deal with groups of n = 10 each and it is reasonable to expect ICC = 0.20:

$$M_N = 1 + [9 * 0.20] = 1 + 1.8 = 2.8.$$

For ICC = 0.10, we would find:

$$M_N = 1 + [9 * 0.10] = 1 + 0.9 = 1.9.$$

In other words, for ICCs in the range of 0.10–0.20 we would need to multiply N in the case of ICC = 0 with a factor 1.9–2.8 depending on the exact ICC value we consider reasonable. Suppose, we are dealing with a three-group experiment and strive for a statistical power of 0.80 for a medium size difference (η^2 = 0.06) using a default F-test at α = 0.05. Under ICC = 0 (i.e., one-way ANOVA), this comes down to N = 159 or n = 53 per condition. Using a multiplication factor of 1.9–2.8, we would need between n = 101 per condition (using 1.9) and n = 149 per condition (using 2.8). Suppose that we are dealing with a situation like in Experiment 2 in Chap. 10: we expect a positive difference of *both* treatment conditions from the control condition (i.e., a one-sided test) but have no clear expectations with regard to how the two treatment conditions differ from one another. Each of three conditions comprises k = 14 learning groups of n = 10 individuals each.

Figures 13.3 and 13.4 present the histogram of the distribution of post-test score (0–100, higher is better) for the three conditions (X = 0: control; X = 1: treatment A; X = 2: treatment B) at the level of the individual (Fig. 13.3) and at level of the group (average per group: Fig. 13.4).

Figure 13.3 provides a preliminary screening of possible extreme cases in post-test score, and Fig. 13.4 can help us to check for eventual extreme groups. Additionally, we could inspect histograms of the distribution of scores per group, to locate eventual extreme cases in terms of in which group(s) they are (it is normally helpful to do that, but adding another 42 histograms to this book is not needed). An equivalent of the latter does not make sense when dealing with pairs, but an equivalent of Fig. 13.4 could help to check for eventual extreme pairs (there are

Fig. 13.3 Histogram of the distribution of post-test score (*S*: 0–100; higher is better) for the three conditions (*X* = 0: control; *X* = 1: treatment A; *X* = 2: treatment B) at the level of the individual (*Jamovi*)

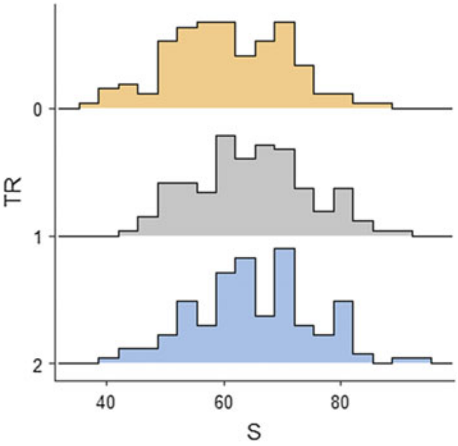
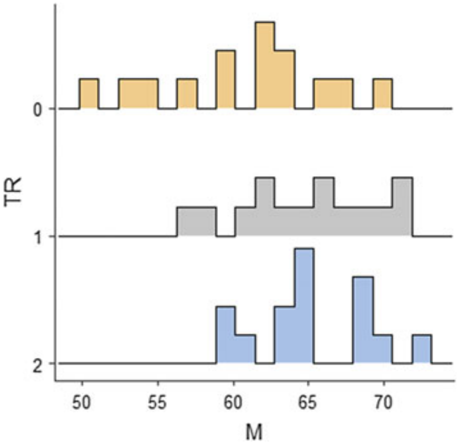


Fig. 13.4 Histogram of the distribution of post-test score (*S*: 0–100; higher is better) for the three conditions (*X* = 0: control; *X* = 1: treatment A; *X* = 2: treatment B) at the level of the group (average per group) (*Jamovi*)



none in Experiment 1). These figures are generally useful, because assumptions of normally distributed residuals are made in mixed-effects models as well: in the population sampled from, we generally assume RIs to be normally distributed around the fixed intercept (i.e., the RIs observed form a random sample from a normally distributed population of RIs), and we assume the individual residuals to be normally distributed around their group-specific RIs. Some departures from these assumptions are not problematic, and even less so when dealing with large samples, but the problem with more extreme departures is that they may substantially inflate at least one of the variance estimates. The quantile plots in Figs. 13.5 (for the distribution of intercepts) and 13.6 (for the distribution of individual residuals around their intercepts) can help us examine the aforementioned assumptions.

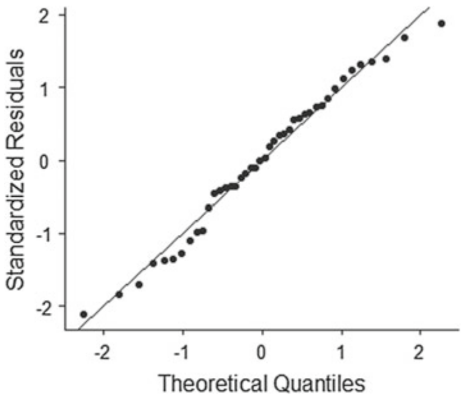


Fig. 13.5 Quantile plot of the residuals of RIs around the fixed intercept (Jamovi)

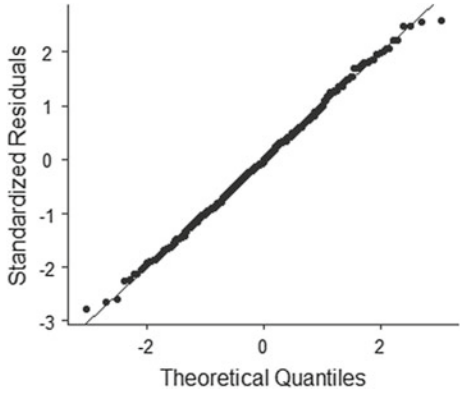


Fig. 13.6 Quantile plot of the individual residuals around the RIs (Jamovi)

These plots do not indicate severe departures from normality. For the three conditions, we find the following *M*s and *SD*s. In the control condition, $M = 60.786$ and $SD = 10.017$; in treatment A, $M = 64.821$ and $SD = 9.822$; in treatment B, $M = 65.243$ and $SD = 10.346$. In other words, the two treatment conditions differ a bit over 0.4 *SD* from the control condition but less than 0.1 *SD* from one another.

We can now proceed with testing and estimating the RI variance. In this experiment, the two hierarchical levels are found in the learning groups (level 2) and the individuals within learning groups (level 1). Hence, the RI variance is about the extent to which the group-specific intercepts vary around the fixed intercept. Using REML in Jamovi, we find $ICC = 0.139$, and the LR test of the *ICC* yields $\chi^2{}_1 = 21.155$, $p < 0.001$. Although the *ICC* in this experiment is quite a bit lower than that in Experiment 1, it is still substantial so we should prefer a mixed-effects model (i.e., accounting for $ICC = 0.139$) over a fixed-effects model (i.e., erroneously assuming $ICC = 0$).

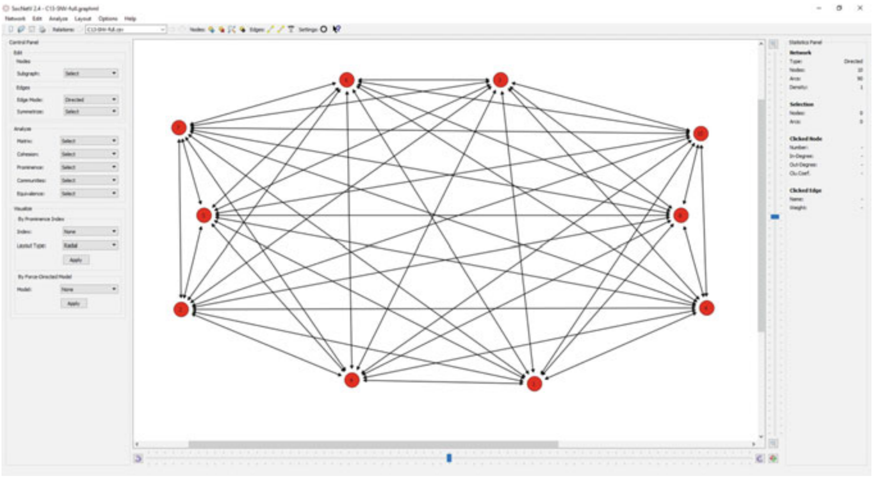
We can now proceed with the fixed-effects part, using FIML in Jamovi. For the treatment factor, we find $R^2{}_M = 0.039$. Using Helmert coding, we find for the first contrast, *both* treatments versus the control: $B = 4.246$, $p = 0.004$ (one-tailed), and a 90% CI of [1.281; 7.212]. For the second contrast, the two treatments compared with one another, we find: $B = 0.421$ (in favour of treatment B), $p = 0.811$ (two-tailed), and a 90% CI of [−3.002; 3.845]. For a model in which all means differ, we find AIC = 3119.636 and BIC = 3139.838. For a model in the control mean differs from the treatment means but the latter do not differ from one another, we find AIC = 3117.695 and BIC = 3133.856. The *ICC*-adjusted 90% CI of d is [−0.298; 0.381]. This is somewhat wider than the 90% CI of d we would obtain in a fixed-effects model: [−0.193; 0.276]. The explanation of the difference is simple: $ICC > 0$ means a reduction of the effective sample size from N to N divided by M_N and a loss of information and precision proportional to it.

Advanced Questions

Although the experiments in this chapter have quantitative outcome variables, the concepts discussed in this chapter also apply to mixed-effects models for categorical outcomes variables (e.g., the extensions of the methods discussed in Chaps. 5–7). Besides, when additional variables are measured and we have sufficiently large samples, we may also estimate *random slopes* (RSs) along with or instead of RIs. For instance, if in Experiment 2 we had a quantitative covariate measured prior to the treatment, we could estimate the fixed-effects slope of the relation between covariate and outcome variable for each condition (i.e., model with interaction) or the fixed-effect slope of the relation between covariate and outcome variable across conditions (i.e., ANCOVA) using FIML but we could additionally estimate learning-group-specific RSs of that relation between covariate and outcome variable. Each learning group may have its own RS and, analogous to the RIs, we assume the RSs observed in our experiment to be a random sample of a population of possible RSs. That said, RS variances are easier to estimate in larger than in smaller clusters; when we deal with say $N = 900$ divided over $K = 30$ clusters of $n = 30$, we may have sufficient information in the data to obtain good estimates for both RIs and RSs; when either K or n is much smaller, estimating random effects becomes more difficult. With small numbers of K (e.g., $K = 10$), there are only a few RIs and RSs to estimate and a substantial deviation from normally distributed residuals may influence at least one of the random effect variances considerably. When K is sufficiently large but cluster size n is small (e.g., the experiments in this chapter), estimating the RIs variance may not be much of a problem but RSs become difficult to estimate and may not even make much sense even if they can be estimated. In experiments that involve repeated measurements, RSs are generally easier to estimate (e.g., Tan, 2010) and may in some cases make sense even if we have only two measurements per participants (e.g., Leppink, 2015a).

Finally, the mixed-effects analysis approach taken in this chapter, and more commonly in experimental educational and psychological research, is based on prespecified clusters. When an experiment explicitly includes a component of interaction and cluster sizes n are small, like in the experiments in this chapter, it appears reasonable to assume that all members of a cluster interact with each other. However, with increasing n, the likelihood of not all members interacting increases as well and it is possible that *cliques* or sub-groups of interacting individuals are formed (i.e., the term ‘cliques’ is also used in Chap. 3 in the context of network analysis as an approach to studying how different items or other variables of interest may be interrelated) and/or that some combinations of individuals yield a higher-frequency and/or higher-intensity interaction than other combinations of individuals. In such cases, social network analysis (Leppink & Pérez-Fuster, 2018; Scott, 1988, 2017; Wasserman & Faust, 1994) may be used to examine which individuals interact with each other and eventually to what extent. Figures 13.7 and 13.8 present two examples of a social network of a group (cluster) of n = 10 individuals who all interact with each other to an equal extent (Fig. 13.7) and who do not all interact with each other (Fig. 13.8).

Especially in times of online interaction, experimenters may collect data with regard to who interacts with whom and how frequently or intensely quite easily and social network analysis can then help to compare prespecified groupings with actual groupings and to make decisions with regard to what (different types of) clustering to use for mixed-effects modelling. In large-sample studies that monitor not only *if* interaction occurs but also the frequency or intensity of interactions in a network, social network analysis may help to compute distance measures that may be used in so-called *spatial multilevel models* (Arcaya, Brewster, Zigler, & Subramanian, 2012; Bingenheimer & Raudenbush, 2004; Dong, Harris, Jones, & Yu 2015; Dong, Ma, Harris, & Pryce, 2016; Ma, Chen, & Dong, 2017) to account for different degrees of interaction in a network. Although these models are commonly associated with geographical distances, different degrees of interaction can also be



Relations: C13-SNV-full.csv

Subgraph: Select

Edge Mode: Directed

Symmetrize: Select

Matrix: Select

Cohesion: Select

Prominence: Select

Communities: Select

Equivalence: Select

Index: None

Layout Type: Radial

Model: None

Type: Directed

Nodes: 10

Arcs: 90

Density: 1

Nodes: 0

Arcs: 0

Number: -

In-Degree: -

Out-Degree: -

Clu.Coef: -

Name: -

Weight: -

Fig. 13.7 Social network of a group of n = 10 individuals in which everyone interacts equally with everyone (SocNetV)

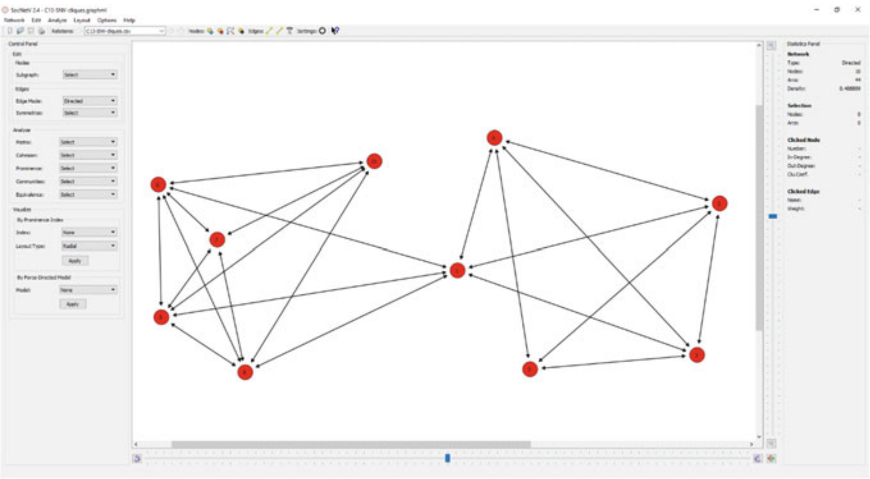


Fig. 13.8 Social network of a group of $n = 10$ individuals in which any individual interacts with only a limited number of other individuals in the group (*SocNetV*)

expressed in terms of distance measures. That said, the more complex the intended models, the more demands on sample size, and in most experimental settings the approach presented for the experiments in this chapter may well be sufficient.

This chapter is the first of four chapters that discusses situations where the assumption of independent observations that underlies the methods discussed in earlier chapters in this book is violated. This chapter presents examples of experiments in which participants are allowed to or even instructed to interact. In Chap. 14, we focus on a second common type of dependence in observations: ratings of the same individuals on an outcome variable of interest by two or more assessors. Even though the assessors are supposed to provide their ratings independently, differences between individuals rated creates a within-individual between-assessors *ICC* proportional to these differences. In Chap. 15, we see a few examples of experiments in which conditions of a treatment factor are measured twice or more times on the same outcome variable of interest. Finally, in Chap. 16, we deal with experiments in which different participants undergo different conditions in different orders. What all chapters in this fourth and final part of the book have in common is that a state of $ICC > 0$ (instead of $ICC = 0$) can be and should be accounted for.

Two or More Raters

14

Abstract

In not so few experiments where performance or another behavioural outcome variable is measured, scores of learners or other individuals result from ratings by two or more independent assessors. These ratings are often averaged into a single score per participant. While this does not always result in incorrect conclusions with regard to treatment effects of interest, a more accurate analytic approach is found in treating the raters or assessors as stations that have to be passed by participants. This chapter discusses some examples of how to run that type of analysis and acquire estimates of treatment effects and inter-rater reliability simultaneously.

Introduction

When an outcome variable of interest results from ratings from two or more independent assessors, different designs with regard to which assessors rate which participants are possible. In one extreme case, each participant is rated by a *different* cluster of independent assessors. In this case, raters can be considered as *nested within* participant, just like in Chap. 13 participants can be considered as nested within pairs or learning groups. Therefore, for this type of designs, the models discussed in Chap. 13 may be used to estimate treatment effects and the resulting *ICC* can provide an indication of inter-rater or inter-assessor reliability. However, this kind of design requires fairly large numbers of assessors and is for that and other reasons quite uncommon in experimental research. At the other end, we find a type of design where there is *one* set of *k* independent assessors that constitute the *k* number of stations to be passed by *all* participants. This type of case, where participants and raters form a fully *crossed* design, is quite common in experimental research and comes with the advantage that different raters can be treated as *repeated*

© Springer Nature Switzerland AG 2019
J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_14

measurements in the mixed-effects models and that allows us to estimate eventual systematic differences in rating tendency between raters. For instance, a stricter rater will likely provide lower scores on a test or fewer ‘passes’ on an exam than a more lenient rater. This kind of design is discussed in this chapter. Finally, there are designs somewhere in between nested and fully crossed. For instance, there are ten raters which form five pairs, and each of these five pairs provides ratings of a different subset of participants in an experiment. Although this is sometimes considered most feasible given the logistics in a particular context and we may in some cases still be able to use the methods discussed in this chapter, the more the balance shifts from crossed to nested, the less information we have to estimate systematic differences between raters and the more we find ourselves back with the methods discussed in the previous chapter. Moreover, in the context of the latter kind of design, if the different subsets of raters are active in only one or part of all conditions and there are considerable differences between raters on the strictness-leniency dimension, even with randomisation there may be considerable differences between conditions on the strictness-leniency dimension in a given experiment for randomisation is done only with a small number of raters. Therefore, this chapter focusses on fully crossed designs, one with two raters and one with four raters.

Experiment 1: Two Raters or Assessors

Suppose, we are dealing with a two-group experiment with $n = 64$ participants per condition and the outcome variable is a type of performance rated on an integer scale from 0 (*min*) to 10 (*max*) by two independent assessors. The assessors are always the same, meaning that all $N = 128$ participants have to pass these two assessors (stations). Figures 14.1 and 14.2 present the histograms of the distributions of the two assessors for each of the two conditions ($X = 0$: control, $X = 1$: treatment).

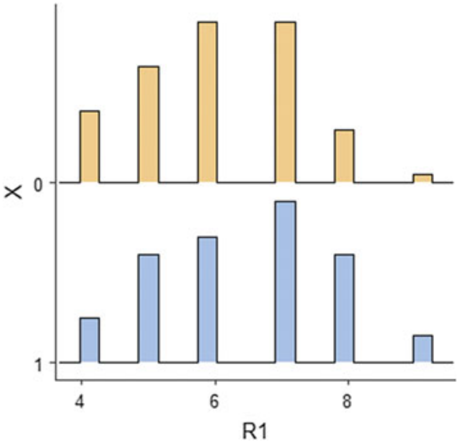
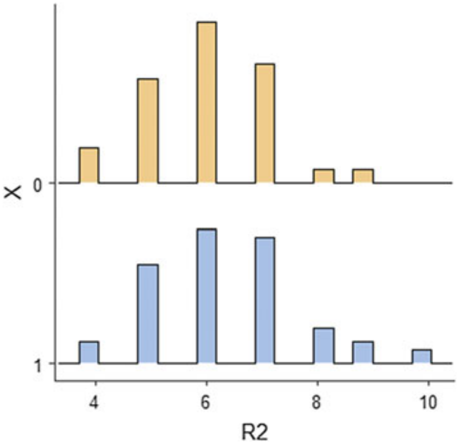


Fig. 14.1 Histogram of the distribution of ratings in the control ($X = 0$) and treatment ($X = 1$) condition for Assessor 1 (R1) (*Jamovi*)

Fig. 14.2 Histogram of the distribution of ratings in the control (*X* = 0) and treatment (*X* = 1) condition for Assessor 2 (R2) (Jamovi)



For Assessor 1, we find *M* = 6.063 (*SD* = 1.233) in the control condition and *M* = 6.453 (*SD* = 1.332) in the treatment condition. For Assessor 2, we find *M* = 6.031 (*SD* = 1.112) in the control condition and *M* = 6.391 (*SD* = 1.341) in the treatment condition. Figure 14.3 presents the scatter plot of the relation between the ratings of the two assessors.

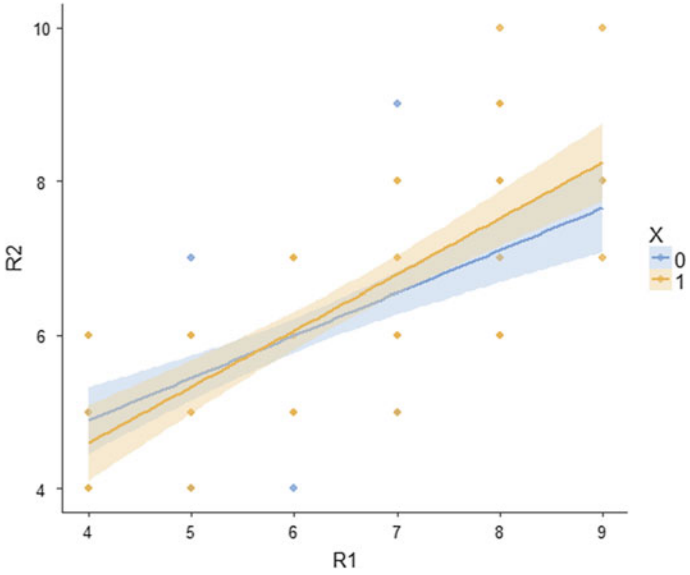


Fig. 14.3 Scatterplot of the relation between the ratings of the two assessors in Experiment 1 (Jamovi)

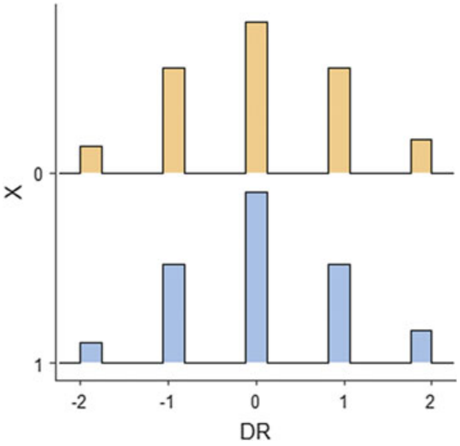


Fig. 14.4 Histogram of the distribution of differences in ratings between assessors (DR = R2 − R1) for the control (X = 0) and treatment (X = 1) condition in Experiment 1 (*Jamovi*)

Pearson's r is 0.612 in the control condition and 0.726 in the treatment condition. Figure 14.4 presents the histogram of the *difference* in ratings (*DR*: Assessor 2 minus Assessor 1) per condition.

In the control condition, M_{DR} = 0.031 and SD_{DR} = 1.038; in the treatment condition, M_{DR} = 0.063 and SD_{DR} = 0.990. Inspecting Figs. 14.1, 14.2, 14.3, 14.4 is useful, because they can inform our statistical modelling choices. The histograms are helpful because they indicate severe departures from normally distributed residuals neither in the univariate distributions of assessors (Figs. 14.1, 14.2) nor in the distribution of differences between assessors (Fig. 14.4), and the relation between ratings can reasonably be described in linear terms (Fig. 14.3). Inspecting the distribution of differences between assessors is important, because even in the case of normally distributed univariate distributions extreme differences between assessors in individual cases may occur and these can substantially influence M_{DR}, SD_{DR}, and r between assessors.

Mixed-Effects Analysis (1): Residual Covariance Structure

Not so few researchers may be inclined to calculate Cronbach's alpha as a measure of inter-rater reliability or the *ICC* from which Cronbach's alpha can be computed using the Spearman-Brown formula (see Chap. 3). If we make no distinction between treatment conditions, the *ICC* used for the computation of Cronbach's alpha assumes CS (cf. RI model): the same *SD* for all items (ratings) and the same correlation for all pairs of items (ratings). The means of items or ratings are allowed to differ as long as the *SD*s are the same. In the experiment at hand, where we have two assessors, there is only one correlation so that part is easy, and the two *SD*s are assumed to be equal. In the resulting mixed-effects model (*SPSS*), where the means of assessors are allowed to vary but the *SD*s are the same, under REML we find *ICC* = 0.682. From the Spearman-Brown formula, we can then compute

Cronbach's alpha and find 0.811. There are two potential problems with this model: it does not allow the means of the treatment and control condition to vary, and the *SD*s of assessors are not much different but not equal either.

There are two possible solutions to the problem of possibly different means for treatment and control condition: to run the aforementioned mixed-effects model for each condition separately or to use the same model as before but to allow the treatment conditions to vary (cf. the *ICC* estimations in Chap. 13). Using the first approach, we find $ICC = 0.609$ ($\alpha = 0.757$) in the control condition and $ICC = 0.726$ ($\alpha = 0.841$) in the treatment condition. In the second approach, we can also allow for condition-by-assessor interaction (i.e., the extent to which assessor differences differ across conditions *casu quo* M_d of the conditions differs across assessors). Doing so, we find $ICC = 0.675$; not much lower than 0.682 but a bit lower. The larger the difference in means between treatment and control condition, the larger the difference in *ICC* in a model that accounts for that difference and a model that does not account for it. As recommended in Chap. 13, it is generally better to account for mean differences when estimating an *ICC*, even if they are small. Next, we can make our model more flexible by allowing the *SD*s of assessors to vary. Doing so, we find $ICC = 0.675$ as well; the difference in *ICC* between a model treating the *SD*s as equal (0.6748) and a model treating them as unequal lies in the fourth decimal (0.6754). This is because the difference in *SD*s is small. A LR test (under REML) for the difference in −2RLL (i.e., the equivalent of −2LL but under REML instead of FIML) of a model with equal (770.494) and a model with unequal (770.887) *SD*s yields $\chi^2_1 = 0.393$, $p = 0.531$. The LR test is one at $df = 1$, because the model treating the *SD*s as equal uses 1 *df* for a common *SD* whereas the other model uses 2 *df* since there are two assessors each of which has its own *SD*. The difference is not statistically significant at any meaningful significance level. Hence, in this case, CS does not fit significantly worse than a more flexible model that allows for unequal *SD*s, and with $ICC = 0.675$ Cronbach's alpha would be about 0.806.

Mixed-Effects Analysis (2): Treatment Effect

Now that we have decided on our residual covariance structure, we can estimate the treatment effect (using FIML). In a fully crossed design, the outcome of the treatment effect is the same whether we also include assessor differences and the condition-by-assessor interaction or not. Using SPSS or Jamovi, we find the following outcomes for the fixed effects. For the condition-by-assessor interaction, we find: $F_{1,\ 128} = 0.031$, $p = 0.861$. For the assessor main effect, we find: $F_{1,\ 128} = 0.278$, $p = 0.599$. For the treatment effect of interest, we find: $F_{1,\ 128} = 3.450$, $p = 0.066$. Jamovi returns $R^2_{\mathrm{M}} = 0.023$ for a full factorial model and $R^2_{\mathrm{M}} = 0.022$ for the treatment effect (i.e., a rather small effect). The 95% CI of M_d for the treatment effect (point estimate: 0.375) is [−0.021; 0.771] and the 90% CI is [0.040; 0.710]. For the full factorial model (i.e., with interaction), we find AIC = 774.192 and BIC = 795.463. For a main-effects-only model, we find AIC = 772.223 and

BIC = 789.949. For a treatment-effect-only model (i.e., no differences between assessors), we find AIC = 770.500 and BIC = 784.681. For an assessor-effect-effect-only model (i.e., no differences between conditions), we find AIC = 773.627 and BIC = 787.808. Finally, for the null model, we find AIC = 771.905 and BIC = 782.540. In other words, AIC prefers the treatment-effect-only model, while BIC prefers the null model.

Some may wonder why we are treating assessor as a fixed effect instead of as a random effect. The answer to that is that although the participants who are rated by the assessors are still assumed to be a random sample of a population of possible participants—and hence the *ICC*, which comes down to a participant-level RI model is treated as a random effect and therefore estimated with REML—the assessors in this design constitute a fixed set of assessors available in a given setting where the experiment takes place rather than a random sample of a population of possible assessors. Besides, even if we drew assessors randomly, with only two or an otherwise very small number (e.g., four in Experiment 2), it is difficult to obtain meaningful estimates for generalisation to a wider population of assessors. Finally, even though our primary interest lies in treatment effects, M_ds between specific assessors across conditions provides us with very useful information about whether a treatment effect is more or less the same across assessors in our experiment (i.e., no or minimal condition-by-assessor interaction).

Experiment 2: Several Raters or Assessors

Suppose, we are dealing with a two-group experiment with $n = 100$ participants per condition and the outcome variable is a type of performance rated on an integer scale from 0 (*min*) to 10 (*max*) by four instead of two independent assessors. The assessors are always the same, meaning that all $N = 200$ participants have to pass these four assessors (stations). The exercise of inspecting histograms and scatter plots should now be done for each of four assessors (i.e., histograms of univariate distributions) and for each of six pairs of assessors (i.e., histograms of differences and scatter plots), and again per condition. Suppose, we do these checks and consider that we are good to continue as in Experiment 1. For Assessor 1, we find $M = 5.560$ ($SD = 0.903$) in the control condition and $M = 6.050$ ($SD = 0.857$) in the treatment condition. For Assessor 2, we find $M = 5.500$ ($SD = 1.087$) in the control condition and $M = 5.950$ ($SD = 1.209$) in the treatment condition. For Assessor 3, we find $M = 5.380$ ($SD = 1.196$) in the control condition and $M = 5.890$ ($SD = 1.163$) in the treatment condition. For Assessor 4, we find $M = 5.840$ ($SD = 1.098$) in the control condition and $M = 6.330$ ($SD = 0.943$) in the treatment condition. In other words, we appear to be dealing with differences around $d = 0.5$ in favour of the treatment condition, albeit that this difference fluctuates a little bit across assessors (largest: $d = 0.557$ with Assessor; smallest: $d = 0.391$ with Assessor 2). Table 14.1 presents the Pearson's r between ratings for each pair of assessors for each condition.

Table 14.1 Correlation of ratings for each pair of assessors for each condition (*Jamovi*)

Control		Treatment	
Pair	Pearson's r	Pair	Pearson's r
Assessors 1 and 2	0.710	Assessors 1 and 2	0.763
Assessors 1 and 3	0.653	Assessors 1 and 3	0.675
Assessors 1 and 4	0.774	Assessors 1 and 4	0.716
Assessors 2 and 3	0.427	Assessors 2 and 3	0.557
Assessors 2 and 4	0.567	Assessors 2 and 4	0.502
Assessors 3 and 4	0.570	Assessors 3 and 4	0.494

In both conditions, the correlation varies a bit across pairs of assessors. The correlations presented in Table 14.1 along with the *M*s and *SD*s inform our modelling of the random part, the residual covariance structure.

Mixed-Effects Analysis (1): Residual Covariance Structure Candidates

Using the approach from Experiment 1, we start with a CS model and find (using REML in *SPSS*): $ICC = 0.594$ and a residual variance $V_{\mathrm{RES}} = 1.134$. In this model, we treat $ICC = 0.594$ as holding for all pairs of assessors and $V_{\mathrm{RES}} = 1.134$ (or in its square root form: SD_{residual}) as holding for all assessors. The −2RLL of this model is 2050.661. When we allow V_{RES} to vary across assessors, we find: $ICC = 0.620$, $V_{\mathrm{RES}} = 0.706$ for Assessor 1, $V_{\mathrm{RES}} = 1.363$ for Assessor 2, $V_{\mathrm{RES}} = 1.470$ for Assessor 3, and $V_{\mathrm{RES}} = 1.062$ for Assessor 4. The −2RLL of this model is 1998.641. A LR test of the difference in −2RLL of the two models reported yields: $\chi_3^2 = 52.02$, $p < 0.001$. This test is at $df = 3$ because the difference between the two models is that the restrictive model uses 1 *df* for V_{RES} (i.e., one for all assessors), whereas the more flexible model uses 4 *df* (i.e., one for each assessor). Another more flexible model is to also let the correlation vary across pairs of assessors. This model is also referred to as the *unstructured* (UN) model (e.g., Field, 2018; Tan, 2010; Verbeke & Molenberghs, 2000), because it does not assume any particular structure such as a single *ICC* or a single V_{RES}. Table 14.2 presents the point estimates and 95% CIs of the correlation and variance estimates obtained with this model.

The −2RLL of this model is 1931.983. To compare if this model can be preferred above another somewhat more flexible model (−2RLL = 1998.641), we have to perform a LR test at $df = 5$. This is because both models use 4 *df* for the *V*-estimates but the UN model uses 6 *df* for the *r*-estimates (i.e., one for each pair of assessors) while the other model—like the restrictive CS model—uses 1 *df* for *r*-estimation (i.e., the *ICC* as *r*-estimate for all pairs of assessors). We find: $\chi_5^2 = 66.658$, $p < 0.001$. In other words, we have reasons to prefer the UN model. Although a drawback of this model is that it does not provide a single *r*-estimate

Table 14.2 V_{RES} and r_{RES} estimates in an UN residual covariance model (*SPSS*): point estimates and 95% CIs (LB, UB)

Term	Point estimate	95% LB	95% UB
V_{RES} A1	0.775	0.636	0.943
V_{RES} A2	1.322	1.086	1.610
V_{RES} A3	1.391	1.142	1.693
V_{RES} A4	1.048	0.861	1.277
r_{RES} A1–A2	0.735	0.664	0.793
r_{RES} A1–A3	0.663	0.578	0.734
r_{RES} A1–A4	0.493	0.381	0.591
r_{RES} A2–A3	0.747	0.679	0.803
r_{RES} A2–A4	0.531	0.423	0.623
r_{RES} A3–A4	0.534	0.427	0.626

like the other two models, providing the reader with Table 14.2 is more informative than presenting a single *ICC* of around 0.6 because the *r*-estimate varies considerably across assessor pairs. For all three pairs that involve Assessor 1, the *r*-estimates are okay (quite well above 0.60) while for the three pairs in which Assessor 1 is *not* present the *r*-estimates are somewhat disappointing (quite well below 0.60; e.g., Stemler & Tsai, 2010).

Although the UN model has the advantage of being more flexible than the other models discussed, an issue with this model is that it uses quite a few *df*, given *k* conditions:

$$df = k + [k * (k-1)/2].$$

For $k = 4$, $df = 10$; for $k = 5$, $df = 15$; and for $k = 8$, $df = 36$. In other words, the flexibility of the UN model comes with a loss of *df*, and that loss may be a problem especially when dealing with somewhat larger numbers of assessors. Another approach to the residual covariance structure, which does not use as many *df* as the UN model, but is more flexible than the other two models discussed thus far, is found in combining RIs (see Chap. 13) and RSs. In a RI model, differences between participants on the outcome variable of interest induce a within-participant between-measurements correlation proportional to these differences. The M_d between two assessors can be thought of as a fixed slope, around which participant-specific RSs are allowed to vary following a certain distribution (i.e., often assumed Normal or at least not too different from Normal). Given *k* assessor *M*s, $k - 1$ *M*s can vary freely, and the same can be said about the fixed slopes. Given four assessors, we can compute six fixed slopes for that is the number of pairs of assessors. However, once we have estimated the fixed slope for each of the three pairs that include Assessor 1, we can from there—through subtraction—find the fixed slope for each of the three pairs that do *not* include Assessor 1. This translates into a model for the residual covariance structure as follows: a RI, a RS for the difference between Assessor 1 and Assessor 2, a RS for the difference between Assessor 1 and Assessor 3, and a RS for the difference between Assessor 1

Table 14.3 V_{RES} and RI/RS estimates in a RI-RS model (*SPSS*): point estimates and 95% CIs (LB, UB)

Term	Point estimate	95% LB	95% UB
V_{RES}	0.041	0.015	0.112
V_{RI}	0.721	0.582	0.892
V_{RS} A1–A2	0.541	0.422	0.692
V_{RS} A1–A3	0.703	0.552	0.894
V_{RS} A1–A4	0.394	0.293	0.530

and Assessor 4. This comes down to using 4 *df*: one for the RI variance, and three for the set of RSs. The RI and RS terms may correlate; if we were to estimate all these correlations, we would need 6 *df* more (i.e., there are three RI-RS pairs and three RS-RS pairs) and we would essentially be back to an UN model (−2RLL of 1931.983).

In the experiment at hand, all these correlations between RI and RS terms are in the [−0.1; 0.1] range, in other words fairly close to zero. Table 14.3 presents the outcomes of a RI-RS model with the correlations between these random effects fixed to zero (i.e., we need 4 *df* instead of 10 *df*).

In this model, the heterogeneity between assessors in terms of their variances and across pairs of assessors in terms of correlations is modelled through a combination of RSs and RIs. The −2RLL of this model is 1937.473, a bit higher than that of the UN model but it also uses 6 *df* less than the UN model. A LR test for the difference in −2RLL yields: $\chi^2_6 = 5.49$, $p = 0.483$. Although a statistically non-significant *p*-value cannot and should not be read that two things are 'the same', this LR test outcome indicates that we do not have sufficient reason to reject the RI-RS model as a significant oversimplification relative to the UN model.

Mixed-Effects Analysis (2): Treatment Effect Under Different Assumptions

Some may argue in favour of the UN model over the RI-RS model pointing at the comprehensive outcomes from Table 14.2 and not having to fix correlations between random effects to zero. Others may vote in favour of the RI-RS model because it is also quite intuitive to understand and if two models can explain the same phenomenon to about the same degree, we may as well prefer the simpler model. In the words attributed to statistician George Box that essentially "*all models are wrong, but some are useful*" (Box & Draper, 1987, p. 424), a compromise solution is to provide the outcomes of the fixed part from both an UN perspective and the RI-RS perspective. Both models perform better than the first two models discussed in this chapter but do not really differ from one another in their outcomes even though they are based on partly different assumptions.

From the UN perspective, we find no statistically significant condition-by-assessor interaction effect: $F_{3,\ 200} = 0.054$, $p = 0.984$. The assessor main effect is statistically significant: $F_{3,\ 200} = 15.156$, $p < 0.001$. The treatment effect is positive (M_d = 0.490 points on the 0–10 scale) and also statistically significant: $F_{1, 200}$ = 15.158, p < 0.001, and a 95% CI of M_d = [0.206; 0.774]. For the full factorial model, we find AIC = 1942.244 and BIC = 2026.567. For the main-effects-only model, we find AIC = 1936.405 and BIC = 2006.674. For the treatment-effect-only model, we find AIC = 1971.373 and BIC = 2027.589. For the assessor-effect-only model, we find AIC = 1950.036 and BIC = 2015.621. Finally, for the null model, we find AIC = 1985.005 and BIC = 2036.536. In other words, as expected, both AIC and BIC prefer the main-effects model.

From the RI-RS perspective, we find for the interaction effect: $F_{1, 294.186}$ = 0.054, p = 0.983; for the assessor main effect, we find: $F_{1, 294.186}$ = 15.323, p < 0.001; and for the treatment main effect, we find $F_{1, 224.585}$ = 14.258, p < 0.001, and a 95% CI of M_d = [0.192; 0.788]. For the full factorial model, we find AIC = 1937.789 and BIC = 1998.689. For the main-effects-only model, we find AIC = 1931.952 and BIC = 1978.799. For the treatment-effect-only model, we find AIC = 1968.442 and BIC = 2001.234. For the assessor-effect-only model, we find AIC = 1945.326 and BIC = 1987.488. Finally, for the null model, we find AIC = 1981.733 and BIC = 2009.841. In other words, as expected, both AIC and BIC prefer the main-effects model.

Other Outcome Variables and Alternative Approaches

Although the experiments in this chapter have quantitative outcome variables, the concepts discussed in this chapter also apply to mixed-effects models for categorical outcomes variables (e.g., the extensions of the methods discussed in Chaps. 5–7). In some cases, for example, assessors may not provide quantitative scores but just make pass/fail decisions or share verdicts like poor/sufficient/good. Such cases ask for binary logistic mixed-effects and ordinal logistic mixed-effects models, respectively. The models discussed in this chapter can also be applied to experiments where items instead of assessors are used. If for instance ten items in a post-test are supposed to measure the same knowledge or skill of interest, one would expect differences between treatment and control conditions to be comparable across items; they do not need to be the same, but it would be odd to find for instance a difference in favour of a control group for one item and a difference in favour of a treatment condition for another item. A closer look at the condition-by-item interaction effect can provide a check on (the absence of) such abnormalities.

When three or larger numbers of assessors or items are available, some may opt for a latent variable modelling approach, such as confirmatory factor analysis or some item-response theory model. These approaches have in common with the methods discussed in this chapter that they can help us to estimate treatment effects of interest while appropriately accounting for the data structure. In latent variable approaches, RIs and RSs can be modelled in terms of latent variables, and these approaches are especially useful alternatives to the methods discussed in this chapter when two or more related but different constructs are measured. For example, if the four assessors in Experiment 2 were to rate participants in two related but different constructs, an appropriate latent variable model could account for the anticipated structure in a two-correlated-latent-variables model and provide additional model fit statistics and reliability estimates. In the mixed-effects approach taken in this chapter, the second related construct could be incorporated as a second factor. In Experiment 2, the design is a 2-by-4 design: there are two treatment conditions (between subjects) and four assessors (within subjects). With another variable to be rated by the four assessors, we would obtain a 2-by-2-by-4 design: two treatment conditions (between subjects), and two variables (within subjects) rated by four assessors (within subjects). That said, the more complex our data structures, the more complex our models, and the larger the sample sizes needed to obtain meaningful estimates of all the effects at play.

The methods discussed in this chapter, as well as the latent variable alternatives, can be extended to situations that include some kind of nested structure as discussed in Chap. 13 by adding another level. For instance, suppose that each pair or each group of students in the example experiments in the previous chapter were rated on their performance by two or more assessors. The result would be a three-level design: pair or group as upper level (level 3), participant within pair or group as middle level (level 2), and the assessors as stations to be crossed by all pairs or groups of participants as lower level (level 1). We could then model the residual covariance structure through combinations of RIs and RSs and eventually their intercorrelations at the upper and middle level. Since at the lowest level, we are always back to single observations—in this example, no participant is rated twice on the outcome variable of interest by the same assessor—at the lowest level we cannot estimate RIs or RSs. If we were to expend the three-level design by another level under the assessors, by having each participant rated on the same outcome variable by the same assessors, we would obtain a four-level design, and the idea of combinations of RIs and RSs would then hold for all levels except for the measurements which would then constitute the lowest level 1 (levels 2, 3, and 4 would then be found in the assessors, the participants, and the pairs or groups, respectively). However, to reiterate, the more complex our designs, the more complex our analyses, and the higher the demands on sample size; designs with more than four levels are certainly possible, and corresponding mixed-effects models are available as well, but with such structures we may need several (ten) thousands to obtain meaningful estimates of all the different fixed and random effects to be modelled.

Group-by-Time Interactions

15

Abstract

Whether we deal with different plots of land that receive different treatments in agricultural experiments or we have groups of human participants who are given different treatments in an educational or psychological experiment, when we measure these plots of land or groups of participants at two or more occasions on the same outcome variable of interest we speak of a split-plot design and we can use split-plot analytic methods to analyse differences between plots of land or groups of participants and how these differences increase or decrease from occasion to occasion. In some cases, there are two measurements (i.e., two occasions) one of which takes place immediately after treatment (i.e., post-test, sometimes also referred to as immediate post-test) and the second sometime after (i.e., follow-up). In other cases, one of the measurements takes place before the treatment (i.e., pre-test) and the other takes place after treatment (i.e., post-test). The latter type is also known as pre-test post-test control-group design. In yet other cases, there are three or more measurements; these measurements may either all take place after treatment, or one or some (though not all) of them take place before the treatment. In this chapter, each of these types of situations is discussed with an example experiment.

Introduction

The experiments discussed in Chaps. 13 and 14 have in common that the measurement of a given participant (Chap. 14) or cluster of participants (Chap. 13) takes place a single point in time; no repeated measurements are involved (though different raters can be conceived as different stations to be passed). However, the concepts of mixed-effects analysis discussed in the previous two chapters also apply to split-plot designs. The histograms and scatterplots discussed in Chap. 14 to examine relations

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_15

between different assessors (or items) and to detect eventual extreme combinations of scores (i.e., a well above-average rating by one assessor and a clearly below-average rating by another assessor) can also be used for comparisons of distributions of scores at different occasions. Treatment effects, effects of occasion, and treatment-by-occasion (i.e., group-by-time) interaction effects are fixed effects, and the residual covariance structure can be modelled in ways discussed in Chap. 14. In this chapter, this exercise of estimating random and fixed effects is done with four example experiments: an immediate post-test follow-up control group design (Experiment 1), a pre-test post-test control-group design (Experiment 2), a several post-treatment measurements design (Experiment 3), and a several pre-treatment and several post-treatment measurements design (Experiment 4).

Experiment 1: Immediate Test and Follow Up

Researchers in a Health Professions Education department are interested in the effect of a new type of simulation training on the development of communication skills among undergraduate medical students. The researchers decide to do an experiment with $N = 128$ in which this new simulation training constitutes the experimental treatment condition and the conventional form of simulation serves as control condition. In both conditions ($n = 64$ per condition), participating undergraduate medical students individually undergo training with the same type of simulated patients. The only way in which the two conditions differ is that specific instructions are provided during the training in the treatment condition but not in the control condition. At the end of the training, participants from both conditions individually complete the same post-test with a simulated patient. For simplicity of the example, this post-test yields a quantitative integer score that can range from 0 to 25, and higher scores indicate better post-test performance. A week later, all participants return and complete the same post-test, because the researchers are not just interested in which of the two conditions performs better immediately after training but also how to two conditions compare a week after training. Suppose, the histograms are okay and Fig. 15.1 presents the scatterplot of the bivariate distribution of post-test and follow-up test for the control ($X = 0$) and treatment ($X = 1$) condition.

Pearson's r for the relation between post-test and follow-up test is 0.706 in the control condition and 0.569 in the treatment condition. Box's homogeneity of covariance matrices test is not statistically significant at any meaningful significance level (i.e., 1, 5, or 10%): $\chi^2_3 = 4.119$, $p = 0.249$; and the same holds for Shapiro–Wilk's multivariate normality test: $W = 0.987$, $p = 0.277$ (*Jamovi*).

In the control condition, we find $M = 15.047$ ($SD = 2.019$) for the post-test, $M = 15.156$ ($SD = 2.721$) for the follow-up test, and $M = 0.109$ ($SD = 1.928$) for the gain from post-test to follow-up test. In the treatment condition, we find $M = 16.172$ ($SD = 1.619$) for the post-test, $M = 14.969$ ($SD = 2.678$) for the follow-up test, and $M = -1.203$ ($SD = 2.205$) for the gain from post-test to follow-up test.

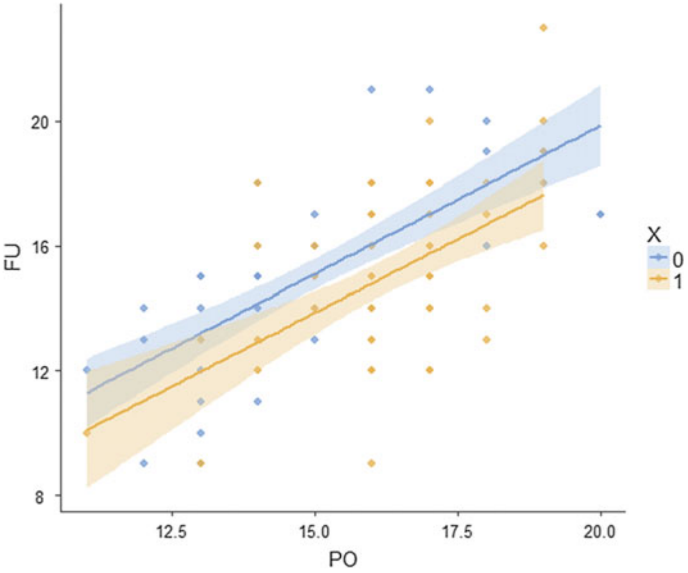


Fig. 15.1 Scatterplot of the bivariate distribution of post-test (*PO*) and follow-up test (*FU*) for the control ($X = 0$) and treatment ($X = 1$) condition in Experiment 1 (*Jamovi*)

Mixed-Effects Analysis (1): Unequal Residual Variance

Like in the experiments discussed in Chaps. 13 and 14, it is important to account for the residual covariance structure and hence *not* to assume all observations to be independent. Erroneously assuming independence of repeated measurements would result in too wide CIs and too high *p*-values for occasion main effects and condition-by-occasion (i.e., group-by-time) interaction effects, because in models that assume independence of all observations no distinction is made between variance *within* and variance *between* participants (i.e., everything is thrown on one common error pile; e.g., Leppink, 2015a; Tan, 2010). With only two measurements, a likely first candidate is CS aka a RI model. Using REML in *SPSS* or *Jamovi*, we find $ICC = 0.597$. The resulting −2RLL is 1097.485. Next, we run a model in which the V_{RES}s of the two occasions are allowed to differ. We find $ICC = 0.642$, $V_{\text{RES}} = 3.349$ for the post-test and $V_{\text{RES}} = 7.289$ for the follow-up test. The resulting −2RLL is 1067.305. A LR test for the difference in −2RLL yields: $\chi_1^2 = 30.180$, $p < 0.001$. In other words, the more flexible model appears to be preferred.

Mixed-Effects Analysis (2): Main and Interaction Effects

To reiterate from Chaps. 13 to 14, in the process of finding an appropriate model for the residual covariance structure (i.e., modelling the random part of our mixed-effects models), we should include our fixed effects of interest as well. RIs, RSs, V_{RES}s, and random effects are based on fluctuations around fixed intercepts and slopes, so we should include these fixed intercepts and slopes in our models as well when estimating random effects. In this case, that comes down to: a main effect of occasion, a main effect of condition, and a condition-by-occasion interaction effect. However, at the stage of estimating the random effects, we use REML, and once we turn to estimating fixed effects, we use FIML. For the full factorial model, we find AIC = 1077.315 and BIC = 1102.132. For the main-effects model, we find AIC = 1087.744 and BIC = 1109.015. For the treatment-effect-only model, we find AIC = 1095.495 and BIC = 1113.221. For the occasion-effect-only model, we find AIC = 1093.716 and BIC = 1111.442. Finally, for the null model, we find AIC = 1101.467 and BIC = 1115.648. Thus, both AIC and BIC prefer the full factorial model. Although the treatment condition performs on average 1.125 points higher on the post-test (the 95% CI ranges from 0.490 to 1.760), that favour has vanished completely after one week: the regression coefficient of this interaction effect is −1.313, and the corresponding 95% CI ranges from −2.031 to −0.594. In other words, evidence in favour of a positive effect of treatment is found only immediately after training but not one week later. This phenomenon is not uncommon in studies that involve learning, yet relatively few experiments in educational research actually include a follow-up test.

Experiment 2: Pre and Post

A second type of split-plot design is found in a pre-test post-test control-group design. In this case, one or more treatment conditions and a control condition are compared in terms of an outcome variable that is measured first *prior to* and then *after* treatment. Although this type of design can be quite problematic in studies on learning because pre-testing can in itself influence learning and as such count as an intervention, there are plenty of outcome variables (e.g., motivation, effort, interest, hours of sleep, blood pressure in psychopharmacological studies) that may lend themselves well for this type of design. There has been quite a bit of debate in the literature with regard to how to analyse data obtained in such a design. Although not so few researchers are inclined to use repeated-measures analysis as in Experiment 1—where *all* measurements are done *after* the treatment—this is, for randomised controlled experiments, not the best approach. The explanation for this is as follows.

Regression to One Common Mean Versus Regression to Different Means

In a quasi-experiment, where pre-existing instead of randomised groups receive different treatments, or in experiments like Experiment 1 in this chapter where all measurements take place after treatment, groups will usually differ on the outcome variable of interest prior to the first measurement. In each group or condition, scores from measurement at subsequent occasions will than show fluctuations following a pattern of regression to their group-specific or condition-specific mean M_c. However, as Twisk et al. (2018, p. 80) note: "*When differences at baseline between the treatment and control group are due to random fluctuations and measurement error, there is a tendency of the average value to go down in the group with the initial highest average value and to go up in the group with the initial lowest average value*." After all, in a true experiment—where, contrary to quasi-experiments—we randomly allocate our participants to the different conditions, *prior to treatment* there is regression to a *common mean M*. As rightly argued by Twisk et al. (2018), not adjusting for the baseline difference in the outcome variable is likely to result in the estimation of an artificial treatment effect. Absent treatment, the participants allocated to the different conditions have been (randomly) sampled from the same population, and any deviation from equal *M*s on an outcome variable prior to any treatment is entirely due to the laws of probability.

Although repeated-measures analysis makes sense in Experiment 1 in this chapter as well as in quasi-experiments with pre-existing groups—albeit it that, for causal inference, quasi-experiments should not be treated as if they were true experiments—it fails to appreciate this characteristic of regression towards a common mean *M* prior to any treatment when (at least) one measurement takes place prior to treatment. In the repeated-measures analysis, where differences between conditions are expected and estimated on the first measurement, we assume regression to condition-specific M_cs instead of to one common *M* (e.g., Van Breukelen, 2006). The regression equation of (the fixed part of) the repeated-measures model for the predicted score *S* on the outcome variable is:

$$S = B_0 + [B_1 * \text{Condition}] + [B_2 * \text{Occasion}] + [B_3 * \text{Condition-by-Occasion}].$$

In this equation, B_0 is the expected score for a participant in the control condition (i.e., Condition = 0) at the first occasion (i.e., Occasion = 0), B_1 is the expected difference in (average) score between conditions at the first occasion, B_2 is the expected change in score in the control condition from first to second occasion (i.e., from Occasion = 0 to Occasion = 1), and B_3 is the difference in expected change in score between conditions. When the first measurement occasion is before the treatment, B_1 is expected to be 0, and any deviation from 0 in a given experiment is entirely due to the laws of probability. When we include B_1, we assume regression to M_cs; when we exclude B_1, we assume regression to *M*. When the first measurement takes place after the treatment (or we deal with pre-existing groups in a quasi-experiment), leaving out B_1 comes at the risk of substantially biased treatment effect estimates, because we erroneously assume regression to M instead of to M_cs. However, in a randomised controlled experiment, when the first occasion is measured before treatment, the assumption of regression to M instead of to M_cs is correct, and hence B_1 should be omitted from the aforementioned equation to obtain the following:

$$S = B_0 + [B_1 * \text{Occasion}] + [B_2 * \text{Condition-by-Occasion}].$$

This equation is the same as the previous but without baseline (i.e., Occasion = 0) difference; B_0 is the common M of all conditions prior to treatment (i.e., at Occasion = 0). The treatment effect can be directly obtained from B_2. In other words, in this model, we have *three* instead of four fixed-effects terms: intercept (B_0), occasion (time: B_1), and condition-by-occasion interaction (the treatment effect of interest: B_2). This second model is like an ANCOVA but, as any mixed-effects model, can handle missing data. Suppose that there are a few participants who do respond at pre-test but not at post-test; in the classical ANCOVA model, these cases are omitted from the analysis unless we apply some form of imputation; in the mixed-effects model here, FIML can (under MAR or MCAR) use all available data and account for the missing data without losing cases and without imputation (see also Chap. 4). For an excellent explanation of why the classical ANCOVA model and this adjusted mixed-effects model—where B_0 serves as common M of all conditions prior to treatment—are to be preferred over the aforementioned repeated-measures approach or a simple change test on pre-post differences, see Twisk et al. (2018).

Why Statistical Tests for Baseline Differences Do Not Make Sense

Many researchers use to argue that an adjustment for baseline differences, through ANCOVA or the adjusted mixed-effects model, only makes sense when the difference at baseline (in our example: pre-test) is statistically significant, and until some years ago I used to think that way as well. However, this approach does not make sense, since given randomisation all differences are baseline are by definition entirely the result of the laws of probability and not of any 'no difference' null hypothesis being incorrect. Given that prior to treatment all participants are from the same source population, the null hypothesis of 'no difference' at baseline is always true, and anytime we find a statistically significant difference we deal with a Type I error or *artefact*. Regardless of what statistical testing criteria we use—p-values, AIC, BIC, BF, or other—testing for baseline differences in a true experiment does not make sense. Therefore, while in quasi-experiments with pre-existing groups or in experiments where the first measurement takes place after treatment the adjusted model (i.e., with B_0 as common M) is not recommended because it is likely to result in substantially biased treatment effect estimates (e.g., Van Breukelen, 2006), in experiments where the first measurement takes place before treatment this model can be expected to be the best option regardless of whether or not the baseline difference is statistically significant.

Different Types of Outcome Variables

The correction discussed here is important for all experiments that involve quantitative outcome variables. Besides, although the baseline constitutes the most important covariate, the regression to M instead of to M_cs argument can also be considered for other relevant covariates (e.g., Kahan, Jairath, Doré, & Morris, 2014). However, when the outcome variable of interest is dichotomous, the situation is more complicated than when we are dealing with quantitative outcome variables. In a linear model, the unexplained variance of the outcome variable of interest decreases when we add a covariate and the explained variance increases with the same amount. In a logistic model, on the contrary, the unexplained variance is fixed; consequently, adding a covariate that contributes to the explanation of the outcome variable of interest will result in an increase of the *total* variance as a sum of explained variance (which increases) and (fixed) unexplained variance (e.g., Twisk et al., 2018). A consequence of this difference between linear and logistic models is that even when the baseline values of two conditions are the same and the baseline value is clearly related to the outcome at later occasions, baseline-difference-unadjusted and -adjusted mixed-effects binary logistic regression models will differ (i.e., non-collapsibility; e.g., Hernan, Clayton, & Keiding, 2011; Newman, 2004). Therefore, when dealing with dichotomous outcome variables, an adjustment for baseline differences in the outcome variable of interest as argued for in the case of quantitative outcome variables mostly appears not necessary (Twisk et al., 2018).

Two Types of Interaction: Condition-by-Occasion and Condition-by-Baseline

Although ANCOVA and the adjusted mixed-effects model constitute more appropriate approaches to pre-test post-test control-group design data, they have in common with the typical repeated-measures and change approaches that they focus on a group-by-time or condition-by-occasion interaction but do *not* consider the possibility that the effect of treatment on the outcome variable of interest may differ substantially across the range of baseline scores: *condition-by-baseline interaction*. In the case of the latter, moderated regression (see also Chap. 12) can provide a more nuanced picture of a treatment effect of interest. Let us look at this a bit closer with an example. Suppose that we are redoing Experiment 1 but with a pre-test and post-test instead of with a post-test and follow-up test; the pre-test is a training session that participants usually take at their stage in the curriculum, then they undergo condition-specific training as in Experiment 1, and immediately after, they do the post-test.

Interaction (1): Condition-by-Occasion

Suppose, the histograms of the distributions pre-test, post-test and pre-post difference are fine and that we find: M = 6.953 (SD = 1.578) for pre-test and M = 16.891 (SD = 1.920) for post-test in the control condition, and M = 6.734 (SD = 1.576) for pre-test and M = 17.531 (SD = 1.781) for post-test in the treatment condition. The scatterplot in Fig. 15.2 presents the bivariate distribution for the relation between pre-test (baseline) and post-test for the control (X = 0) and treatment (X = 1) condition.

Pearson's r for the correlation between pre-test and post-test is 0.355 in the control condition and 0.718 in the treatment condition. Using *SPSS*, in ANCOVA with treatment as factor and pre-test (baseline) as covariate, we find for the treatment effect of interest: B = 0.777, p = 0.006, with a 95% CI of [0.223; 1.330]. In the adjusted mixed-effects model, assuming CS, we find for the treatment effect of interest: B = 0.755, p = 0.004, with a 95% CI of [0.247; 1.263]. In the adjusted mixed-effects model, allowing for unequal V_{RES} for pre-test and post-test (i.e., LR test indicates that this residual covariance structure is to be preferred over CS:

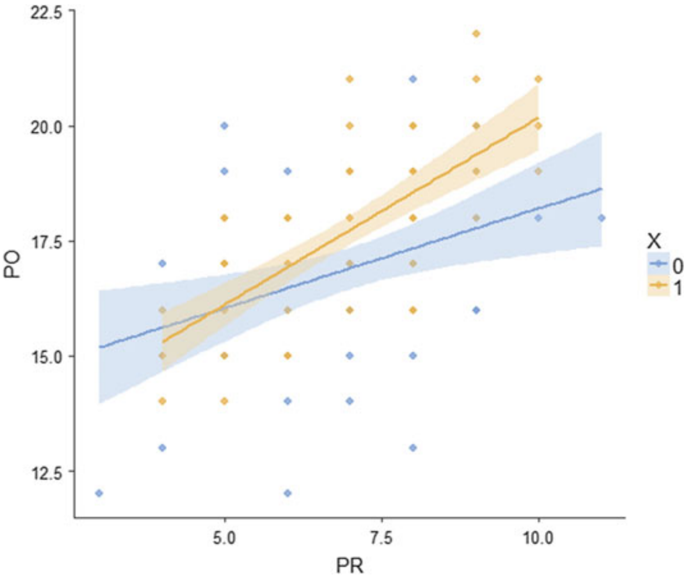


Fig. 15.2 Scatterplot of the bivariate distribution of pre-test (PR) and post-test (PO) for the control (X = 0) and treatment (X = 1) condition in Experiment 2 (*Jamovi*)

$\chi_1^2 = 4.556$, $p = 0.033$), we find for the treatment effect of interest: $B = 0.777$ (i.e., same as for ANCOVA given no missing data), $p = 0.006$, with a 95% CI of [0.231; 1.322]. In other words, a positive treatment effect.

Interaction (2): Condition-by-Baseline

The methods discussed in the previous paragraph do not go beyond the condition-by-occasion interaction and therefore provide only a *single* treatment effect estimate. However, in Fig. 15.2, we see some signs of a condition-by-baseline interaction effect: learners with above-average baseline scores appear to benefit from the treatment more than their below-average baseline peers. To examine whether a model accounting for the apparent condition-by-baseline interaction seen in Fig. 15.2 is recommended, we can use AIC/BIC model comparison approach as in Chap. 12, Experiment 5. For Model 0 (null model), we find AIC = 526.830 and BIC = 532.534. For Model 1 (condition as only predictor), we find AIC = 524.998 and BIC = 533.554. For Model 2 (baseline as only predictor), we find AIC = 490.624 and BIC = 499.180. For Model 3 (main effects of condition and baseline), we find AIC = 484.960 and BIC = 496.368. Finally, for Model 4 (full factorial, with condition-by-baseline interaction), we find AIC = 482.211 and BIC = 496.471. Performing a classical significance test on the interaction effect (*JASP*), we find: $B = 0.380$, $p = 0.032$, with a 95% CI of [0.033; 0.728]. If we test at $\alpha = 0.05$, AIC and the outcome of the statistical significance test indicate in favour of Model 4. BIC indicates a slight preference for Model 3. In a Bayesian ANCOVA in *JASP* (default prior), we find a BF of about 1.43 in favour of Model 4 over Model 3, in other words: not spectacular. The effect size estimates of the interaction effect (*JASP* or *Jamovi*) indicate that the condition-by-baseline interaction effect is rather small: $\eta^2 = 0.026$, partial $\eta^2 = 0.036$, and $\omega^2 = 0.020$. Some researchers may argue, based on combinations of BIC, BF, and/or effect size estimates, that we do not need to go beyond ANCOVA. Others may argue that we need to account for the interaction, and a middle solution is to report both. Applying the simple slope analysis approach from Chap. 12 (Experiment 5) to the data from the experiment at hand, we find the following outcomes. For one *SD* below the average of pre-test, we find: $B = 0.180$, $p = 0.643$, and 95% CI = [−0.581; 0.941]. For average pre-test, we find: $B = 0.777$, $p = 0.005$, and 95% CI = [0.236; 1.317]. For one *SD* above the average of pre-test, we find: $B = 1.373$, $p < 0.001$, and 95% CI = [0.612; 2.135]. In other words, we have insufficient evidence to suggest a treatment effect at one *SD* below the average of pre-test.

Experiment 3: Several Post-treatment Measurements

Variants on Experiments 1–2 are found in having more than two measurements of the same outcome variable of interest. For instance, in Experiment 2, researchers could have chosen for a pre-test, post-test, and follow-up test. The arguments concerning regression to one versus several means, baseline corrections, and the two types of interaction also apply to these extensions, and can be dealt with by extensions of the models discussed in the context of Experiment 2. Another variant on Experiment 1 is to have more than two measurements after treatment and no measurements prior to treatment. Suppose, we are dealing with three different randomised groups (n = 53 per group) in an experiment that focusses on differences between two treatment conditions (Groups 1 and 2) and a control condition (Group 3) in a number of outcome variables including the experienced difficulty of learning tasks carried out as part of the grammar training. In each of three conditions, participants complete four learning tasks and rate the experienced difficulty of each task right after completing it on a VAS from 0 (min) to 100 (max). Suppose, the histograms of distributions at each occasion and those of the differences in scores between occasions look fine, and the findings are as follows. In Group 1, we find: M = 7.604 (SD = 1.964) after Task 1, M = 11.019 (SD = 2.422) after Task 2, M = 12.208 (SD = 2.938) after Task 3, and M = 12.019 (SD = 3.692) after Task 4. In Group 2, we find: M = 7.698 (SD = 2.198) after Task 1, M = 11.340 (SD = 3.019) after Task 2, M = 12.302 (SD = 3.719) after Task 3, and M = 12.547 (SD = 4.107) after Task 4. In Group 3, we find: M = 7.660 (SD = 2.019) after Task 1, M = 10.623 (SD = 3.001) after Task 2, M = 11.792 (SD = 3.466) after Task 3, and M = 11.660 (SD = 4.296) after Task 4. Figure 15.3 presents Ms and SDs of experienced task difficulty per condition (group) after each task.

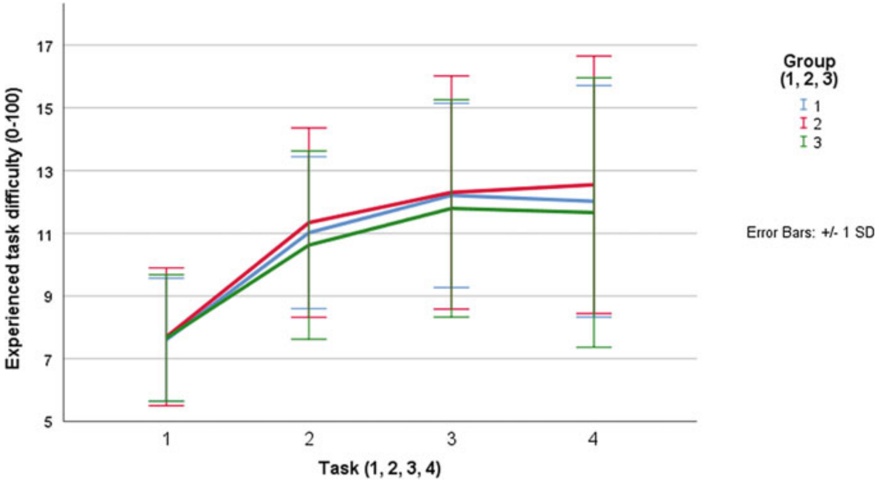


Fig. 15.3 Ms and SDs (error bars) of experienced difficulty per condition after each task in Experiment 3 (SPSS)

In short, small differences after Task 1, increases in *M*s and M_ds between conditions from Task 1 to Task 2, some increases in *M*s from Task 2 to Task 3 as well, and generally *SD*s increase from task to task. The researchers are interested in the main effect of time, the main effect of group, and the group-by-time (i.e., condition-by-task) interaction effect.

Suppose, the scatterplots of the bivariate relations for each pair of tasks do not indicate any strange patterns (normally, this would involve six scatterplots: one for each of six pairs of tasks). Box's homogeneity of covariance matrices test is not statistically significant at any meaningful significance level: $\chi^2_{20} = 17.832$, $p = 0.598$; and the same holds for Shapiro-Wilk's multivariate normality test: $W = 0.989$, $p = 0.253$ (*Jamovi*).

Mixed-Effects Analysis (1): Random Part

With four tasks in a particular temporal order, we have several options, and even more so when the tasks are equidistant in time (i.e., equal time intervals between adjacent tasks). Four possible residual covariance structures can be taken from Chap. 14 (Experiment 2): CS, its extension allowing the V_{RES} to vary across tasks, RIs and/or RSs (not all may be needed), and UN. When we are dealing with equidistant tasks, other possible candidates for the residual covariance structure are the following. Firstly, with tasks in a given order, it is well possible that the residuals of adjacent tasks are correlated more strongly than the residuals of non-adjacent tasks, and that the residual correlation (r_{RES}) decreases as the distance between tasks decreases. In one type of residual covariance structure, named after the German mathematician Otto Toeplitz (1881–1940) and therefore also known as *Toeplitz* covariance structure (e.g., Bareiss, 1969; Littell, Pendergast, & Natarajan, 2000; Lu & Mehrotra, 2009), one r_{RES} is estimated for adjacent tasks, one for tasks that have one other task in between them, one for tasks that have two other tasks in between them, and so forth. For four tasks, this comes down to three r_{RES} estimates: one for tasks 1 and 2, tasks 2 and 3, and tasks 3 and 4; one for tasks 1 and 3 and tasks 2 and 4; and one for tasks 1 and 4. There are two variants of Toeplitz: equal and unequal V_{RES} across tasks. The same holds for another popular type of residual covariance structure: AR1 (e.g., Lu & Mehrotra, 2009; Tan, 2010). In AR1, we need only 1 *df* for r_{RES}, because r_{RES} between a given pair of tasks is a simple mathematical function of the number of tasks *k* between that pair of tasks:

$$r_{\text{RES}} = r^{k+1}.$$

For adjacent tasks—in our case: 1 and 2, 2 and 3, and 3 and 4—there are no tasks between them, hence $k = 0$, and $r_{\text{RES}} = r$. For the next shortest distance—tasks 1 and 3 and tasks 2 and 4—there is one task in between, hence $k = 1$, and $r_{\text{RES}} = r^2$. Finally, for tasks 1 and 4, there are two tasks in between, hence $k = 2$, and $r_{\text{RES}} = r^3$. Given that $r < 1$, r_{RES} by definition *decreases* according to the power function with the number of tasks between a pair of tasks. Sometimes, a combination of RI and AR1 can be used (e.g., Tan, 2010). Whether we assume equal or unequal V_{RES} across tasks, Toeplitz and AR1 only make sense when we have equidistance between tasks. After all, what reason would we have to assume r_{RES} between tasks 1 and 2 to be the same for 2 and 3 and for 3 and 4 if the distances are different? Further, given that both CS and AR1 only use 1 *df* for r_{RES}, they are often too simplistic. Simultaneously, while the UN model may be one of the most likely candidates when dealing with only a few tasks, Toeplitz may well constitute a good compromise between the too simplistic CS and AR1 on the one hand and the *df*-consuming UN on the other hand (e.g., Lu & Mehrotra, 2009). When tasks are not equidistant, combinations of RI and RS terms may constitute a useful alternative to the UN model. Not all RI and RS terms may be needed. For instance, in this experiment, the RS variance for the change from Task 1 to Task 2 is as close to zero that we may as well call it zero. Also, we do not always need to estimate all correlations between all RI and RS terms or treat all these correlations as being zero; in some cases, for instance, a single average correlation between the different RI and RS terms may do. Finally, there are at least two other covariance structures worth considering when dealing with this kind of designs: first-order ante-dependence (AD1) and HF (e.g., Eyduran & Akbaş, 2010). In HF, all r_{RES}s are a function of the V_{RES}s of the two tasks that constitute a pair and a constant factor λ that holds for all pairs. In other words, it uses as many *df* as CS with unequal V_{RES} across tasks and as many *df* as AR1 with unequal V_{RES} across tasks (i.e., the number of tasks plus one). In AD1, we allow the V_{RES} to vary across tasks and we estimate a r_{RES} for each pair of adjacent tasks; r_{RES} of any pair of non-adjacent tasks is the product of all adjacent correlations in between. This way, AD1 may constitute a more parsimonious alternative to the UN model that comes with the advantage over AR1 and Toeplitz in that it does not need tasks to be equidistant.

Suppose, the tasks in this experiment are equidistant. Table 15.1 provides a comparison between different residual covariance structures, using the same fixed part (i.e., main effect of condition, main effect of task, and condition-by-task interaction effect): (R1) UN; (R2) RI aka CS; (R3) CS with unequal V_{RES} across tasks; (R4) RI, RS for the difference between Task 1 and Task 3, and RS for the difference between Task 1 and Task 4; (R5) the same as R4 but with an average correlation between these random effects; (R6) Toeplitz; (R7) Toeplitz with unequal V_{RES} across tasks; (R8) AR1; (R9) AR1 with unequal V_{RES} across tasks; (R10) AD1; and (R11) HF.

In this comparative approach, we use the UN model—the most flexible one of all—as starting point and see how all other models, which are less flexible than the UN model in one way or another, compare. A statistically significant LR test outcome in that case indicates a significant loss of information and hence not a good simplification of the UN model. In this case, the only model that does not yield a statistically significant LR test outcome is AD1. How come AD1 performs so well? The r_{RES}s estimated in the UN model are as follows: 0.710 for Tasks 1 and 2, 0.598 for Tasks 1 and 3, 0.837 for Tasks 2 and 3, 0.550 for Tasks 1 and 4, 0.740 for Tasks 2 and 4, and 0.860 for Tasks 3 and 4. When we multiply 0.710 by 0.837, we obtain 0.594 as estimate for r_{RES} for Tasks 1 and 3, which is almost the same as the 0.598

Table 15.1 Comparison of eight residual covariance structures in Experiment 3 (SPSS): (R1) UN, (R2) RI aka CS, (R3) CS with unequal V_{RES} across tasks; (R4) RI, RS for the difference between Task 1 and Task 3, and RS for the difference between Task 1 and Task 4; (R5) the same as R3 but with an average correlation between these random effects; (R6) Toeplitz; (R7) Toeplitz with unequal V_{RES} across tasks; (R8) AR1; (R9) AR1 with unequal V_{RES} across tasks; (R10) AD1; and (R11) HF: number of estimated parameters (df), −2RLL, and LR test χ^2_{df} and *p*-value for any of R2–R11 tested against R1

Structure	*df*	−2RLL	LR test χ^2_{df}	LR test *p*-value
R1	10	2675.520	–	–
R2	2	2889.728	$\chi^2_8 = 214.208$	<0.001
R3	5	2771.988	$\chi^2_5 = 96.468$	<0.001
R4	4	2848.548	$\chi^2_6 = 173.028$	<0.001
R5	5	2762.062	$\chi^2_5 = 86.542$	<0.001
R6	4	2765.684	$\chi^2_6 = 90.164$	<0.001
R7	7	2695.772	$\chi^2_3 = 20.252$	<0.001
R8	2	2768.692	$\chi^2_8 = 93.172$	<0.001
R9	5	2696.030	$\chi^2_5 = 20.510$	0.001
R10	7	2676.862	$\chi^2_3 = 1.342$	0.719
R11	5	2796.654	$\chi^2_5 = 121.134$	<0.001

obtained in the UN model. Likewise, when we multiply, 0.837 by 0.860, we obtain 0.720 as estimate for r_{RES} for Tasks 2 and 4, which is not far away from the 0.740 obtained in the UN model. Finally, the multiplication 0.710 * 0.837 * 0.860 ≈ 0.511 as estimate for r_{RES} for Tasks 1 and 4, which is not far away from the 0.550 obtained in the UN model.

If the most complex model (UN: *df* = 10) and a simplification of that model (AD1: *df* = 7) perform about equally well, we may as well prefer the simplification. Therefore, based on the outcomes of Table 15.1, we may as well prefer AD1. Perhaps some others will still prefer UN. The middle solution is to provide the fixed-effects outcomes for both.

Mixed-Effects Analysis (2): Fixed Part

Using the UN model for the random part, we find the following outcomes for the fixed effects (SPSS). For the main effect of task, we find: $F_{3,\ 159} = 171.411$, $p < 0.001$. For the main effect of condition, we find: $F_{2,\ 159} = 0.511$, $p = 0.601$. For the condition-by-task interaction effect, we find: $F_{6,\ 159} = 0.849$, $p = 0.534$. Using the AD1 model for the random part, we find the following outcomes for the fixed effects (SPSS). For the main effect of task, we find: $F_{3,\ 222.480} = 170.919$, $p < 0.001$. For the main effect of condition, we find: $F_{2,\ 160.351} = 0.515$, $p = 0.598$. For the condition-by-task interaction effect, we find: $F_{6,\ 222.480} = 0.858$, $p = 0.526$.

In other words, either way insufficient evidence for an interaction effect or for a treatment main effect. The only fixed effect for which we find sufficient evidence is for the main effect of task, and Fig. 15.3 illustrates what that effect looks like.

Experiment 4: Several Pre-treatment and Several Post-treatment Measurements

An in experimental educational and psychological research slightly less common type of design (i.e., compared to the other designs discussed in this chapter) is that where both *before* and *after* treatment there are *multiple* measurements on the same outcome variable of interest. The methods discussed for Experiments 1–3 still apply, but the control for baseline differences discussed in Experiment 2 now applies for *several* baseline measurements. Consider the following example. Some traffic psychologists are interested in the effect of a new type of drug on—among others—cognitive performance within the first half an hour after consumption and want to test this by a comparing a treatment group (n = 51) and control group (n = 51) in performance on a computer task that requires a participant to use the front (speed up), back (slow down), left (steer left), and right (steer right) arrows on a computer keyboard to drive a car on a very busy four-lane highway that randomly bends left and right, sometimes more sometimes less firmly, and is full of cars that occasionally move to another lane to overtake a car in front of them. The participant is supposed to avoid hitting any of the cars, to respect a distance between cars in front of them (i.e., to overtake timely), and to not scratch the walls on the left and right of the highway. Depending one's performance, the score on this 5-min task is somewhere between 0 (min) and 25 (max). Based on previous research, the researchers argue that to accurately estimate the effect of the drug and not have too much noise from inexperience with this type of task, participants should first do three *practice trials* separated by 3-min breaks, then receive condition-specific treatment (i.e., another 3-min break)—an orange juice in the control condition, and an orange juice with a standardised dose of the drug in the treatment condition—and then do another four *post-treatment trials* separated by 3-min breaks. Suppose, the usual histograms and scatterplots indicate no abnormalities and Fig. 15.4 presents *M*s and *SD*s of task performance per condition (group) after each task (i.e., Tm2, Tm1, and T0 being the three practice trials and T1–T4 being the post-treatment trials).

At Tm2, we find M = 8.745 (SD = 2.096) in the control condition and M = 8.725 (SD = 2.246) in the treatment condition. At Tm1, we find M = 9.902 (SD = 2.274) in the control condition and M = 9.863 (SD = 2.458) in the treatment condition. At T0, we find M = 10.078 (SD = 2.505) in the control condition and M = 9.863 (SD = 2.254) in the treatment condition. At T1, we find M = 9.874 (SD = 2.602) in the control condition and M = 11.020 (SD = 2.429) in the

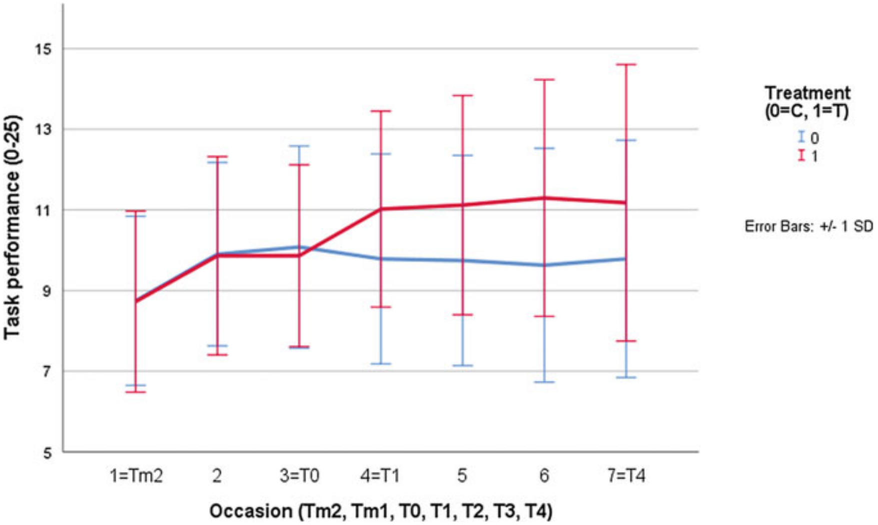


Fig. 15.4 *M*s and *SD*s (error bars) of task performance per condition after each task in Experiment 3 (*SPSS*)

treatment condition. At T2, we find *M* = 9.745 (*SD* = 2.606) in the control condition and *M* = 11.118 (*SD* = 2.718) in the treatment condition. At T3, we find *M* = 9.627 (*SD* = 2.898) in the control condition and *M* = 11.294 (*SD* = 2.935) in the treatment condition. Finally, at T4, we find *M* = 9.784 (*SD* = 2.942) in the control condition and *M* = 11.176 (*SD* = 3.428) in the treatment condition. The standard deviations increase somewhat over time and, as to be expected, the conditions only start to visibly differ after treatment (i.e., starting T1); any differences between conditions on the practice trials is a matter of random fluctuation, while differences on the last four trials can result from treatment differences along with random fluctuation. Therefore, the following regression equation of (the fixed part of) the repeated-measures model for the predicted score *S* on the outcome variable can be used:

$$\begin{aligned} S = {} & B_0 + [B_1 * \mathrm{Tm1}] + [B_2 * \mathrm{T0}] + [B_3 * \mathrm{T1}] + [B_4 * \mathrm{T2}] + [B_5 * \mathrm{T3}] + [B_6 * \mathrm{T4}] \\ & + [B_7 * \mathrm{T1} * \mathrm{Treat}] + [B_8 * \mathrm{T2} * \mathrm{Treat}] + [B_9 * \mathrm{T3} * \mathrm{Treat}] + [B_{10} * \mathrm{T4} * \mathrm{Treat}]. \end{aligned}$$

This model is nothing more than an extension of the adjusted mixed-effects model regression equation in Experiment 2.

Mixed-Effects Analysis (1): Random Part

Table 15.2 provides a comparison between different residual covariance structures, using the same fixed part (cf. the regression equation just mentioned): (R1) UN;

Table 15.2 Comparison of eight residual covariance structures in Experiment 4 (*SPSS*): (R1) UN; (R2) CS; (R3) CS with unequal V_{RES} across tasks; (R4) RI, and RS for the difference between first and last task; (R5) the same as R4 but with correlation between these random effects; (R6) Toeplitz; (R7) Toeplitz with unequal V_{RES} across tasks; (R8) AR1; (R9) AR1 with unequal V_{RES} across tasks; (R10) AD1; and (R11) HF: number of estimated parameters (*df*), −2RLL, and LR test χ^2_{df} and *p*-value for any of R2–R11 tested against R1

Structure	*df*	−2RLL	LR test χ^2_{df}	LR test *p*-value
R1	28	2225.138	–	–
R2	2	2572.709	χ^2_{26} = 347.571	<0.001
R3	8	2508.713	χ^2_{20} = 283.575	<0.001
R4	3	2544.560	χ^2_{25} = 319.422	<0.001
R5	4	2533.052	χ^2_{24} = 307.914	<0.001
R6	7	2266.993	χ^2_{21} = 41.855	0.013
R7	13	2239.182	χ^2_{15} = 14.044	0.522
R8	2	2269.031	χ^2_{26} = 43.893	0.016
R9	8	2241.338	χ^2_{20} = 16.200	0.704
R10	13	2238.200	χ^2_{15} = 13.062	0.598
R11	8	2544.477	χ^2_{20} = 319.339	<0.001

(R2) CS; (R3) CS with unequal V_{RES} across tasks; (R4) RI, and RS for the difference between first and last task; (R5) the same as R4 but with correlation between these random effects; (R6) Toeplitz; (R7) Toeplitz with unequal V_{RES} across tasks; (R8) AR1; (R9) AR1 with unequal V_{RES} across tasks; (R10) AD1; and (R11) HF.

Four solutions appear to deserve a closer look: R1 (−2RLL = 2225.138), R7 (−2RLL = 2239.182; χ^2_{15} = 14.044, p = 0.522), R9 (−2RLL = 2241.338; χ^2_{20} = 16.200, p = 0.704), and R10 (−2RLL = 2238.200; χ^2_{15} = 13.062, p = 0.598). Each of R7, R9, and R10 can be tested against R1, because each of these three alternatives to R1 is a simplification aka *special case* of R1. R7 and R10 *cannot* be viewed in these terms; neither is a special case of the other, and besides the use the same number of *df*; for these reasons, a LR test for a comparison between R7 and R10 is not an option. However, R9 *can* be viewed as a special case of R7 and as a special case of R10: these three models all allow the V_{RES} to vary across tasks, but R9 uses only 1 *df* for r_{RES} while R7 and R10 use several *df* for estimating r_{RES}s. When we compare R9 against R7, we find: χ^2_5 = 2.156, p = 0.827. When we compare R9 against R10, we find: χ^2_5 = 3.138, p = 0.679. In other words, we do not have sufficient evidence to reject the assumption that R9 is a valid simplification of R1, R7, and R10. In R9, the r_{RES} for adjacent tasks is estimated to be 0.920, and from here r_{RES} for any pair of non-adjacent tasks can be computed following the power function mentioned in Experiment 3 (e.g., Tasks 1 and 3: 0.846; Tasks 1 and

Table 15.3 Fixed-effect outcomes in Experiment 4 (*SPSS*): *B*, *df*, *p*, and 95% CI (LB, UB)

Term	*B*	*df*	*p*-value	95% LB	95% UB
B_0	8.735	110.019	<0.001	8.298	9.173
B_1	1.147	282.964	<0.001	0.962	1.332
B_2	1.235	390.451	<0.001	0.982	1.488
B_3	0.947	426.325	<0.001	0.583	1.310
B_4	0.913	433.994	<0.001	0.467	1.358
B_5	0.795	375.994	0.004	0.260	1.331
B_6	0.952	271.734	0.003	0.319	1.584
B_7	1.440	337.605	<0.001	1.059	1.821
B_8	1.567	339.165	<0.001	1.033	2.101
B_9	1.860	308.420	<0.001	1.180	2.540
B_{10}	1.587	234.199	<0.001	0.758	2.416

7: 0.606). V_{RES} is estimated to be 5.016 for Task 1, 5.979 for Task 2, 5.949 for Task 3, 6.352 for Task 4, 6.766 for Task 5, 7.900 for Task 6, and 9.474 for Task 7.

Mixed-Effects Analysis (2): Fixed Part

Using R9 for the random part, Table 15.3 presents the outcomes for the fixed effects (*SPSS*).

In Table 15.3, terms B_0–B_{10} are from the aforementioned regression equation: B_0–B_6 indicate the trajectory in the control condition across tasks cf. Fig. 15.4, whereas B_7–B_{10} indicate the difference between the two conditions for each of the tasks after treatment, with positive differences indicating higher scores in the treatment condition.

Revisiting the Main-Interaction-Simple Effects Distinction

Unless we deal with measurements prior to treatment in a randomised controlled experiment, we should *always* include all main effects underlying an interaction effect. In a two-way or three-way factorial experiment like in Chap. 11 or in a moderated regression like in Chap. 12 and in Experiment 2 in this chapter, we cannot interpret the interaction term as the interaction effect of interest without having the underlying main effects in the model. Whether we deal with measurements prior to or after treatment in a quasi-experiment or with measurements after treatment in a randomised controlled experiment, leaving out the treatment term in attempt to control for baseline differences is generally asking for trouble. However, such control (i.e., correction) *does* make sense when it comes to measurements prior to treatment in a randomised controlled treatment (at least as far as quantitative outcome variables are concerned), whether we have one such baseline measurement (Experiment 2) or several (Experiment 4).

16 Models for Treatment Order Effects

Abstract

In the experiments discussed thus far, treatment is a between-subjects factor: participants in one condition do not also participate in another condition. However, in some cases, different groups may receive different treatments at different occasions, the order of treatment varies across groups, and there is a measurement of an outcome variable of interest at each occasion. In the simplest setup, there are two treatments, A and B, which are taken in a different order by each of two groups: A-B in one group (with a measurement after A followed by a measurement after B), B-A in the other group (with a measurement after B followed by a measurement after A). In other cases, there are more treatments and more orders for a larger number of groups with it, or there are only a few—and perhaps only two—treatments which can vary in each of a larger number of trials. At each trial, there is a measurement of an outcome variable of interest. Consider students who are asked to read ten articles, each article is read in each of two possible formats determined in a random order, and after each article students are asked to rate on a VAS how much effort it took to read the article. These are all examples of situations where treatment varies both between and within participants and a measurement of an outcome variable of interest takes place in each trial (e.g., after each condition or for each article). As in Chaps. 14 and 15, the occasions or trials can still be viewed as stations to be passed by each of the participants, but there is something that varies from station to station that has to be accounted for in our models. This final chapter provides examples for how to do that.

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_16

Introduction

There are not so few examples of empirical research articles that state 'effectiveness' of a treatment based on a study in which only *one group* was measured on an outcome variable of interest prior to and after treatment. Over the years, I have found myself in many discussions with authors and co-authors about interpretation and wording in such cases. We appear to be wired such that we tend to speak in terms of *X* 'impacting' or 'influencing' *Y*, or *X* 'resulting in' higher or lower *Y* even when the design of our study does not allow for such inference. Treating such a *within-subjects* comparison as if it was a true, *between-subjects*, experiment is based on the assumption that the between-subjects model holds for within-subjects studies. This is a very strong assumption; we do not really know if this can ever be true and we actually have good reasons to believe, at least in most settings, that it *cannot* be true.

In a pre-test post-test single-group design, many phenomena may explain a difference that we would like to attribute to a treatment effect. Some of these phenomena may relate to maturation or to other influences unrelated to the treatment of interest, such as regression to the mean (see also Chap. 15). Besides, earlier measurements may in some cases influence later measurements. For instance, in studies on learning, researchers have to realise that a pre-test of knowledge or skill prior to treatment may in itself function as a treatment and influence participants' scores on a post-test of that knowledge or skill (see also the assessment and learning paradox in Chap. 12). Adding pre-test as an additional factor to a randomised controlled experiment may provide a way to estimate the effect of such a pre-test. If we originally had in mind to compare a control condition and two treatment conditions, with adding a pre-test factor, we would obtain a three-by-two design, with half of the conditions obtaining a pre-test and half of the conditions not obtaining a pre-test. Finally, even when there is no pre-test to worry about, participants undergoing different conditions comes at the risk of so-called *carryover effects*: participants' performance or behaviour of interest otherwise in one condition is influenced by the fact that they just participated in at least one other condition (e.g., Gravetter & Forzano, 2006). Examples of many possible sources of carryover effects are learning, fatigue, adaptation, sensitisation, and habituation. All these within-subjects design-related effects have in common that they threaten the internal validity of a study and make outcomes of interest more difficult to interpret.

In a between-subjects design, when the experimental condition and the treatment condition result from random (sampling and) allocation, absent confounding resulting from flaws in the experimental setup, the between-subjects causal account of the treatment-control contrast can be defended via the conditions of Mill (1843): there is *covariation* between treatment factor and outcome variable of interest, differences in treatment precede the measurement of the outcome variable (i.e., *temporal order*), and there are no other factors that could reasonably explain the covariation (i.e., *no alternative explanations*). In this case, two logical methods based on Mill's work together constitute the basis of the treatment-control causal inference: the method of agreement (i.e., treatment condition) or "*if X, then Y*", and the method of difference (i.e., control condition, absent treatment) or "*if not X, then not Y*" (e.g., Rosnow & Rosenthal, 2005). This is consistent with most ideas about causality as well as with ideas of repeated sampling, sampling distributions, and expected values.

An important advantage of within-subjects over between-subjects designs is that they enable us to separate within-subjects from between-subjects variance. This is the reason why treating within-subjects data as if they were between-subjects will result in substantially larger *SE*s and CIs compared to appropriate treating within-subjects data as within-subjects. However, when we apply Mill's methods to a within-subjects design, we run into trouble. Firstly, although there may still be covariation between treatment factor and outcome variable of interest, the models we use to estimate treatment effects usually do not deal with covariation at the level of the individual participant; rather, we have a model based on a group of participants and assume that this model is the same for individual participants (i.e., *local homogeneity*). Secondly, the idea of temporal order from the between-subjects design no longer applies: at least one of the conditions is preceded by a measurement of the outcome variable of interest. Thirdly, a modification of this temporal order and other factors may largely if not completely explain the covariation.

We may never fully get rid of these problems, but may somewhat reduce the trouble by including several groups in our experiment and having each group undergo different conditions in different orders. Given *k* conditions, we have *k!* possible orders of conditions. In the simplest case, where we have two possible conditions A and B, we can have each of two groups undergo these conditions in a different order: A-B (first group) and B-A (second group). With three possible conditions, we have 3 * 2 * 1 = 6 possible orders: A-B-C, A-C-B, B-A-C, B-C-A, C-A-B, and C-B-A. The process of having different groups for different orders of conditions is also called *counterbalancing* (e.g., Gravetter & Forzano, 2006; Rosnow & Rosenthal, 2005). When we have one group for each possible order, we speak of *complete* counterbalancing; when we have fewer groups than possible orders, we can only use a limited number of the orders, and in such cases, we speak of *partial* counterbalancing. A common misunderstanding appears to be that counterbalancing eliminates carryover effects, but this is not the case, and generally speaking, when we have reasons to assume (substantial) carryover from one condition to another we may better opt for a between-subjects design (i.e., have one group for each condition, instead of one group per different order of conditions).

There is an excellent read on counterbalancing and so-called *Latin-square designs* by Richardson (2018). A Latin square is a matrix with *k* number of rows and columns (e.g., two by two, three by three, or six by six), in which sequences of conditions are presented such that each condition occurs only once in each row and only once in each column. Fisher (1925) already recommended the use of Latin squares to control for effects of extraneous variables. In line with one of the core messages of this book, which is based on Fisher (1925), when using Latin squares, the choice of the exact square should match the research design and should be accounted for in the data analysis as well. In this context, Grant (1948) pointed out that a simple two-by-two Latin square, where two conditions are presented in one order to half of the participants but in the reverse order to the other half of the participants, does not control for interactions between the treatment variable and the counterbalanced variable. Later on, others indicated that this problem also exists for larger Latin squares (e.g., Poulton & Edwards, 1979; Richardson, 2018). All these considerations have in common that although Latin-square designs constitute a potentially powerful tool for experimental researchers in a wide range of settings, it is important to account for their use in the statistical analysis (i.e., the bridge between design and analysis). In this chapter, we have a look at two experiments where the order of conditions varies: one with a simple two-order two-group comparison, and one with a multi-order design.

Experiment 1: Two Orders

Sometimes, experiments that include a within-subjects component arise from a particular interest in an *order* effect. Take the following example. Some clinicians are interested in comparing two different orders of tasks in simulation training in terms of clinical reasoning activity during each of these tasks. They randomly sample $N = 100$ Clinical Psychology students and randomly assign them to two possible orders: first Task 1 then Task 2 ($n = 50$), or first Task 2 then Task 1 ($n = 50$). Both tasks take 10 min and involve a participant to carry out physical examination manoeuvres on a human simulated patient while thinking aloud. All sessions are video-recorded and given a consensus-based clinical reasoning score by a team of experienced clinicians who are blind with regard to any possible expectations of which type of task or which order of tasks might yield better scores. Suppose, the histograms of the distributions of scores for each of the task and for the difference between the two tasks do not indicate substantial deviations from normality and the *M*s and *SD*s are as follows. In the condition in which the order is Task 1–Task 2, $M = 10.580$ and $SD = 1.592$ for Task 1, and $M = 11.280$ and $SD = 2.110$ for Task 2. In the condition in which the order is Task 2–Task 1, $M = 11.660$ and $SD = 3.041$ for Task 2, and $M = 11.480$ and $SD = 2.367$ for Task 1. Figure 16.1 presents the scatterplot of the relation between clinical reasoning score at Occasion 1 (Y_1: Task 1 for the order Task 1–Task 2; Task 2 for the order Task 2–Task 1) and Occasion 2 (Y_2: Task 2 for the order Task 1–Task 2; Task 1 for the order Task 2–Task 1) per condition ($X_1 = 0$: order is Task 1–Task 2; $X_1 = 1$: order is Task 2–Task 1).

In the condition where Task 1 is done first, Pearson's correlation between the two tasks is $r = 0.726$, and in the condition where Task 2 is done first, we find $r = 0.794$.

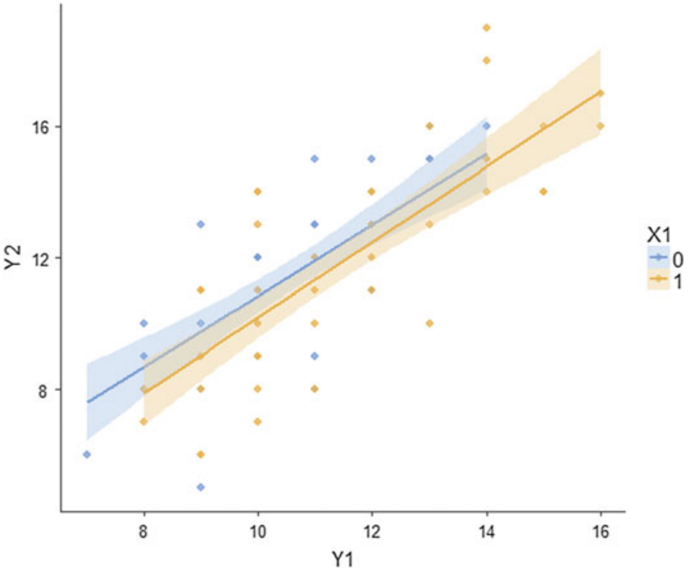


Fig. 16.1 Scatterplot of the relation between clinical reasoning score at Occasion 1 (Y_1: Task 1 for the order Task 1–Task 2; Task 2 for the order Task 2–Task 1) and Occasion 2 (Y_2: Task 2 for the order Task 1–Task 2; Task 1 for the order Task 2–Task 1) per condition (X_1 = 0: order is Task 1–Task 2; X_1 = 1: order is Task 2–Task 1)

Mixed-Effects Analysis (1): Equal Versus Unequal Residual Variance

We can now turn to determining the random part of our mixed-effects model, knowing that the fixed effects to include are the following: main effect of occasion (i.e., first vs. second occasion), main effect of task (i.e., first vs. second task), and the occasion-by-task interaction effect. For the simplest model, CS, we find −2RLL = 833.665 (*SPSS*). For the extension allowing for unequal V_{RES}, we find −2RLL = 803.165. The resulting LR test yields: $\chi_1^2 = 30.500, p < 0.001$. In CS, we find: $ICC = 0.717$ and $V_{\text{RES}} = 5.459$. In the extension with unequal V_{RES}, we find: $ICC = 0.769$, $V_{\text{RES}} = 3.492$ at Occasion 1, and $V_{\text{RES}} = 7.246$ at Occasion 2.

Mixed-Effects Analysis (2): Task, Occasion, and Task Order Effects

In this model, the occasion-by-task interaction effect can be used as an estimate of the task order effect: if the order of tasks does not really matter, the interaction effect should be minimal. Next, the occasion main effect can be used as an estimate of change from one practice occasion to another, and the task main effect can be used as an estimate of the difference between two tasks. In some studies, the main interest lies in the treatment main effect, which would here be the task main effect. In other studies, including the one in this example, the interest lies in the task order effect. However, even if the interest in a particular study lies in the treatment main effect, we will still need to check for interaction first because main effects are generally hard to interpret when there is substantial interaction. For the full factorial model, we find: AIC = 813.549, BIC = 836.637, and for the interaction effect, $p = 0.307$. For the two main-effects model, we find: AIC = 812.597 and BIC = 832.387. For the task main-effect model, we find: AIC = 815.995 and BIC = 832.486. For the occasion main-effect model, we find: AIC = 823.267 and BIC = 839.759. Finally, for the null model, we find: AIC = 826.441 and BIC = 839.634. In other words, the two main-effects model appears to be preferred. In this model, we find for occasion (i.e., higher at Occasion 2): $B = 0.640$, $p < 0.001$, and 95% CI = [0.295; 0.985]; and we find for task (i.e., higher for Task 2): $B = 0.352$, $p = 0.020$, and 95% CI = [0.057; 0.648].

Experiment 2: Several Orders

Experiment 1 can be easily extended to more orders when more than two tasks are to be compared. However, sometimes, researchers take a different approach to dealing with task order effects, even when only two types of tasks are involved. For example, in a recent study by Martin et al. (2018), 72 participants read eight articles in either of two formats—infographic or text-only—in either of eight possible orders. Several outcome variables were measured, one of which was viewing time. The 72 participants were randomly assigned to each of the orders (i.e., $n = 9$ per order). Had the group of participants been larger (e.g., a larger-scale online experiment), an approach to randomisation could have been as follows. Given k tasks and c options per task, the number of possible orders N_{O} is:

$$N_{\mathrm{O}} = c^k.$$

With $c = 2$ and $k = 8$ (cf. Martin et al., 2018), this comes down to 256 different orders. If we are not interested in estimating the effects of different orders (which are difficult to estimate with 72 participants and 8 orders as well) and we have $N = 256$ participants to be randomly assigned to these orders, we could have one participant per order. Advantages of this is that the occurrence of each of the formats is exactly equal (i.e., 50% of participants receiving infographic, 50% of participants receiving text-only) at each occasion and that the correlation between occasions in terms of format occurrence is exactly 0; these factors facilitate estimation. Another factor that facilitates estimation is sample size. With only 72 participants, several of the residual covariance structure models that use relatively many *df*—UN, Toeplitz with unequal V_{RES}, and AD1, and RI-RS models that include several RS terms and correlations between RI and RS terms—may not provide stable estimates, and we may need to use more restrictive models such as

Table 16.1 *Ms* and *SDs* of viewing time per format per occasion (*Jamovi*)

Occasion	Text-only *M (SD)*	Infographic *M (SD)*
1	7.955 (0.457)	8.400 (0.513)
2	7.695 (0.597)	7.924 (0.543)
3	7.573 (0.549)	7.802 (0.574)
4	7.417 (0.538)	7.689 (0.544)
5	7.478 (0.578)	7.573 (0.610)
6	7.423 (0.560)	7.459 (0.569)
7	7.349 (0.524)	7.480 (0.591)
8	7.389 (0.628)	7.359 (0.563)

CS or AR1 with or without unequal V_{RES} even if these more restrictive models in somewhat larger samples usually perform worse than their more flexible alternatives. Therefore, in this final experiment of this book, we have a look at a simulated example of an experiment like the one by Martin et al. (2018), in which viewing time (in minutes) serves as the outcome variable, but where the sample size is larger.

Suppose, the usual histograms indicate no strange cases or shapes, and Table 16.1 presents the *M*s and *SD*s of viewing time per format per occasion.

The findings in Table 16.1 appear to indicate a format-by-occasion interaction effect: the difference between formats in average viewing time appears to decrease from occasion to occasion.

Mixed-Effects Analysis (1): Random

The fixed effects to be included are: main effect of format (i.e., infographic vs. text-only), main effect of occasion (i.e., Occasions 1–8), and the format-by-occasion interaction effect. Table 16.2 provides a comparison between different residual covariance structures, using the same fixed part (i.e., format and occasion main effects and their interaction effect): (R1) UN; (R2) CS; (R3) CS with unequal V_{RES} across tasks; (R4) RI, and RS for the difference between first and last task; (R5) Toeplitz; (R6) Toeplitz with unequal V_{RES} across tasks; (R7) AR1; (R8) RI and AR1; (R9) AR1 with unequal V_{RES} across tasks; (R10) RI and AR1 with unequal V_{RES} across tasks; (R11) AD1; and (R12) HF.

For R6–R11, the LR test is not statistically significant at the 5% level, although for R7 and R8 the outcome *is* statistically significant if we test one-sided (see also Chap. 13). For the comparisons of random part models in the following, the outcomes are or are not statistically significant at 5% regardless of whether we test one-sided or two-sided, so let us stick with two-sided testing. The most restrictive of R6–R11 is R7, and this is a special case of each of R6 and R8–R11. For R7 versus R8 (i.e., the nearest to R7 in terms of additional *df*), the LR test yields: $\chi_1^2 = 0.207$, $p = 0.649$. For R7 versus R9 (i.e., the second-nearest to R7 in terms of

Table 16.2 Comparison of eight residual covariance structures in Experiment 2 (*SPSS*): (R1) UN; (R2) CS; (R3) CS with unequal V_{RES} across tasks; (R4) RI, and RS for the difference between first and last task; (R5) Toeplitz; (R6) Toeplitz with unequal V_{RES} across tasks; (R7) AR1; (R8) RI and AR1; (R9) AR1 with unequal V_{RES} across tasks; (R10) RI and AR1 with unequal V_{RES} across tasks; (R11) AD1; and (R12) HF: number of estimated parameters (*df*), −2RLL, and LR test χ^2_{df} and *p*-value for any of R2–R12 tested against R1

Structure	*df*	−2RLL	LR test χ^2_{df}	LR test *p*-value
R1	36	3018.991	–	–
R2	2	3324.434	$\chi^2_{34} = 305.443$	<0.001
R3	9	3310.937	$\chi^2_{27} = 291.946$	<0.001
R4	3	3319.352	$\chi^2_{33} = 300.361$	<0.001
R5	8	3062.013	$\chi^2_{28} = 43.022$	0.035
R6	15	3046.163	$\chi^2_{21} = 27.172$	0.165
R7	2	3064.081	$\chi^2_{34} = 45.090$	0.097
R8	3	3063.874	$\chi^2_{33} = 44.883$	0.081
R9	9	3048.765	$\chi^2_{27} = 29.774$	0.324
R10	10	3048.517	$\chi^2_{26} = 29.526$	0.288
R11	15	3045.703	$\chi^2_{21} = 26.712$	0.181
R12	9	3307.284	$\chi^2_{27} = 288.293$	<0.001

additional *df*), the LR test yields: $\chi^2_7 = 15.316$, $p = 0.032$. For R7 versus R10 (i.e., third-nearest to R7 in terms of additional *df*), the LR test yields: $\chi^2_7 = 15.564$, $p = 0.049$. For R7 versus R11, the LR test yields: $\chi^2_{13} = 18.378$, $p = 0.019$. Finally, for R7 versus R6, the LR test yields: $\chi^2_{13} = 17.918$, $p = 0.161$.

In both R7 versus R8 and R9 versus R10, the difference is a RI term, and the outcome of the LR test in R9 versus R10 is similar to that of R7 versus R8: $\chi^2_1 = 0.248$, $p = 0.618$. For R9 versus R11, we find: $\chi^2_6 = 3.062$, $p = 0.801$. Finally, for R9 versus R6, we find: $\chi^2_6 = 2.354$, $p = 0.884$. We cannot perform a LR test on R6 versus R11, since neither is a special case of the other and the difference in $df = 0$ for these two models. However, there appears to be no need to compare R6 and R11 anyway, for we have insufficient evidence to assume that R9 is too much of a simplification of R6 and R11.

In R9, r_{RES} for adjacent tasks is estimated to be 0.462, and V_{RES} varies from 0.239 for Article 1 to 0.361 for Article 8.

Mixed-Effects Analysis (2): Fixed

Using R9 for the random part, we find the following outcomes for the fixed effects. For the full factorial model, we find: AIC = 3027.051, BIC = 3167.666, and $p < 0.001$ for the interaction effect. For the two main-effects model, we find: AIC = 3040.553 and BIC = 3141.796. For the article-effect-only model, we find:

Table 16.3 *EMM* and *SE* per format per article in Experiment 2 (*SPSS*)

Article	Text-only *EMM* (*SE*)	Infographic *EMM* (*SE*)
1	7.996 (0.041)	8.359 (0.041)
2	7.713 (0.046)	7.906 (0.046)
3	7.564 (0.045)	7.812 (0.045)
4	7.436 (0.043)	7.670 (0.043)
5	7.446 (0.047)	7.606 (0.047)
6	7.420 (0.045)	7.462 (0.045)
7	7.352 (0.045)	7.477 (0.045)
8	7.372 (0.050)	7.376 (0.050)

AIC = 3112.117 and BIC = 3207.795. For the format-effect-only model, we find: AIC = 3300.732 and BIC = 3362.603. Finally, for the null model, we find: AIC = 3356.145 and BIC = 3412.391. Although BIC prefers the two main-effects model, AIC and the *p*-value associated with the interaction effect indicate a preference of the full factorial model. Table 16.3 presents the *EMM*s and *SE*s for each format per article.

The additional viewing time needed for infographic decreases with the number of articles.

To Conclude

In Experiment 2, as in several experiments in previous chapters, we see that BIC sometimes prefers a simpler model—for instance a two main-effects instead of a full factorial model—where AIC and perhaps *p* and/or JZS-prior-based BF appear to hint at a more complex model (e.g., full factorial instead of two main effects), especially when the additional complexity requires quite a few more *df* (i.e., considerably more parameters to be estimated). For the effects in Experiment 2, for example, the main effect of format requires 1 *df* while the main effect of article and the format-by-article interaction effect each require 7 *df*. Conversely, AIC may in some cases slightly prefer a more complex model where none of the other criteria may support that. This underlines the importance of not relying on a single criterion for model selection but to base model-selection decisions on combinations of criteria. As seen in previous chapters, when both AIC and BIC indicate a preference towards a more complex model relative to simpler alternatives, the *p*-value from a 'no difference' null hypothesis significance test will likely be statistically significant at $\alpha = 0.05$ (two-tailed testing) and JZS-prior-based BF will normally be in favour of the more complex model. When neither AIC nor BIC indicate a preference towards a more complex model, the *p*-value from a 'no difference' null hypothesis significance test will normally not be statistically significant at $\alpha = 0.05$ and

JZS-prior-based BF will probably indicate at least some preference towards (one of) the simpler model(s). This is *not* to say that the simpler model is then ‘true’, but we have insufficient evidence to assume that the more complex model provides a better explanation of the outcome variable of interest.

While effect size estimates such as Cohen’s *d* and associated CIs and overall model fit indices such as R^2 are easy to compute in fixed-effects designs and are also fairly easy to obtain in group-nesting cases such as discussed in Chap. 13 and in experiments with two or more assessors where CS holds such as in Experiment 1 in Chap. 14, things can become tricky when more complex covariance structures are to be preferred. When the random part consists of only an RI term (cf. CS), V_{Random} is the RI variance estimated. However, when for instance an RI term and an RS term are included in the random part, V_{Random} is the sum of the RI variance, the RS variance, and the RI-RS covariance. This is still fairly straightforward, but when we deal with multiple occasions and add several RS terms or use any of AR1 (with or without varying V_{RES}), Toeplitz (with or without varying V_{RES}), HF or AD1 to model the random part, things become more complicated.

Some have proposed pseudo-R^2-statistics similar to R^2_{McF}, R^2_{CS}, and R^2_{N}, based on differences in LL or differences in −2LL. Given the relation between −2LL differences and LR tests, some refer to this kind of pseudo-R^2-statistics as LR-based pseudo-R^2-statistics (e.g., Bartoń, 2018). While under FIML, such statistics will provide differences in outcomes between different fixed-effects solutions (e.g., one vs. two main effects, or two main effects vs. two main effects plus interaction effect) in mixed-effects modelling that are consistent with differences between these solutions in fixed-effects modelling (i.e., increasing not decreasing R^2 when adding one or more fixed-effects terms), these statistics and differences are generally lower —and sometimes quite a bit lower—than what we obtain through R^2 in fixed-effects models or R^2_{M} in mixed-effects models. Consequently, we cannot really interpret a difference in LR-based pseudo-R^2-statistic for two competing models—such as two main effects versus two main effects plus interaction effect—in terms of a difference in proportion of variance explained in the outcome variable as we do with the conventional R^2 in fixed-effects models or with R^2_{M} in mixed-effects models. However, given its intuitive concept, applicability across fixed-effects and mixed-effects models, and consistency with R^2 and R^2_{M}, we *can* use the difference in deviance (−2LL) of any model versus Model 0 to acquire a *deviance reduction factor* (DRF)—which in fixed-effects categorical outcome variable models is also known as R^2_{McF}—that quantifies the *proportion of reduction in deviance* by an alternative model (Model 1) at hand relative to Model 0, given the same random part for the two models (i.e., the random part is modelled, using REML, before the fixed part is taken care of, using FIML):

$$\text{DRF}_{10} = [-2\text{LL Model } 0 - -2\text{LLModel } 1]/-2\text{LLModel } 0.$$

This factor can also be used for comparing other models, one of which is a special case of the other, for instance: a two main-effects model (special case: Model 3) versus a two main-effects plus interaction effect model (Model 4); DRF_{43} is then the proportion of reduction in deviance by Model 4 relative to Model 3:

$$\text{DRF}_{43} = [-2\text{LLModel } 3 - -2\text{LLModel } 4]/-2\text{LLModel } 3.$$

When a fixed effect adds absolutely nothing, for instance: the *M*s of treatment and control condition in an experiment are *exactly* the same, $\text{DRF}_{10} = 0$. In the other extreme (and highly unlikely) case, where *all* of the variance in an outcome variable can be explained by a particular model, the deviance is reduced from non-zero to zero, hence $\text{DRF}_{10} = 1$ (see for instance the examples of perfect fit in Chap. 5, where $R^2_{\text{McF}} = 1$). This factor can provide a consistent approach to model comparison—along with criteria such as *p*, AIC, BIC, and BF, and confidence intervals and optionally credible intervals (and eventually R^2, adjusted R^2, R^2_{M} or η^2 and ω^2 when dealing with quantitative outcome variables)—for categorical and quantitative outcome variable situations discussed in this book.

That said, regardless of the simplicity or complexity of computations of (pseudo-) R^2-statistics, even when effect size estimates can be computed easily (i.e., most fixed-effects and CS mixed-effects models), they need to be evaluated in the context at hand and in the light of differences on the original scale (e.g., points or minutes). A Cohen's *d* of 0.50 may make sense when on an outcome variable of interest that can range from 0 to 5 the control group has $M = 3.0$ and $SD = 1.0$ and the treatment group has $M = 3.5$ and $SD = 1.0$. However, when on that same outcome variable (0–5) $M = 4.50$ and $SD = 0.20$ in the control group and $M = 4.60$ and $SD = 0.20$ in the treatment group, Cohen's *d* is still 0.50 but the actual difference between groups may not have much if any practical meaning, at least for many outcome variables. TOST, ROPE, and FOST can be applied to both standardised (e.g., Cohen's *d*) and unstandardised (e.g., points or minutes) differences, provided that researchers can reasonably agree on the equivalence bounds. The 90% CIs are not provided in the experiments in Chaps. 15 and 16, because they can be computed from the 95% CIs or *SE*s and the focus in these last two chapters lies on concepts not discussed in earlier chapters.

This final chapter also once more stresses that, whether our primary interest lies in a main effect or in an interaction effect, we always need to check for interaction before we draw any conclusions about main effects. It is well possible, for example, that researchers who do an experiment similar to Experiment 2 in this chapter are first of all interested in a main effect of format, and may even have an expectation with regard to what that effect looks like. However, the outcome of the interaction effect and the pattern of *EMM*s and *SE*s in Table 16.3 indicates that the two conditions under comparison—infographic and text-only—appear to slowly converge to no difference as we read more articles. We have no idea what would happen if participants read more than eight articles, but at the eighth article in the series of eight, we observe a difference in viewing time of about 0.004 min, which is less than a quarter of a second. Averaged across all articles, we find $EMM = 7.533$

(SE = 0.022) for text-only and EMM = 7.712 (SE = 0.022) for infographic—about 0.179 min or about 10.74 s, $p < 0.001$—but we may wonder if 10.74 s on an average of a bit over 7.5 min is a big deal, and this difference is representative for *neither* the first articles (larger differences, for instance 0.363 min or 21.78 s for the first article) *nor* later articles (slowly converging to zero).

Part V
General Recommendations

A General Pragmatic Approach to Statistical Testing and Estimation

17

Abstract

This final chapter provides a synthesis of the other sixteen chapters in this book in the form of a general pragmatic approach to statistical testing and estimation: a coherent set of general recommendations and guidelines on core questions discussed throughout this book: design and sample size (single-level vs. two- or multilevel designs); gaining power (contrasts and sequential testing); missing data (MAR or MCAR vs. MNAR); psychometrics of measurement instruments (levels of measurement and easy alternatives to Cronbach's alpha, among others via a general mixed-effects modelling approach); testing and estimating treatment effects (several criteria and concepts); and dealing with covariates (general guidelines and baseline measurements as a special case). As such, this chapter provides a concise overview and summary of the other sixteen chapters in this book.

Introduction

There ought to be a logical connection between questions, design, and analysis (cf. QDA introduced in Chap. 1). More complex questions generally call for more complex designs and analyses, and that puts higher demands on the sample size. Chapter 2 provides a variety of sources to assist researchers to perform required sample size calculations, assuming a particular effect size, statistical significance level, and desired statistical power. Whether we use *p*-values, information criteria, or BFs, evidence against a $\boldsymbol{H_0}$ in favour of an alternative $\boldsymbol{H_1}$ is—keeping other factors constant—more easily established in larger than in smaller samples. Although we cannot use *p*-values to obtain evidence in favour of a $\boldsymbol{H_0}$ (after all, it is a conditional probability in which the condition is that $\boldsymbol{H_0}$ is true), whether we use information criteria or BFs, evidence in favour of a $\boldsymbol{H_0}$ relative to an alternative $\boldsymbol{H_1}$

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4_17

is—keeping other factors constant—more easily established in larger than in smaller samples. Whether we adopt a Frequentist approach or a Bayesian approach to interval estimation, CIs and CRIs are—keeping other factors constant—narrower in larger than in smaller samples. Whether we call the approach we use TOST, ROPE or FOST, evidence against or in favour of relative or practical equivalence is more easily established in larger than in smaller samples.

Design and Sample Size

The more groups we wish to compare in our experiments, the larger our total sample should be. When we are interested in main and interaction effects of two or more factors, sample size demands also go up.

Single-Level Designs

Another factor that influences sample size demands is that of the type of outcome variable considered. Generally, categorical outcome variables (Chaps. 5, 6 and 7) are more demanding than quantitative outcome variables (Chap. 8), and the demand increases with the number of categories of a categorical outcome variable. Zero cells and low frequency (e.g., 1 or 2 observations) cells can pose serious threats to the validity of testing and estimation outcomes; some statistics may not be computed, and other statistics may be highly inaccurate. It is important to be aware of this before instead of only after data collection.

Two- or Multilevel Designs

Apart from the type of outcome variable, the number of levels and nature of these levels in a design also influence the required sample size. Chapters 1 and 13 provide a useful formula for estimating by what factor the sample size required in a single-level design should be multiplied to acquire a sample size required in a two-level design, where the levels are cluster (level 2) and participant within cluster (level 1). Some of the sources referred to in Chap. 2 also enable researchers to do required sample size calculations for other types of two- or multilevel designs, including where the multilevel structure—or part of it—is due to participants being measured repeatedly or by multiple assessors on the same outcome variable of interest. Just like with cluster designs we should not consider k clusters times n participants the total sample size, we should not treat N participants times k measurements as a total sample size of $N * k$ either. A feature inherent to multilevel designs (Chaps. 13, 14, 15 and 16) is that the effective sample size is always smaller than the products just mentioned, and how much smaller depends on the *ICC*.

Gaining Power

There are ways to gain statistical power relative to conventional practices: planned contrasts through directed hypotheses and/or factorial designs, and sequential testing. Although factorial designs are used quite commonly, they are not always treated appropriately by researchers in the analytic stage. Directed hypotheses and sequential testing remain underused, possibly partly due to misconceptions around these concepts, but that may change in the near future.

Contrasts

Although in the overwhelming majority of studies, researchers engage in two-sided testing and apply Bonferroni or some other correction for multiple testing when several tests are carried out (Chap. 9), in not so few cases neither of these two conventional practices may be appropriate. When directed hypotheses are available and explicitly formulated in grant proposals and (pre)registered reports, one-sided testing—with two groups, or with several groups when for instance a Helmert-type or ordinal hypothesis is available—and refraining from multiple testing (Chap. 10) are not only defendable but a moral and ethical obligation as well. After all, if we can achieve a higher statistical power with a given number of participants or we can achieve the same statistical power with a lower number of participants, that is certainly a good thing. We should not use more participants and resources than needed.

Another practice that results in an unnecessary loss of statistical power and precision is found in treating two- or three-way design data as one-way. Unfortunately, there are full professors with whom I have had several discussions on this matter and to whom I explained—with numbers and main/interaction effect distinctions—why the one-way practice is incorrect, but who nevertheless prefer to continue treating their two- or three-way design data as one-way and hence voluntarily choose to obtain incorrect outcomes and misinform educational practice. Chapter 11 demonstrates how dramatic the loss of statistical power and precision can be. We are all human and all make mistakes, but to prolong a practice that we can reasonably know to be wrong is unethical and irresponsible.

Sequential Testing

Another way to gain statistical power, which has commonly been associated with Bayesian statistics but—when applied appropriately—can be justified in a Frequentist approach as well, is sequential testing (Chaps. 2 and 10). Contrary to popular belief, as discussed in Chaps. 2 and 10, sequential testing has been around in Frequentist statistics for decades, and sequential analyses can be perfectly valid from a Frequentist perspective, provided that they are carefully planned a priori and statistical significance levels are adjusted appropriately. Given its potential in terms of power gain and required sample size reduction (i.e., using as few participants and resources as possible), it is high time we start implementing the concept of sequential analysis in educational and psychological experimental research.

Missing Data

As discussed in Chap. 4, several methods for dealing with missing data are encountered in the literature, some of which are better than others. How to deal with missing data depends on a variety of factors outlined in Chap. 4, but the following guidelines are overall safe.

Random or Completely Random

Under MCAR and small percentages of missing (i.e., less than 10%), listwise deletion may be applied. This may be useful especially when missingness occurs on one of the predictor variables, as that type of missing poses a threat to the validity of model comparisons. Consider a two-way factorial experiment with one covariate, and some missing on the covariate. In models that do not include the covariate, all data is used; in models that do include the covariate, all data minus the cases who have the covariate missing is used. This is problematic, because it affects SEs and therefore CIs and *p*-values, and invalidates comparisons of models with versus without covariate in terms of AIC and BIC and other criteria (e.g., BF and SABIC), because the latter requires that *exactly* the same data is used in *all* models compared. However, if missing is not MCAR, listwise deletion comes with biased estimates. As long as the sample size is—from the start or due to the amount of missingness—not too small, FIML and MI may both provide useful alternatives under MCAR and under MAR, with FIML possibly being somewhat less biased than MI in the case of somewhat smaller samples.

Not Random

Under MNAR, MI appears the best approach to dealing with missing data, although when sample sizes are—from the start or due to the amount of missingness—small MI may be problematic as well. As explained in Chap. 4, the problem with MAR and MNAR is that they cannot be empirically tested. Although MCAR is often easy to recognise and can be tested empirically, whether missingness is MAR or MNAR is often difficult to tell. From a pragmatic standpoint, one might—in carefully designed and carried out randomised controlled experiments—want to treat missingness as MAR if MCAR is not the case and use FIML. An advantage of FIML over MI is that no imputation is carried out; the missingness is handled in the analysis. When missingness is MNAR, there are challenges to be faced anyhow; unless samples are sufficiently large, good auxiliary variables are available, and we are willing to make perhaps rather strong assumptions, even MI may not really help us deal with our worries.

Psychometrics of Measurement Instruments

Chapters 3, 14, 15 and 16 cover questions concerning the psychometrics of our measurement instruments and approaches. Whether we use items or assessors at a given point in time or we deal with outcome variables measured repeatedly for the same participants, we deal with some important questions. The mixed-effects modelling approach presented in Chaps. 14, 15 and 16 enables researchers to test and estimate treatment effects of interest while accounting for the residual covariance structure without having to rely on latent variables.

Levels of Measurement

Although the examples discussed in Chaps. 14, 15 and 16 use quantitative outcome variables, the concepts and methods discussed in these chapters also apply to categorical outcome variables. This allows researchers to put the Cronbach's alpha as a considered 'default' estimator of reliability or internal consistency in a broader framework of residual covariance structures regardless of the level of measurement of the items, ratings or repeatedly measured outcome variable of interest. Also, traditional computations of Cronbach's alpha and its alternatives require no missing data. On the contrary, a mixed-effects modelling approach can—at least under MAR and MCAR—obtain valid estimates in the presence of missing data as well. Now that we have seen the multilevel designs and residual covariance structures, let us take a final look at this.

Easy Alternatives to Cronbach's Alpha

Suppose, in a random sample of $N = 300$ students, we administer two questions, each of which require a response on a VAS with negative and positive scores and with 0 as middle point (i.e., neutral). The two items correlate $r = 0.469$, each item has $M = 0$, and item 1 has $SD = 1$. For the SD of item 2, let us look at three scenarios: Scenario 1: $SD = 1$; Scenario 2: $SD = 1.5$; Scenario 3: $SD = 2$. Item 2 from Scenario 2 is obtained by multiplying item 2 from Scenario 1 by 1.5, and item 2 from Scenario 3 is obtained by multiplying item 2 from Scenario 1 by 2. Thus, the three versions of item 2 correlate perfectly.

In each of the three scenarios, McDonald's omega is 0.638. However, for Cronbach's alpha, we find: 0.638 in Scenario 1, 0.604 in Scenario 2, and 0.545 in

Scenario 3. The reason for this is that *ICC* based on CS is different in each of the three scenarios: 0.469 in Scenario 1, 0.432 in Scenario 2, and 0.375 in Scenario 3. There is an easy solution to this problem: to estimate *ICC* under a more flexible alternative to CS, namely allowing V_{RES} to vary across items. Doing so, we find $ICC = 0.469$ in all three scenarios, and hence the V_{RES}-adjusted alpha is equal to McDonald's omega not only in Scenario 1 but in all three scenarios: 0.638. This difference between Cronbach's alpha and its two more appropriate alternatives does not matter only for reporting on the reliability of the set of items but matters for questions like 'how many items would we need to obtain a reliability of 0.7, 0.8 or 0.9' as well. Based on the *ICC* obtained under CS, we would need at least 3 items for a reliability of 0.7, at least 5 items for a reliability of 0.8, and at least 11 items for a reliability of 0.9 in Scenario 1. In Scenario 2, the numbers would be: 4 items (0.7), 6 items (0.8), and 12 items (0.9). And in Scenario 3, the numbers would be: 4 items (0.7), 7 items (0.8), and 15 items (0.9). Using *ICC* under the more flexible CS with varying V_{RES}, the numbers would be the same for all three scenarios: 3 (0.7), 5 (0.8), and 11 (0.9). Especially for a reliability of 0.9, this is quite a difference.

Due to the CS assumption, the upper bound of Cronbach's alpha is smaller than 1 when items are perfectly correlated but standard deviations are different. If we take item 2 from Scenario 1 and item 2 from Scenario 2, we obtain a Cronbach's alpha of 0.960, and the same exercise for item 2 from Scenario 1 and item 2 from Scenario 3 yields a Cronbach's alpha of 0.889. When we express reliability in terms of consistency, as Cronbach's alpha and its alternatives are supposed to do, perfect correlation should result in a reliability of 1. However, the upper bound of Cronbach's alpha decreases as the difference in *SD*s of items involved increases.

The same holds when three or more items are involved. When for example three items are involved and both the correlations and *SD*s are more or less the same, CS may be realistic. However, when *SD*s and/or correlations differ, alternatives are better. Table 17.1 provides an example for three different scenarios (again: a random sample of $N = 300$ in each scenario) with three items: (1) CS is realistic, (2) CS with varying V_{RES} is realistic, and (3) UN is best.

When the correlation is fairly similar across pairs of items but the *SD*s are quite different (i.e., Scenario 2 in Table 17.1), V_{RES}-adjusted α and McDonald's ω yield about the same outcome, and when correlations differ as well (cf. Scenario 3 in Table 17.1) McDonald's ω yields the highest

As discussed in Chap. 3, McDonald's ω and other alternatives to Cronbach's α, such as GLB, put some demands on the sample size. If researchers worry about their sample size not being large enough to use McDonald's ω or GLB, an easy alternative to Cronbach's α that accounts for varying V_{RES} and provides estimates more in line with McDonald's ω is found in V_{RES}-adjusted α, especially when the correlation is fairly similar across pairs of items. This is not to say that we can throw McDonald's ω, GLB, and other alternatives to Cronbach's α in the bin, but we do have an alternative to Cronbach's α when there is some worry about whether or not the sample size is large enough for McDonald's ω and other alternatives. When in

Table 17.1 Cronbach's alpha, V_{RES}-adjusted alpha, and Cronbach's alpha in three scenarios (*JASP*, *SPSS*): (1) CS is realistic, (2) CS with varying V_{RES} is realistic, and (3) UN is best

Scenario	(1): CS	(2): varying V_{RES}	(3): UN
ICC based on CS	0.343	0.309	0.343
ICC with varying V_{RES}	0.343	0.328	0.356
Cronbach's α	0.610	0.573	0.610
V_{RES}-adjusted α	0.610	0.595	0.621
McDonald's ω	0.612	0.599	0.665
Average inter-item r	0.342	0.328	0.352
Item 1–Item 2 r	0.315	0.275	0.477
Item 1–Item 3 r	0.379	0.351	0.373
Item 2–Item 3 r	0.333	0.359	0.206
SD Item 1	0.782	0.598	1.259
SD Item 2	0.703	0.711	1.152
SD Item 3	0.758	1.036	1.437

doubt, reporting a combination of criteria (e.g., Cronbach's α and McDonald's ω, Cronbach's alpha and GLB, or Cronbach's alpha and V_{RES}-adjusted α), especially in combination with information on item standard deviations and correlations between pairs of items, is always an option.

Testing and Estimating Treatment Effects

The mixed-effects modelling approach has another advantage: we can simultaneously estimate reliability and treatment effects of interest without potentially overestimating the reliability. If we were to compute Cronbach's α, V_{RES}-adjusted α, or McDonald's ω for the *full* sample—different treatment conditions merged together—we would likely overestimate the reliability. Estimating the reliability for each condition separately would then be more appropriate but may be difficult when sample sizes are on the smaller side. In the mixed-effects modelling approach, one common reliability estimate can be obtained while appropriately accounting for different conditions, and the problem of overestimated reliability is avoided.

For making decisions on the residual covariance structure, an LR testing approach under REML in which simpler structures are tested against UN, and perhaps some additional tests are carried out for structures that can be considered special cases of other structures, constitutes a robust approach. The problem with AIC and BIC when dealing with random effects is that many *df* can be involved. The more *df* are involved, the more AIC and BIC may diverge, because AIC will tend more towards increased complexity while BIC will tend more towards as simple as possible even if the latter makes no sense (e.g., preferring independence of residuals when there is dependence). When dealing with fixed effects, AIC and BIC are normally useful, unless *df* associated with a treatment or interaction effect is rather large (i.e., many conditions or groups); in the latter case, AIC and BIC may in some cases quite easily disagree.

A Deviance-Reduction/Information-Theoretic Approach

When dealing with quantitative outcome variables, R^2, adjusted R^2, η^2, ω^2, and R^2_{M} may provide useful statistics for the proportion of variance explained in an outcome variable depending on what type of design is employed. When dealing with categorical outcome variables, I recommend R^2_{McF} as the default pseudo-R^2-statistic for reasons outlined in Chaps. 5 and 6. The DRF discussed in Chap. 16 provides a natural extension (i.e., deviance-reduction equivalent) of R^2_{McF} for multilevel designs for both categorical and quantitative outcome variables.

Differences in AIC and BIC are a direct function of deviance reduction and a criterion-specific penalty for model complexity (and in the case of BIC: sample size). Although AIC and BIC do not require researchers to specify prior distributions as in Bayesian analysis, one way to view AIC and BIC is as a form of Bayesian analysis using different priors. The prior associated with BIC is a bit wider (i.e., less informative) than that associated with AIC and therefore less easily prefers a more complex model than AIC. BFs based on JZS priors tend to yield an outcome somewhere in between the pseudo-BFs one would obtain based on AIC and BIC. In cases where AIC and BIC agree with regard to which model to prefer, the decision is fairly easy. In cases where AIC and BIC disagree, AIC indicates a preference towards a more complex model than does BIC; other statistics—including R^2-statistics—may then be used to support decision making, although the strength of evidence in favour of a possible 'best' model is in such disagreement cases often limited at best (i.e., unless the difference in *df* between competing models is rather large). As such AIC and BIC provide easy alternatives to *p*-values (which cannot provide evidence in favour of $\boldsymbol{H_0}$ relative to $\boldsymbol{H_1}$) and BFs (which require the specification of prior distributions). When two-sided testing is considered, AIC and BIC may serve as primary criteria and *p*-values and JZS-prior-based BFs may be considered additionally. However, in the case of one-sided testing, one-sided BFs and one-sided *p*-values make more sense than AIC or BIC.

Two Types of Intervals: '$1 - \alpha$' and '$1 - 2\alpha$'

In several chapters in this book, 90 and 95% CIs are used, and in some chapters 95% CRIs are used as well. While 95% CIs constitute the default in educational and psychological experimental research because usually two-sided tests at $\alpha = 0.05$ are performed, 95% CRIs are also starting to be used and can be used for ROPE. For TOST, 90% CIs (i.e., $1 - 2\alpha$) are used. In FOST, both 90 and 95% CIs as well as 95% CRIs can be used. As argued in Chap. 9, 90% CIs can be used as a default, 95% CRIs are—in terms of width—usually somewhere in between the 90 and 95% CIs, and 95% CIs can be reported additionally when some correction for multiple testing is to be considered. As argued in Chap. 9, reporting 90% and optionally 95% CIs (and additionally or alternatively: 95% CRIs) is always a good idea, even in the form of a follow up in multi-condition comparisons (i.e., FOST-OF).

Provided that we can reach agreement on a region of relative or practical equivalence, in a study—or usually, and better, across a series of studies—three outcomes are possible: (a) evidence in favour of relative equivalence (i.e., rejecting both $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$, hence rejecting both substantial negative and substantial positive differences, and thus rejecting the whole non-relative-equivalence range), (b) evidence against relative equivalence (i.e., when either $\boldsymbol{H_{0.3}}$ or $\boldsymbol{H_{0.4}}$ can be rejected), or (c) inconclusive (i.e., when at least one of $\boldsymbol{H_{0.1}}$ and $\boldsymbol{H_{0.2}}$ cannot be rejected, and none of $\boldsymbol{H_{0.3}}$ and $\boldsymbol{H_{0.4}}$ can be rejected). In single studies, when samples are of a smaller size, (c) is usually much more likely than (a) or (b); the larger our samples, the easier (a) and (b) can be established, although we should always bear in mind that statistics is normally about relative not absolute evidence and that there is always a real chance that our conclusions are wrong. Whenever possible, it is better not to draw conclusions based on single studies but to invest in replication instead, and even then, we should not interpret findings as absolute evidence.

Dealing with Covariates

Caution is a virtue when practicing statistics, and one of the topics where that applies especially is when dealing with covariates. As discussed in Chaps. 12 and 15, third variables can have a variety of different roles depending on the study design and context, and their roles influence how we should treat them.

General Guidelines

A first general guideline for the treatment of a third variable or ‘covariate’ is that when it is measured after the start of treatment it may well be affected by that treatment and is therefore to be treated as a mediator not as a confounder. Treating such a variable as a covariate as if it was a confounder can then result in a substantial or even severe distortion of testing and estimation outcomes regarding a treatment effect of interest; treating it as a mediator is then more appropriate. Besides, regardless of whether a third variable is or is not a mediator, it may moderate a treatment effect of interest, meaning that the magnitude of a treatment effect may differ substantially across the range of the third variable. Failing to account for that moderation may result in meaningless outcomes with regard to the treatment effect of interest. Finally, checking for treatment-by-third-variable moderation always requires both the main effect of treatment and the main effect of the third variable to be included in our model, unless the third variable is a baseline measurement: a first measurement of a response variable of interest that takes place prior to the start of treatment (e.g., pre-test post-test control-group design).

Baseline Measurements: A Special Case

Experiments 2 and 4 in Chap. 15 demonstrate how to deal with baseline measurements. While the approach taken there would likely result in biased estimates in the case of non-randomised groups, in the case of randomised controlled experiments it constitutes a best practice. The rationale behind this differential treatment for non-randomised versus randomised studies is regression to group-specific *M*s in non-randomised versus regression to a *common M* in randomised studies. Therefore, although the general guideline is to have both treatment and covariate main effects in a model that also includes the interaction term, in the special case of baseline measurements it is recommended *not* to include the treatment term but only the covariate term and the treatment-by-covariate interaction term. The latter treatment correctly treats the samples in conditions prior to treatment as coming from one common source population and hence one common *M* to be regressed to.

To Conclude

Like Education, Psychology, Medicine, and other disciplines, the discipline of Statistics is in continuous development; new work on the behaviour of existing statistical methods under different circumstances, on new statistical methods, and other topics, is published almost on a daily basis. It is a real challenge keep up with all the developments, and whichever discipline we find ourselves in, we can learn new things every day. Models are always a reduction of reality, and in the words attributed to statistician George Box that essentially "*all models are wrong, but some are useful*" (Box & Draper, 1987, p. 424); instead of being 'right' most of the time, we may well be 'wrong' most of the time. Moreover, a model fitting or performing well in one context should not be taken as guarantee that it fits in other contexts as well. For instance, if some statistical models work well in studies that include novice learners but not advanced learners, we should not generalise the performance of our statistical models to more advanced learners; instead, we need studies that include advanced learners as well. Even if we deal with random samples and allocate participants in these random samples to conditions in an experiment at random, we must be aware of the context in which our experiments are carried out and think carefully to what extent the models and findings from these experiments are generalisable or transferable to other contexts. And when possible, collecting data in other contexts is the best way to study that generalisability or transferability.

References

Abraham, W. T., & Russell, D. W. (2004). Missing data: A review of current methods and applications in epidemiological research. *Current Options in Psychiatry, 17,* 315–321. https://doi.org/10.1097/01.yco.0000133836.34543.7e.

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and the Family, 67,* 1012–1028. https://doi.org/10.1111/j.1741-3737.2005.00191.x.

Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement, 21*(1), 1–23. https://doi.org/10.1177/0146621697211001.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). New York: Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csáki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281), Tsahkadsor, Armenia, USSR, 2–8 September 1971. Budapest: Akadémiai Kiadó.

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In S. Kotz & N. Johnson (Eds.), *Breakthroughs in statistics* (pp. 610–624). New York: Springer.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10*(3), 317–328.

Allison, P. D. (2002). *Missing data* (Vol. 136). Thousand Oaks, CA: Sage.

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician, 32*(3), 307–317. https://doi.org/10.2307/2987937.

Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. New York: Springer.

Anderson, C. J., & Rutkowski, L. (2010). Multinomial logistic regression. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 26, pp. 390–409). London: Sage.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581–594. https://doi.org/10.1177/014662167800200413.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 143–166). Maple Grove, MN: JAM Press.

Arcaya, M., Brewster, M., Zigler, C. M., & Subramanian, S. V. (2012). Area variations in health: A spatial multilevel modeling approach. *Health Place, 18*(4), 824–831. https://doi.org/10.1016/j.healthplace.2012.03.010.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General), 132,* 235–244. https://doi.org/10.2307/2343787.

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40–49. https://doi.org/10.1002/mpr.329.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. London: Macmillan.

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554. https://doi.org/10.1177/1745691612459060.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5–37. https://doi.org/10.1016/j.jsp.2009.10.001.

Bareiss, E. H. (1969). Numerical solution of linear equations with Toeplitz and Vector Toeplitz matrices. *Numerische Mathematik, 13*(5), 404–424. https://doi.org/10.1007/BF02163269.

Barlow, N. M., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change*. Boston: Pearson Education.

Barnard, J., & Meng, X. L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research, 8,* 17–36. https://doi.org/10.1177/096228029900800103.

Baron, R. B., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182.

Bartoń, K. (2018). *Package 'MuMIn'. R package version 1.42.1*. Retrieved April 17, 2019, from https://cran.r-project.org/web/packages/MuMIn/MuMIn.pdf.

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods, 16*(4), 373–390. https://doi.org/10.1037/a0025813.

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, *53*, 370–418. https://doi.org/10.1098/rstl.1763.0053.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., et al. (2018). Redefine statistical significance. *Nature: Human Behaviour*, *2*, 6–10. https://www.nature.com/articles/s41562-017-0189-z.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika, 93*(3), 491–507. https://doi.org/10.1093/biomet/93.3.491.

Berry, K. J., Johnston, J. E., Zahran, S., & Mielke, P. W. (2009). Stuart's tau measure of effect size for ordinal variables: Some methodological considerations. *Behavior Research Methods, 41*(4), 1144–1148. https://doi.org/10.3758/brm.41.4.1144.

Bingenheimer, J. B., & Raudenbush, S. (2004). Statistical and substantive inferences in public health: Issues in the application of multilevel models. *Annual Review of Public Health, 25,* 53–77. https://doi.org/10.1146/annurev.publhealth.25.050503.153925.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8.

Bland, J. M., & Altman, D. G. (2003). Applying the right analysis: Analyses of measurement studies. *Ultrasound in Obstetrics & Gynaecology, 22*(1), 85–93. https://doi.org/10.1102/uog.122.

Bloom, H. S. (2008). In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (Chapter 9, pp. 115–133). London: Sage.

Bohrer, R. (1967). On sharpening Scheffé bounds. *Journal of the Royal Statistical Society*, *29*(1), 110–114. https://www.jstor.org/stable/2984571.

Bond, T., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum. https://doi.org/10.1186/1471-2377-13-78.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo della probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8,* 3–62.

Bordens, K. S., & Abbott, B. B. (2011). *Research design and methods: A process approach*.

Bourdieu, P. (1984). *Distinction: A social critique of the judgment of taste*. Cambridge, MA: Harvard University Press.

Bovaird, J. A., & Embretson, S. E. (2008). Modern measurement in the social sciences. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (Chapter 16, pp. 269–289). London: Sage.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika, 36*(3), 317–346. https://doi.org/10.2307/2332671.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: Wiley.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Briggs, D. C., & Wilson, M. (2004). An introduction to multidimensional measurement using Rasch models. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 322–341). Maple Grove, MN: JAM Press.

Brown, A. M. (2001). A step-by-step guide to non-linear regression analysis of experimental data using a Microsoft Excel spreadsheet. *Computer Methods and Programs in Biomedicine, 65*(3), 191–200. https://doi.org/10.1016/S0169-2607(00)00124-3.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behaviour of some statistics which test the equality of several means. *Technometrics, 16*(1), 129–132. https://doi.org/10.1080/00401706.1974.10489158.

Buchner, A., Erdfelder, E., Faul, F., & Lang, A. G. (2009). *G*Power version 3.1.2*. Retrieved April 17, 2019, from http://www.gpower.hhu.de/.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304. https://doi.org/10.1177/0049124104268644.

Burns, R. B., & Dobson, C. B. (1981). *Experimental psychology: Research methods and statistics*. New York: Springer.

Cameron, C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge, UK: Cambridge University Press.

Cavanaugh, J. E. (1997). Unifying the deviations of the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters, 31,* 201–208. https://doi.org/10.1016/s0167-7152(96)00128-9.

Center for Open Science. (2018). Registered reports: Peer review before results are known to align scientific values and practices. *Center for Open Science*. Retrieved April 17, 2019, from https://cos.io/rr/.

Champoux, J. E., & Peters, W. S. (1987). Form, effect size and power in moderated regression analysis. *Journal of Occupational and Organizational Psychology, 60*(3), 243–255. https://doi.org/10.1111/j.2044-8325.1987.tb00257.x.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide*. New York: Wiley.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.

Coe, R., Waring, M., Hedges, L. V., & Arthur, J. (2017). *Research methods and methodologies in education* (2nd ed.). London: Sage.

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin, 70,* 213–220. https://doi.org/10.1037/h0026256.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. New York: Routledge.

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist, 45*(12), 1304–1312.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003.

Cole, J. C. (2010). How to deal with missing data: Conceptual overview and details for implementing two modern methods. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 15, pp. 214–238). London: Sage.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: Wiley.

Collins, L. M. J. L., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330–351.

Comfrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Erlbaum.

Cook, D. A. (2015). Much ado about differences: Why expert-novice comparisons add little to the validity argument. *Advances in Health Sciences Education, 20*(3), 829–834. https://doi.org/10.1007/s10459-014-9551-3.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago, IL: Rand McNally.

Cook, T. D., & Wong, V. C. (2008). Better quasi-experimental practice. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (Chapter 10, pp. 134–165). London: Sage.

Cox, D. R. (1958). *Planning of experiments*. New York: Wiley.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). New York: Chapman & Hall.

Cragg, J. G., & Uhler, R. S. (1970). The demand for automobiles. *The Canadian Journal of Economics, 3*(3), 386–406. https://doi.org/10.2307/133656.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). New York: Pearson.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30,* 1–14. https://doi.org/10.1037/0003-066X.30.1.1.

Cronbach, L. J. (1976). Equity in selection: When psychometrics and political philosophy meet. *Journal of Educational Measurement, 13,* 31–41. https://doi.org/10.1111/j.1745-3984.1976.tb00179.x.

Cronbach, L. J. (1988). Internal consistency of tests: Analyses old and new. *Psychometrika, 53*(1), 63–70. https://doi.org/10.1007/BF02294194.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418. https://doi.org/10.1177/0013164404266386.

Crutzen, R., & Peters, G. J. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review, 11*(3), 242–247. https://doi.org/10.1080/17437199.2015.1124240.

Darrow, A. L., & Kahl, D. R. (1982). A comparison of moderated regression techniques considering strength of effect. *Journal of Management, 8*(2), 35–47. https://doi.org/10.1177/014920638200800203.

Delattre, M., Lavielle, M., & Poursat, M. A. (2014). A note on BIC in mixed-effects models. *Electronic Journal of Statistics, 8,* 456–475. https://doi.org/10.1214/14-EJS890.

Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement, 77*(2), 185–203. https://doi.org/10.1177/0013164416658325.

De Rooij, M. (2018). Transitional modeling of experimental longitudinal data with missing data. *Advances in Data Analysis and Classification, 12*(1), 107–130. https://doi.org/10.1007/s11634-015-0226-6.

Dey, S., Raheem, E., & Lu, Z. (2016). Multilevel multinomial logistic regression model for identifying factors associated with anemia in children 6-59 months in northeastern states of India. *Cogent Mathematics, 3*(1), 1–12. https://doi.org/10.1080/23311835.2016.1159798.

Ding, C. S. (2018). *Fundamentals of applied multidimensional scaling for educational and psychological research*. New York: Springer. https://doi.org/10.1007/978-3-319-78172-3.

Dodge, H. F., & Romig, H. G. (1929). A method of sampling inspection. *Bell System Technical Journal, 8*(4), 613–631. https://doi.org/10.1002/j.1538-7305.1929.tb01240.x.

Dong, G., Harris, R., Jones, K., & Yu, J. (2015). Multilevel modelling with spatial interaction effects with application to an emerging land market in Beijing, China. *PloS ONE*. https://doi.org/10.1371/journal.pone.0130761.

Dong, N., Kelcey, B., & Spybrook, J. (2017). Power analyses of moderator effects in three-level cluster randomized trials. *Journal of Experimental Education, 86*(3), 489–514. https://doi.org/10.1080/00220973.2017.1315714.

Dong, N., Kelcey, B., Spybrook, J., & Maynard, R. A. (2016). *Designing and analyzing multilevel experiments and quasi-experiments for causal evaluation (Version 1.07)*. Retrieved April 17, 2019, from https://www.causalevaluation.org/power-analysis.html.

Dong, G., Ma, J., Harris, R., & Pryce, G. (2015). Spatial random slope multilevel modeling using multivariate conditional autoregressive models: A case study of subjective travel satisfaction in Beijing. *Annals of the American Association of Geographers, 106*(1), 19–35. https://doi.org/10.1080/00045608.2015.1094388.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required samples sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143.

Dunn, O. J. (1979). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. https://doi.org/10.1111/bjop.12046.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50*(272), 1096–1121. https://doi.org/10.1080/01621459.1955.10501294.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242. https://doi.org/10.1037/h0044139.

Eekhout, I., De Vet, H. C. W., Twisk, J. W. R., Brand, J. P. L., De Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology, 67*(3), 335–342. https://doi.org/10.1016/j.jclinepi.2013.09.009.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7,* 1–26. https://doi.org/10.1007/978-1-4612-4380-9_41.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 63,* 589–599. https://doi.org/10.1093/biomet/68.3.589.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: The Society of Industrial and Applied Mathematics.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association, 78,* 316–331. https://doi.org/10.1080/01621459.1983.10477973.

Efron, B. (1998). R. A. Fisher in the 21st century. *Statistical Science*, *13*(2), 95–122. https://doi.org/10.1214/ss/1028905930.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Enders, C. K., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Eaquation Modeling: A Multidisciplinary Journal, 15*(1), 75–95. https://doi.org/10.1080/10705510701758281.

Engelhard, G. (1994). Historical views of the concept of invariance in measurement theory. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 73–99). Norwood, NJ: Ablex.

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. http://hdl.handle.net/11245/1.380173.

Etherton, J. L., Osborne, R., Stephenson, K., Grace, M., Jones, C., & De Nadai, A. S. (2018). Bayesian analysis of multimethod ego-depletion studies favours the null hypothesis. *British Journal of Social Psychology, 57*(2), 367–385. https://doi.org/10.1111/bjso.12236.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review, 25*(1), 5–34. https://doi.org/10.3758/s13423-017-1262-3.

Eyduran, E., & Akbaş, Y. (2010). Comparison of different covariance structure used for experimental design with repeated measurement. *The Journal of Animal & Plant Sciences, 20*(1), 44–51.

Fasano, G., & Francheschini, A. (1987). A multidimensional version of the Kolmogorov-Smirnov test. *Monthly Notices of the Royal Astronomical Society, 225,* 155–170. https://doi.org/10.1093/mnras/225.1.155.

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or *t*-test? *Statistics Surveys, 4,* 1–39. https://doi.org/10.1214/09-SS051.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*(6), 661–669. https://doi.org/10.1177/1745691612462587.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). London: Sage.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374. https://doi.org/10.1016/0001-6918(73)90003-6.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3–32.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1938). Presidential Address. *Sankhyā: The Indian Journal of Statistics*, *4*(1), 14–17. https://www.jstor.org/stable/40383882.

Fisher, R. A. (1960). *The design of experiments* (7th ed.). New York: Hafner.

Fisher Box, J. (1987). Guinness, Gosset, Fisher, and small samples. *Statistical Science*, *2*(1), 45–52. https://www.jstor.org/stable/2245613.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New York: Wiley.

Fleming, T. R., Harrington, D. P., & O'Brien, P. C. (1984). Designs for group sequential tests. *Contemporary Clinical Trials*. https://doi.org/10.1016/S0197-2456(84)80014-8.

Forbes, S. A., Ross, M. E., & Chesser, S. S. (2011). Single-subject designs and action research in the K-12 setting. *Educational Research and Evaluation, 17,* 161–173. https://doi.org/10.1080/13803611.2011.599555.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology, 44,* 205–231.

Fox-Wasylyshyn, S. M., & El-Masri, M. M. (2005). Handling missing data in self-report measures. *Research in Nursing & Health, 28*(6), 488–495. https://doi.org/10.1002/nur.20100.

Freud, S. (1920). *A general introduction to psychoanalysis*. New York: Boni and Liveright.

Frome, E. L., & Checkoway, H. (1985). Use of Poisson regression models in estimating rates and ratios. *American Journal of Epidemiology, 121*(2), 309–323. https://doi.org/10.1093/oxfordjournals.aje.a114001.

Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. *Journal of Educational and Behavioral Statistics, 1*(2), 113–125. https://doi.org/10.3102/10769986001002113.

Games, P. A., Keselman, H. J., & Clinch, J. J. (1979). Tests for homogeneity of variance in factorial designs. *Psychological Bulletin, 86*(5), 978–984. https://doi.org/10.1037/0033-2909.86.5.978.

Gehan, E. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika, 52*(1/2), 203–223. https://doi.org/10.2307/2333825.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association, 70*(350), 320–328. https://doi.org/10.1080/01621459.1975.10479865.

Gelman, A., & Carlin, J. (2017). Some natural solutions to the *p* value communication problem—And why they won't work. *Journal of the American Statistical Association, 112*(519), 899–901. https://doi.org/10.1080/01621459.2017.1311263.

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics, 15*(3), 373–390. https://doi.org/10.1007/s001800000040.

Giraud, C. (2015). *Introduction to high-dimensional statistics*. Boca Raton, FL: CRC.

Glass, G. V. (1966). Note on rank biserial correlation. *Educational and Psychological Measurement, 26*(3), 623–631. https://doi.org/10.1177/001316446602600307.

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology, 63*(3), 527–537. https://doi.org/10.1348/000711009X475853.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215–231. https://doi.org/10.1093/biomet/61.2.215.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. H. Hoyle (Ed.), *Statistical strategies for small sample size* (pp. 1–29). Thousand Oaks, CA: Sage.

Grant, D. A. (1948). The Latin square principle in the design and analysis of psychological experiments. *Psychological Bulletin, 45*(5), 427–442. https://doi.org/10.1037/h0053912.

Gravetter, F. J., & Forzano, L. A. B. (2006). *Research methods for the behavioral sciences* (2nd ed.). London: Thomson Wadsworth.

Green, S. G., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74,* 169–173. https://doi.org/10.1007/s11336-008-9098-4.

Greenwood, J. D. (1989). *Explanation and experiment in social psychological science*. New York: Springer.

Gruijters, S. L. K. (2016). Baseline comparisons and covariate fishing: Bad statistical habits we should have broken yesterday. *European Health Psychologist, 18,* 205–209.

Grynszpan, O., Weiss, P. L., Perez-Diaz, F., & Gal, E. (2014). Innovative technology-based interventions for autism spectrum disorders: A meta-analysis. *Autism, 18*(4), 346–361. https://doi.org/10.1177/1362361313476767.

Guilford, J. (1936). *Psychometric methods*. New York: McGraw-Hill.

Gunel, E., & Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika, 61*(3), 545–557. https://doi.org/10.1093/biomet/61.3.545.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255–282. https://doi.org/10.1007/BF02288892.

Hacking, I. (1965). *Logic of statistical inference*. Cambridge: Cambridge University Press.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society, 30B*, 67–82. https://www.jstor.org/stable/2984459.

Hambleton, R. K. (1978). On the use of cutoff scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement, 15,* 277–290. https://doi.org/10.1111/j.1745-3984.1978.tb00075.x.

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore: Johns Hopkins University Press.

Hambleton, R. K. (1983). *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care, 38*(9, Suppl. II), II60–II65. https://www.jstor.org/stable/3768063.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage.

Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika, 69*(3), 553–566. https://doi.org/10.1093/biomet/69.3.553.

Hatcher, L. (1994). *A step-by-step approach to using the SAS® system for factor analysis and structural equation modelling*. Cary, NC: SAS Institute.

Hauck, W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics, 12*(1), 83–91. https://doi.org/10.1007/BF01063612.

Hayes, A. F. (2017). *The PROCESS macro for SPSS and SAS*. Retrieved April 17, 2019, from http://www.processmacro.org/index.html.

Hayes, A. F. (2018). Partial, conditional, and moderated moderated mediation: Quantification, inference, and interpretation. *Communication Monographs, 85*(1), 4–40.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication, 16*(3), 354–367. https://doi.org/10.1177/0741088399016003004.

Hayes, R., & McArdle, J. J. (2017). Should we impute or should we weight? Examining the performance of two CART-based techniques for addressing missing data in small sample research with nonnormal variables. *Computational Statistics and Data Analysis, 115,* 35–52. https://doi.org/10.1016/j.csda.2017.05.006.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9, Suppl. II), II28–II42. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1815384/.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine, 22*(9), 1433–1446. https://doi.org/10.1002/sim.1522.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50*(4), 933–944. https://doi.org/10.2307/2533433.

Hedeker, D., & Gibbons, R. D. (1996). MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine, 49*(2), 157–176. https://doi.org/10.1016/0169-2607(96)01720-8.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107.

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness, 11*(1), 1–21. https://doi.org/10.1080/19345747.2017.1375583.

Hernan, M. A., Clayton, D., & Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology, 40*(3), 780–785. https://doi.org/10.1093/ije/dyr041.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733.

Horowitz, J. L., & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association, 95*(449), 77–84. https://doi.org/10.1080/01621489.2000.10473902.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice. *The American Statistician, 55*(3), 244–254. https://doi.org/10.1198/000313001317098266.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time-to-event data*. New York: Wiley.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and Applications* (3rd ed.). New York: Taylor & Francis.

Howell, D. C. (2010). Best practices in the analysis of variance. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 23, pp. 341–357). London: Sage.

Howell, D. C. (2017). *Statistical methods for psychology* (8th ed.). Boston: Cengage.

Hoyle, R. H. (2000). Confirmatory factor analysis. In H. E. A. Tinsley, & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (Chapter 16, pp. 465–497). Academic Press. https://doi.org/10.1016/B978-012691360/50017-3.

Hoyle, R. H. (2008). Latent variable models of social research data. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (Chapter 23, pp. 395–413). London: Sage.

Huitema, B. E. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies* (2nd ed., Part VII, pp. 565–617). New York: Wiley.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76,* 297–307. https://doi.org/10.1093/biomet/76.2.297.

Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistician, 65*(332), 1582–1589. https://doi.org/10.1080/01621459.1970.10481187.

Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational and Behavioral Statistics, 1*(1), 69–82. https://doi.org/10.3102/10769986001001069.

IBM Corporation. (2017). *SPSS version 25*. Retrieved April 17, 2019, from https://www-01.ibm.com/support/docview.wss?uid=swg24043678.

Iramaneerat, C., Smith, E. V., Jr., Smith, R. M. (2010). An introduction to Rasch measurement. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 4, pp. 50–70). London: Sage.

Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294*(2), 218–228. https://doi.org/10.1001/jama.294.2.218.

Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS, Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124.

Jamovi Project. (2019). *Jamovi version 0.9.5.16*. Retrieved April 17, 2019, from https://www.jamovi.org/.

Janssen, K. J. M., Donders, A. R. T., Harrell, F. E., Vergouwe, Y., Chen, Q., Grobbee, D. E., et al. (2010). Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology, 63*(7), 721–727. https://doi.org/10.1016/j.jclinepi.2009.12.008.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*(3), 1070–1085. https://doi.org/10.3758/s13428-015-0631-y.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26. https://doi.org/10.3102/0013189X033007014.

Justel, A., Peña, D., & Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters, 35*(3), 251–259. https://doi.org/10.1016/S0167-7152(97)00020-5.

Kahan, B. C., Jairath, V., Doré, C. J., & Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Methodology, 15,* 139. https://doi.org/10.1186/1745-6215-15-139.

Kalamaras, D. V. (2018). *Social network visualizer version 2.4*. Retrieved April 17, 2019, from http://socnetv.org/.

Kalyuga, S., & Singh, A. M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review, 28*(4), 831–852. https://doi.org/10.1007/s10648-015-9352-0.

Kane, J. S. (2013). *Beyond ANCOVA: A new method for excluding the influence of covariates in comparing group means*. Retrieved April 17, 2019, from https://www.prostatservices.com/articles/beyond-ancova-a-new-method-for-excluding-the-influence-of-covariates-in-comparing-group-means.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete data. *Journal of the American Statistical Association, 53*(282), 457–481. https://doi.org/10.2307/2281868.

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26,* 379–424. https://doi.org/10.1080/07370000802212669.

Kapur, M. (2011). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science, 39,* 561–579. https://doi.org/10.1007/s11251-010-9144-3.

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science, 38,* 1008–1022. https://doi.org/10.1111/cogs.12107.

Kapur, M., & Rummel, N. (2012). Productive failure in learning from generation and invention activities. *Instructional Science, 40,* 645–650. https://doi.org/10.1007/s11251-012-9235-4.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association, 90*(431), 928–934. https://doi.org/10.1080/01621459.1995.10476592.

Kelcey, B., Dong, N., Spybrook, J., & Cox, K. (2017). Statistical power for causally defined indirect effects in group-randomized trials with individual-level mediators. *Journal of Educational and Behavioral Statistics, 42*(5), 499–530. https://doi.org/10.3102/1076998617695506.

Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017). Experimental power for indirect effects in group-randomized studies with group-level mediators. *Multivariate Behavioral Research, 52*(6), 699–719. https://doi.org/10.1080/00273171.2017.1356212.

Kellow, J. T., & Willson, V. L. (2010). Setting standards and establishing cut scores on criterion-referenced assessments: Some technical and practical considerations. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 2, pp. 15–28). London: Sage.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1–2), 81–89. https://doi.org/10.1093/biomet/30.1-2.81.

Kendall, M. G. (1962). *Rank correlation methods*. New York: Hafner.

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Prentice Hall.

Kirschner, P. A., & Van Merriënboer, J. J. G. (2013). Do learners really know best? Urban legends in education. *Educational Psychologist, 48*(3), 169–183. https://doi.org/10.1080/00461520.2013.804395.

Kish, L. (1965). *Survey sampling*. New York: Wiley. https://doi.org/10.1002/bimj.19680100122.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: APA.

Kleinbaum, D. G. (1996). *Survival analysis, a self learning text*. New York: Springer.

Konstantopoulos, S. (2010). An introduction to meta-analysis. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 12, pp. 177–196). London: Sage.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609–622. https://doi.org/10.1037/0278-7393.32.3.609.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x.

Kruschke, J. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. https://doi.org/10.1177/1745691611406925.

Kruschke, J. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573–603. https://doi.org/10.1037/a0029146.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Boston: Academic Press.

Kruschke, J. (2018). *BEST: Bayesian estimation supersedes the t-test (R package)*. Retrieved April 17, 2019, from https://cran.r-project.org/web/packages/BEST/BEST.pdf.

Kruschke, J., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association, 53*(284), 814–861. https://doi.org/10.2307/2281954.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391.

Kurtz, A. K. (1948). A research test of Rorschach test. *Personnel Psychology, 1,* 41–53. https://doi.org/10.1111/j.1744-6570.1948.tb01292.x.

Kvålseth, T. O. (1985). Cautionary note about R^2. *The American Statistician, 39,* 279–285. https://doi.org/10.1080/00031305.1985.10479448.

Lai, T. L., Shih, M. C., & Zhu, G. (2006). Modified Haybittle-Peto group sequential designs for testing superiority and non-inferiority hypotheses in clinical trials. *Statistics in Medicine, 25,* 1149–1167. https://doi.org/10.1002/sim.2357.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44,* 701–710. https://doi.org/10.1002/ejsp.2023.

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177.

Lakens, D. (2018). *TOSTER: Two one-sided tests (TOST) equivalence testing*. Retrieved April 17, 2019, from https://CRAN.R-project.org/package=TOSTER.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., et al. (2018). Justify your alpha. *Nature: Human Behaviour*, *2*, 168–171. https://www.nature.com/articles/s41562-018-0311-x.

Lanza, L. M., & Collins, S. T. (2008). A new SAS procedure for latent transition analysis: Transitions in dating and sexual behaviour. *Developmental Psychology, 42*(2), 446–456. https://doi.org/10.1037/0012-1649.44.2.446.

Laplace, P. (1812). *Analytique des probabilités*. Paris: Courcier.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Le, C. T. (1998). *Applied categorical data analysis*. New York: Wiley.

Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed.). New York: Wiley.

Lee, C. B., Hanham, J., & Leppink, J. (2019). *Instructional design principles for high-stakes problem-solving environments*. Singapore: Springer. https://doi.org/10.1007/978-981-13-2808-4.

Leppink, J. (2015a). Data analysis in medical education research: A multilevel perspective. *Perspectives on Medical Education, 4*(1), 14–24. https://doi.org/10.1007/s40037-015-0160-5.

Leppink, J. (2015b). On causality and mechanisms in medical education research: An example of path analysis. *Perspectives on Medical Education, 4*(2), 66–72. https://doi.org/10.1007/s40037-015-0174-z.

Leppink, J. (2017a). When I say … time on task. *Medical Education, 51,* 1101–1102. https://doi.org/10.1111/medu.13298.

Leppink, J. (2017b). Science fiction in medical education: The case of learning styles. *Journal of Graduate Medical Education, 9*(3), 394. https://doi.org/10.4300/JGME-D-16-00637.1.

Leppink, J. (2018a). A pragmatic approach to statistical testing and estimation (PASTE). *Health Professions Education, 4*(3), 329–339. https://doi.org/10.1016/j.hpe.2017.12.009.

Leppink, J. (2018b). Analysis of covariance (ANCOVA) vs. moderated regression (MODREG): Why the interaction matters. *Health Professions Education*, *4*(3), 225–232. https://doi.org/10.1016.j.hpe.2018.04.001.

Leppink, J. (2018c). The art of acknowledging that we know nearly nothing. *Health Professions Education, 4*(2), 67–69. https://doi.org/10.1016/j.hpe.2018.03.004.

Leppink, J., O'Sullivan, P. S., & Winston, K. (2017). The bridge between design and analysis. *Perspectives on Medical Education, 6*(4), 265–269. https://doi.org/10.1007/s40037-017-0367-8.

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods, 45*(4), 1058–1072. https://doi.org/10.3758/s13428-013-0334-1.

Leppink, J., Paas, F., Van Gog, T., Van der Vleuten, C. P. M., & Van Merriënboer, J. J. G. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction, 30,* 32–42. https://doi.org/10.1016/j.learninstruc.2013.12.001.

Leppink, J., & Pérez-Fuster, P. (2017). We need more replication research—A case for test-retest reliability. *Perspectives on Medical Education, 6*(3), 158–164. https://doi.org/10.1007/s40037-017-0347-z.

Leppink, J., & Pérez-Fuster, P. (2018). Social networks as an approach to systematic review. *Health Professions Education*, online ahead of print. https://doi.org/10.1016/j.hpe.2018.09.002.

Leppink, J., & Pérez-Fuster, P. (2019). Mental effort, workload, time on task, and certainty: Beyond linear models. *Educational Psychology Review*, online ahead of print. https://doi.org/10.1007/s10648-018-09460-2.

Leppink, J., & Van den Heuvel, A. (2015). The evolution of cognitive load theory and its application to medical education. *Perspectives on Medical Education, 4*(3), 119–127. https://doi.org/10.1007/s40037-015-0192-x.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions fo probability and statistics: Essays in honor of Harold Hotelling*. Bloomington: Stanford University Press.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communications Research, 28*(4), 612–625. https://doi.org/10.1111/j.1468-2958.2002.tb00828.x.

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 26–46. https://doi.org/10.1037/amp0000151.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 1–55.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine, 19,* 1793–1819. https://doi.org/10.1002/1097-0258(20000715)19:13%3c1793:AID-SIM482%3e3.0.CO;2-Q.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202. https://doi.org/10.1080/01621459.1988.10478722.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 56,* 362–375. https://doi.org/10.1016/j.jmp.2008.03.002.

Liu, X. (2016). *Applied ordinal logistic regression using Stata: From single-level to multilevel modeling*. New York: Sage.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Lorenzo-Seva, U., & Ferrando, P. J. (2013). FACTOR 9.2: A comprehensive program for fitting exploratory and semiconfirmatory factor analysis and IRT models. *Applied Psychological Measurement*, *37*(6), 497–498. https://doi.org/10.1177/0146621613487794.

Love, J., Selker, R., Marsman, M., et al. (2018). *JASP version 0.9.2.0*. Retrieved April 17, 2019, from https://jasp-stats.org/.

Lu, K., & Mehrotra, D. V. (2009). Specification of covariance structure in longitudinal data analysis for randomized clinical trials. *Statistics in Medicine, 29*(4), 474–488. https://doi.org/10.1002/sim.3820.

Luo, Y., Szolovits, P., Dighe, A. S., & Baron, J. M. (2017). 3D-MICE: Integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. *Journal of the American Medical Informatics Association, 25*(6), 645–653. https://doi.org/10.1093/jamia/ocx133.

Ly, A., Verhagen, A. J., & Wagenmakers, E. J. (2016). Harold Jeffrey's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72,* 19–31. https://doi.org/10.1016/j.jmp.2015.06.004.

Ma, J., Chen, Y., & Guanpeng, D. (2017). Flexible spatial multilevel modeling of neighborhood satisfaction in Beijing. *The Professional Geographer, 70*(1), 11–21. https://doi.org/10.1080/00330124.2017.1298453.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36,* 611–637. https://doi.org/10.1207/S15327906MBR3604_06.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

Mair, P., & Wilcox, R. R. (2018). *WRS2: A collection of robust statistical methods*. R package version 0.10-0. Retrieved April 17, 2019, from https://cran.r-project.org/web/packages/WRS2/index.html.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*(1), 50–60. https://www.jstor.org/stable/2236101.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports, 50*(3), 163–170. https://doi.org/10.1093/jnci/22.4.719.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748.

Martin, A. J., Bobis, J., Anderson, J., Way, J., & Vellar, R. (2011). Patterns of multilevel variance in psycho-educational phenomena: Comparing motivation, engagement, climate, teaching, and achievement factors. *Zeitschrift für Pädagogische Psychologie, 25,* 49–61. https://doi.org/10.1024/1010-0652/a000029.

Martin, L. J., Turnqvist, A., Groot, B., Huang, S. Y. M., Kok, E. M., Thoma, B., & Van Merriënboer, J. J. G. (2018). Exploring the role of infographics for summarizing medical literature. *Health Professions Education*, online ahead of print. https://doi.org/10.1016/j.hpe.2018.03.005.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272.

Masters, G. N., & Wright, B. D. (1996). The partial credit model. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–122). New York: Springer.

Maxwell, S. E., Delaney, H. D., & Manheimer, J. M. (1985). ANOVA of residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs. *Journal of Educational and Behavioral Statistics, 10*(3), 197–209. https://doi.org/10.3102/10769986010003197.

McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary psychological research. In J. Schinka & W. Velicer (Eds.), *Comprehensive handbook of psychology: Research methods in psychology* (Vol. 2, pp. 447–480). New York: Wiley.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological), 42*(2), 109–142. https://www.jstor.org/stable/2984952.

McCutcheon, A. L. (1987). *Latent class analysis*. London: Sage.

McDonald, R. P. (1978). Generalizability in factorable domains: Domain validity and generalizability. *Educational and Psychological Measurement, 38*(1), 75–79. https://doi.org/10.1177/001316447803800111.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers in econometrics*. Berkeley, CA: Academic Press.

McQuarrie, A. D. R., & Tsai, C. L. (1998). *Regression and time series model selection*. World Scientific.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician, 54*(1), 17–24. https://doi.org/10.1080/00031305.2000.10474502.

Meredith, W. M., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55,* 107–122. https://doi.org/10.1007/BF02294746.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463–477. https://doi.org/10.1016/j.jml.2004.12.001.

Metsämuuronen, J. (2017). *Essentials of research methods in human sciences*. London: Sage.

Michiels, B., Heyvaert, M., Meulders, A., & Onghena, P. (2017). Confidence intervals for single-case effect size measures based on randomization test inversion. *Behavior Research Methods, 49,* 363–381. https://doi.org/10.3758/s13428-016-0714-4.

Michiels, B., & Onghena, P. (2018). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods*, online ahead of print. https://doi.org/10.3758%2Fs13428-018-1084-x.

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive being a connected view of the principles of evidence, and the methods of scientific investigation*. London: Harrison and co.

Miller, R. G. (1997). *Survival analysis*. New York: Wiley.

Mittlbock, M., & Schemper, M. (1996). Explained variation in logistic regression. *Statistics in Medicine, 15,* 1987–1997. https://doi.org/10.1002/(SICI)1097-0258(19961015)15:19%3c1987:AID-SIM318%3e3.0.CO;2-9.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. New York: De Gruyter.

Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies*. Hoboken, NJ: Wiley.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. Berlin: Springer.

Montgomery, A. A., Peters, T. J., & Little, P. (2003). Design, analysis and presentation of factorial randomised controlled trials. *BMC Medical Research Methodology, 3,* 26–30. https://doi.org/10.1186/1471-2288-3-26.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11,* 5–11.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Mueller, R. O., & Hancock, G. R. (2010). Best practices in structural equation modelling. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 32, pp. 488–508). London: Sage.

Mulaik, S. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research, 22*(3), 267–305. https://doi.org/10.1207/s15327906mbr2203_3.

Muthén, L. K., & Muthén, B. (2017). *Mplus user's guide, version 8*. Retrieved April 17, 2019, from https://statmodel.com.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691–692.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), 1–11. https://doi.org/10.6084/m9.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4,* 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Newman, S. C. (2004). Commonalities in the classical, collapsibility and counterfactual concepts of confounding. *Journal of Clinical Epidemiology, 57*(4), 325–329. https://doi.org/10.1016/j.jclinepi.2003.07.014.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15,* 625–632. https://doi.org/10.1007/s10459-010-9222-y.

Normand, M. P. (2016). Less is more: Psychologists can learn more by studying fewer people. *Frontiers in Psychology, 7,* 934. https://doi.org/10.3389/fpsyg.2016.00934.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: MacGraw-Hill.

Nussbaum, E. M., Elsadat, S., & Khago, A. M. (2010). In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 21, pp. 306–323). London: Sage.

Onghena, P., Maes, B., & Heyvaert, M. (2018). Mixed methods single case research: State of the art and future directions. *Journal of Mixed Methods Research*, online ahead of print. https://doi.org/10.1177/1558689818789530.

Osborne, J. W. (2010a). Best practices in data transformation: The overlooked effect of minimum values. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 13, pp. 197–204). London: Sage.

Osborne, J. W. (2010b). Creating valid prediction equations in multiple regression: Shrinkage, double cross-validation, and confidence intervals around predictions. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 20, pp. 299–305). London: Sage.

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2010). Best practices in exploratory factor analysis. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 6, pp. 86–99). London: Sage.

Osborne, J. W., & Overbay, A. (2010). Best practices in data cleaning: How outliers and "fringeliers" can increase error rates and decrease the quality and precision of your results. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 14, pp. 205–213). London: Sage.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skills in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9,* 105–119. https://doi.org/10.1111/j.1539-6053.2009.01038.x.

Patall, E. A., & Cooper, H. (2008). Conducting a meta-analysis. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (Chapter 32, pp. 536–554). London: Sage.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 50*(5), 157–175. https://doi.org/10.1080/14786440009463897.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ, 316*(7139), 1236–1238. https://doi.org/10.1136/bmj.316.7139.1236.

Peters, G. J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist, 16*(2), 56–69.

Peters, G. J. Y. (2017). *Userfriendlyscience: Quantitative analysis made accessible*. R package version 0.7-0. Retrieved April 17, 2019, from https://userfriendlyscience.com/.

Peterson, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association, 72*(360), 854–858. https://doi.org/10.2307/2286474.

Peto, R., & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, *135*(2), 185–207. https://doi.org/10.2307/2344317.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., et al. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II. Analysis and examples. *British Journal of Cancer, 35,* 1–39. https://doi.org/10.1038/bjc.1977.1.

Phan, H. P., & Ngu, B. H. (2017). Undertaking experiments in social sciences: Sequential, multiple time series designs for consideration. *Educational Psychology Review, 29,* 847–867. https://doi.org/10.1007/s10648-016-9368-0.

Plackett, R. L. (1950). Some theorems in least squares. *Biometrika, 37*(1–2), 149–157. https://doi.org/10.1093/biomet/37.1-2.149.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*(2), 191–199. https://doi.org/10.1093/biomet/64.2.191.

Poulton, E. C., & Edwards, R. S. (1979). Asymmetric transfer in within-subjects experiments on stress interactions. *Ergonomics, 22*(8), 945–961. https://doi.org/10.1080/00140137908924669.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods: Instruments & Computers, 36*(4), 717–731. https://doi.org/10.3758/BF03206553.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society: Series B (Methodological), 11,* 18–84. https://doi.org/10.1017/S0305004100025123.

Rao, C. R. (1958). Some statistical methods for the comparison of growth curves. *Biometrics*, *14*, 1–17. https://doi.org/0.2307/2527726.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.

R Core Team. (2018). *R: Language and environment for statistical computing*. R foundation for statistical computing, version 3.5.0. Retrieved April 17, 2019, from www.r-project.org/.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74,* 145–154. https://doi.org/10.1007/s11336-008-9102-z.

Richardson, J. T. E. (2018). The use of Latin-square designs in educational and psychological research. *Educational Research Review, 24,* 84–97. https://doi.org/10.1016/j.edurev.2018.03.003.

Rodgers, J. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34,* 441–456. https://doi.org/10.1207/S15327906MBR3404_2.

Rosnow, R. L., & Rosenthal, R. (2005). *Beginning behavioral research: A conceptual primer* (5th ed.). London: Pearson Prentice-Hall.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271–282. https://doi.org/10.1177/014662169001400305.

Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *British Journal of Mathematical and Statistical Psychology, 44,* 75–92. https://doi.org/10.1111/j.2044-8317.1991.tb00951.x.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*(3), 537–560. https://doi.org/10.1111/j.1744-6570.1994.tb01736.x.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1,* 43–46.

Rouder, J. N., Engelhard, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in ANOVA. *Psychonomic Bulletin & Review, 23*(6), 1779–1786. https://doi.org/10.3758/s13423-016-1026-5.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research, 47*(6), 877–903.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001.

Rouder, J. N., Morey, R. D., Verhagen, A. J., Swagman, A. R., & Wagenmakers, E. J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22*(2), 304–321. https://doi.org/10.1037/met0000057.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2012). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225.

Royall, R. M. (1997). *Statistical evidence: A Likelihood paradigm*. London: Chapman & Hall.

Royall, R. M. (2004). The likelihood paradigm for statistical evidence. In M. L. Taper & S. R. Lele (Eds.), *The nature of scientific evidence*. Chicago: University of Chicago Press.

RStudio Team. (2018). *RStudio version 1.1.456*. Retrieved April 17, 2019, from https://www.rstudio.com/.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*(1), 34–58. https://www.jstor.org/stable/2958688.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rummel, R. J. (1970). *Applied factor analysis*. Evanston: Northwestern University Press.

Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229. https://doi.org/10.1147/rd.33.0210.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schagen, I., & Elliot, K. (2004). *But what does it mean? The use of effect sizes in educational research*. London: National Foundation for Educational Research.

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Scher, A. M., Young, A. C., & Meredith, W. M. (1960). Factor analysis of the electrocardiogram: Tests of electrocardiographic theory: Normal hearts. *Circulation Research, 8,* 519–526. https://doi.org/10.1161/01.RES.8.3.519.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6,* 461–465.

Scott, J. (1988). Social network analysis. *Sociology, 22*(1), 109–127. https://doi.org/10.1177/0038038588022001007.

Scott, J. (2017). *Social network analysis* (4th ed.). London: Sage.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician, 55*(1), 62–71. https://doi.org/10.1198/000313001300339950.

Sewell, J. L., Maggio, L. A., Ten Cate, O., Van Gog, T., Young, J. Q., & O'Sullivan, P. S. (2018). Cognitive load theory for training health professionals in the workplace: A BEME review of studies among diverse professions: BEME Guide No. 53. *Medical Teacher*, online ahead of print. https://doi.org/10.1080/0142159X.2018.1505034.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Sherman, R. (2014). *phack: An R function for examining the effects of p-hacking*. Retrieved April 17, 2019, from http://rynesherman.com/blog/phack-an-r-function-for-examining-the-effects-of-p-hacking/.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120. https://doi.org/10.1007/s11336-008-9101-0.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Silberzahn, R., Uhlman, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356. https://doi.org/10.1177/2515245917747646.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics, 19,* 279–281. https://doi.org/10.1214/aoms/1177730256.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*(4), 510–550. https://doi.org/10.1037/a0029312.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modelling* (2nd ed.). London: Sage.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x.

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics, 41*(6), 605–627. https://doi.org/10.3102/1076998616655442.

StataCorp. (2017). *Stata Statistical Software: Release 15.1*. College Station, TX: StataCorp LLC. Retrieved April 17, 2019, from https://www.stata.com.

Stemler, S. E., & Tsai, J. (2010). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 3, pp. 29–49). London: Sage.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association, 69*(347), 730–737. https://doi.org/10.2307/2286009.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680. http://www.jstor.org/stable/1671815.

Stone, M. (1974). Cross validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B (Methodological)*, *26*, 111–147. https://www.jstor.org/stable/pdf/2984809.

Stuart, E. A., & Rubin, D. B. (2010). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 11, pp. 155–176). London: Sage.

Sweller, J. (2018). Measuring cognitive load. *Perspectives on Medical Education, 7*(1), 1–2. https://doi.org/10.1007/s40037-017-0395-4.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296. https://doi.org/10.1023/A:1022193728205.

Tarone, R. E., & Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, *64*, 156–160. https://doi.org/10.93/biomet/64.1.156.

Tacq, J. J. A. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*. London: Sage.

Tacq, J. J. A., & Nassiri, V. (2011). How ordinal is ordinal in ordinal data analysis? In *Proceedings of the 58th World Statistical Congress*. Dublin, Ireland. http://2011.isiproceedings.org/papers/950432.pdf.

Tan, F. E. S. (2010). Best practices in analysis of longitudinal data: A multilevel approach. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 30, pp. 451–470). London: Sage.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*(4), 613–625. https://doi.org/10.1007/BF02289858.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 1024–1037. https://doi.org/10.1037/0278-7393.25.4.1024.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association. https://doi.org/10.1037/10694-000.

Thorndike, R. L. (1904). *An introduction to the theory of mental and social measurements*. New York: Science Press.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.

Tiffin, P. A., & Paton, L. W. (2018). Rise of the machines? Machine learning approaches and mental health: Opportunities and challenges. *The British Journal of Psychiatry, 213,* 509–510. https://doi.org/10.1192/bjp.2018.105.

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, *41*(59), 472–505. https://doi.org/10.1177/0193841X6655665.

Tjur, T. (2009). Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician, 63*(4), 366–372. https://doi.org/10.1198/tast.2009.08210.

Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models*. Greenwich, CT: Information Age Publishing.

Toon, E., Timmerman, C., & Worboys, M. (2016). Text-mining and the history of medicine: Big Data, big questions? *Medical History, 60*(2), 294–296. https://doi.org/10.1017/mdh.2016.18.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7,* 769. https://doi.org/10.3389/fpsyg.2016.00769.

Tryon, W. W. (2018). Mediators and mechanisms. *Clinical Psychological Science, 6*(5), 619–628. https://doi.org/10.1177/2167702618765791.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika, 23,* 19–23. https://doi.org/10.1007/BF02288975.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*(2), 99–114. https://doi.org/10.2307/3001913.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics, 29,* 614.

Twisk, J. W. R., Bosman, L., Hoekstra, T., Rijnhart, J., Welten, M., & Heymans, M. (2018). Different ways to estimate treatment effects in randomised controlled trials. *Contemporary Clinical Trials Communications, 10,* 80–85. https://doi.org/10.1016/j.conctc.2018.03.008.

Twisk, J. W. R., Hoogendijk, E. O., Zwijsen, S. A., & De Boer, M. R. (2016). Different methods to analyze stepped wedge trials designs revealed different aspects of intervention studies. *Journal of Clinical Epidemiology, 72,* 75–83. https://doi.org/10.1016/j.clinepi.2015.11.004.

Van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., et al. (2014). A new method for constructing networks from binary data. *Scientific Reports, 4* (5918), 1–10. https://doi.org/10.1038/srep05918.

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline: More power in randomized studies, more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59*(9), 920–925. https://doi.org/10.1016/j.jclinepi.2006.02.007.

Van Breukelen, G. J. P., & Van Dijk, K. R. A. (2007). Use of covariates in randomized controlled trials. *Journal of the International Neuropsychological Society, 13*(5), 903–904. https://doi.org/10.1017/S1355617707071147.

Van Buuren, S. (2012). *Flexible imputation of missing data*. New York: Chapman & Hall.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45,* 1–67.

Van der Eijk, C., & Rose, J. (2015). Risky business: Factor analysis of survey data—Assessing the probability of incorrect dimensionalisation. *PLoS ONE, 10*(3), 1–31. https://doi.org/10.1371/journal.pone.0118900.

Van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamic model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842.

Van der Zee, T., & Reich, J. (2018). Open education science. *AERA Open, 4*(3), 1–15. https://doi.org/10.1177/2332858418787466.

Van Ginkel, J. R., Sijtsma, K., Van der Ark, L. A., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology, 6,* 17–30. https://doi.org/10.1027/1614-2241/a000003.

Van Merriënboer, J., & Kirschner, P. (2018). *Ten steps to complex learning* (3rd Rev. ed.). New York: Routledge.

Van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis*. Thousand Oaks: Sage.

Veenman, M. V., Prins, F. J., & Verheij, J. (2003). Learning styles: Self-reports versus thinking-aloud measures. *British Journal of Educational Psychology, 73*(3), 357–372. https://doi.org/10.1348/000709903322275885.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Viechtbauer, W. (2010). Analysis of moderator effects in meta-analysis. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 31, pp. 471–487). London: Sage.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p*-values. *Psychonomic Bulletin & Review, 14,* 779–804. https://doi.org/10.3758/BF03194105.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist, 31*(6–7), 1113–1126. https://doi.org/10.1080/13854046.2017.1317364.

Weakliem, D. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research, 27*(3), 359–397. https://doi.org/10.1177/0049124199027003002.

Weitzenhoffer, A. M. (1951). Mathematical structures and psychological measurement. *Psychometrika, 16,* 387–406. https://doi.org/10.1007/BF02288802.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34,* 28–35. https://doi.org/10.2307/2332510.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38*(3/4), 330–336. https://doi.org/10.2307/2332579.

West, S. G., & Thoemmes, F. (2008). Equating groups. In P. Alasuutari, L. Bickman, & J. Brannen (Eds.), *The SAGE handbook of social research methods* (Chapter 24, pp. 414–430). London: Sage.

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika, 84*(2), 419–427. https://doi.org/10.1093/biomet/84.2.419.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for ANOVA designs. *The American Statistician, 66*(2), 104–111. https://doi.org/10.1080/00031305.2012.695956.

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review, 19*(6), 1057–1064. https://doi.org/10.3758/s13423-012-0295-x.

Whisman, M. A., & McClelland, G. H. (2005). Designing, testing, and interpreting interactions and moderator effects in family research. *Journal of Family Psychology, 19,* 111–120.

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine, 29,* 2920–2931. https://doi.org/10.1002/sim.3944.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Burlington, MA: Elsevier.

Wilcox, R. R., & Tian, T. (2011). Measuring effect size: A robust heteroscedastic approach for two or more groups. *Journal of Applied Statistics, 38*(7), 1359–1368. https://doi.org/10.1080/02664763.2010.498507.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80–83. https://doi.org/10.2307/3001968.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics, 9*(1), 60–62. https://doi.org/10.1214/aoms/1177732360.

Willson, V. L. (1976). Critical values of the rank-biserial correlation coefficient. *Educational and Psychological Measurement, 36*(2), 297–300. https://doi.org/10.1177/001316447603600207.

Wilson, M. (1989). Saltus: A psychometric model for discontinuity in cognitive development. *Psychological Bulletin, 105,* 276–289.

Winer, B. J. (1970). *Statistical principles in experimental design: international* (student ed.). London: McGraw-Hill.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Wolfe, E. W., & Dobria, L. (2010). Applications of the multifaceted Rasch model. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 6, pp. 71–85). London: Sage.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials, 1*(4), 368–376. https://doi.org/10.1191/1740774504cn032oa.

Yerkes, R. M. (1921). Psychological examining in the United States Army. In W. Dennis (Ed.), *Reading in the history of psychology* (p. 1948). New York: Appleton-Century-Crofts.

Yu, C. H. (2010). Resampling: A conceptual and procedural introduction. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Chapter 19, pp. 283–298). London: Sage.

Yuan, K. H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research, 41*(4), 598–629. https://doi.org/10.1177/0049124112460373.

Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. *Biometrika, 61*(1), 165–170. https://doi.org/10.1093/biomet/61.1.165.

Zhang, Z. Y., & Yuan, K. H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement, 76*(3), 387–411. https://doi.org/10.1177/0013164415594658.

Contact and Website

For a quick tour through the book and data files, syntax, and worked examples of studies discussed in this book, please visit: https://wordpress.com/view/research489962293.wordpress.com. The website also has a contact form, which allows you to send emails to me. For any questions, comments, or suggestions (which I will consider for an eventual next edition), please use the contact form on the website or get in touch with me directly via j.leppink@gmail.com (Gmail) or hyjl17@hyms.ac.uk (my Outlook email address at Hull York Medical School).

© Springer Nature Switzerland AG 2019

J. Leppink, *Statistical Methods for Experimental Research in Education and Psychology*, Springer Texts in Education,
https://doi.org/10.1007/978-3-030-21241-4