Springer Proceedings in Mathematics & Statistics

Marie Wiberg · Steven Culpepper
Rianne Janssen · Jorge González
Dylan Molenaar *Editors*

Quantitative Psychology

The 82nd Annual Meeting of the Psychometric Society, Zurich, Switzerland, 2017

Springer

Springer Proceedings in Mathematics & Statistics

Volume 233

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Marie Wiberg · Steven Culpepper
Rianne Janssen · Jorge González
Dylan Molenaar
Editors

Quantitative Psychology

The 82nd Annual Meeting of the Psychometric Society, Zurich, Switzerland, 2017

Springer

Editors
Marie Wiberg
Umeå School of Business,
Economics and Statistics
Umeå University
Umeå
Sweden

Jorge González
Faculty of Mathematics
Pontificia Universidad Católica
de Chile
Santiago
Chile

Steven Culpepper
Department of Statistics
University of Illinois
at Urbana-Champaign
Champaign, IL
USA

Dylan Molenaar
Department of Psychology
University of Amsterdam
Amsterdam
The Netherlands

Rianne Janssen
Faculty of Psychology
and Educational Sciences
KU Leuven
Leuven
Belgium

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-77248-6 ISBN 978-3-319-77249-3 (eBook)
https://doi.org/10.1007/978-3-319-77249-3

Library of Congress Control Number: 2018934883

Mathematics Subject Classification (2010): 62P15, 62-06, 62H12, 62F15, 62-07

Printed on acid-free paper

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

Preface

This volume represents presentations given at the 82nd annual meeting of the Psychometric Society, organized by the University of Zurich, and held in Zurich, Switzerland, during July 17–21, 2017. The meeting was one of the largest Psychometric Society meetings in the Society's history, both in terms of participants and number of presentations. It attracted 521 participants, with 295 papers being presented, of which 91 were part of a symposium. There were 105 poster presentations, 3 pre-conference workshops, 3 keynote presentations, 4 invited presentations, 2 career award presentations, 4 state-of-the-art presentations, 1 dissertation award winner, and 22 symposia.

Since the 77th meeting in Lincoln, Nebraska, Springer publishes the proceedings volume from the annual meeting of the Psychometric Society so as to allow presenters to quickly make their ideas available to the wider research community, while still undergoing a thorough review process. The first five volumes of the meetings in Lincoln, Arnhem, Madison, Beijing, and Asheville were received successfully, and we expect a successful reception of these proceedings too.

We asked authors to use their presentation at the meeting as the basis of their chapters, possibly extended with new ideas or additional information. The result is a selection of 34 state-of-the-art chapters addressing a diverse set of psychometric topics, including item response theory, factor analysis, causal inference, Bayesian statistics, test equating, cognitive diagnostic models, and multistage adaptive testing.

Umeå, Sweden — Marie Wiberg
Champaign, IL, USA — Steven Culpepper
Leuven, Belgium — Rianne Janssen
Santiago, Chile — Jorge González
Amsterdam, The Netherlands — Dylan Molenaar

Contents

Optimal Scores as an Alternative to Sum Scores 1
Marie Wiberg, James O. Ramsay and Juan Li

Disentangling Treatment and Placebo Effects in Randomized Experiments Using Principal Stratification—An Introduction 11
Reagan Mozer, Rob Kessels and Donald B. Rubin

Some Measures of the Amount of Adaptation for Computerized Adaptive Tests 25
Mark D. Reckase, Unhee Ju and Sewon Kim

Investigating the Constrained-Weighted Item Selection Methods for CD-CAT 41
Ya-Hui Su

Modeling Accidental Mistakes in Multistage Testing: A Simulation Study 55
Thales A. M. Ricarte, Mariana Cúri and Alina A. von Davier

On the Usefulness of Interrater Reliability Coefficients 67
Debby ten Hove, Terrence D. Jorgensen and L. Andries van der Ark

An Evaluation of Rater Agreement Indices Using Generalizability Theory 77
Dongmei Li, Qing Yi and Benjamin Andrews

How to Select the Bandwidth in Kernel Equating—An Evaluation of Five Different Methods 91
Gabriel Wallin, Jenny Häggström and Marie Wiberg

Evaluating Equating Transformations from Different Frameworks 101
Waldir Leôncio and Marie Wiberg

An Alternative View on the NEAT Design in Test Equating 111
Jorge González and Ernesto San Martín

Simultaneous Equating of Multiple Forms 121
Michela Battauz

Incorporating Information Functions in IRT Scaling 131
Alexander Weissman

Reducing Conditional Error Variance Differences in IRT Scaling 147
Tammy J. Trierweiler, Charles Lewis and Robert L. Smith

An IRT Analysis of the Growth Mindset Scale 163
Brooke Midkiff, Michelle Langer, Cynthia Demetriou and A. T. Panter

Considering Local Dependencies: Person Parameter Estimation for IRT Models of Forced-Choice Data 175
Safir Yousfi

Elimination Scoring Versus Correction for Guessing: A Simulation Study 183
Qian Wu, Tinne De Laet and Rianne Janssen

Three-Way Generalized Structured Component Analysis 195
Ji Yeh Choi, Seungmi Yang, Arthur Tenenhaus and Heungsun Hwang

Combining Factors from Different Factor Analyses Based on Factor Congruence 211
Anikó Lovik, Vahid Nassiri, Geert Verbeke and Geert Molenberghs

On the Bias in Eigenvalues of Sample Covariance Matrix 221
Kentaro Hayashi, Ke-Hai Yuan and Lu Liang

Using Product Indicators in Restricted Factor Analysis Models to Detect Nonuniform Measurement Bias 235
Laura Kolbe and Terrence D. Jorgensen

Polychoric Correlations for Ordered Categories Using the EM Algorithm 247
Kenpei Shiina, Takashi Ueda and Saori Kubo

A Structural Equation Modeling Approach to Canonical Correlation Analysis 261
Zhenqiu (Laura) Lu and Fei Gu

Dealing with Person Differential Item Functioning in Social-Emotional Skill Assessment Using Anchoring Vignettes 275
Ricardo Primi, Daniel Santos, Oliver P. John, Filip De Fruyt and Nelson Hauck-Filho

Random Permutation Tests of Nonuniform Differential Item Functioning in Multigroup Item Factor Analysis 287
Benjamin A. Kite, Terrence D. Jorgensen and Po-Yi Chen

Using Credible Intervals to Detect Differential Item Functioning in IRT Models 297
Ya-Hui Su, Joyce Chang and Henghsiu Tsai

Bayesian Network for Modeling Uncertainty in Attribute Hierarchy 305
Lihong Song, Wenyi Wang, Haiqi Dai and Shuliang Ding

A Cognitive Diagnosis Method Based on Mahalanobis Distance 319
Jianhua Xiong, Fen Luo, Shuliang Ding and Huiqiong Duan

An Joint Maximum Likelihood Estimation Approach to Cognitive Diagnosis Models 335
Youn Seon Lim and Fritz Drasgow

An Exploratory Discrete Factor Loading Method for Q-Matrix Specification in Cognitive Diagnostic Models 351
Wenyi Wang, Lihong Song and Shuliang Ding

Identifiability of the Latent Attribute Space and Conditions of Q-Matrix Completeness for Attribute Hierarchy Models 363
Hans-Friedrich Köhn and Chia-Yi Chiu

Different Expressions of a Knowledge State and Their Applications 377
Shuliang Ding, Fen Luo, Wenyi Wang, Jianhua Xiong, Heiqiong Duan and Lihong Song

Accuracy and Reliability of Autoregressive Parameter Estimates: A Comparison Between Person-Specific and Multilevel Modeling Approaches 385
Siwei Liu

A Two-Factor State Theory 395
John Tisak, Guido Alessandri and Marie S. Tisak

SPARK: A New Clustering Algorithm for Obtaining Sparse and Interpretable Centroids 407
Naoto Yamashita and Kohei Adachi

Optimal Scores as an Alternative to Sum Scores

Marie Wiberg, James O. Ramsay and Juan Li

Abstract This paper discusses the use of optimal scores as an alternative to sum scores and expected sum scores when analyzing test data. Optimal scores are built on nonparametric methods and use the interaction between the test takers' responses on each item and the impact of the corresponding items on the estimate of their performance. Both theoretical arguments for optimal score as well as arguments built upon simulation results are given. The paper claims that in order to achieve the same accuracy in terms of mean squared error and root mean squared error, an optimally scored test needs substantially fewer items than a sum scored test. The top-performing test takers and the bottom 5% test takers are by far the groups that benefit most from using optimal scores.

Keywords Optimal scoring · Item impact · Sum scores · Expected sum scores

1 Introduction

Test scores must estimate the abilities of the test takers in a manner that is both accurate and unbiased, since they are used in many settings to make decisions about test takers. Sum scores (or number correct scores) have in the past been a common test score choice as they are easy for test takers to interpret and are easy to compute. Scores built on parametric item response theory (IRT; see Lord 1980; Birnbaum 1968) have also been used, although almost exclusively by test constructors,

M. Wiberg (✉)
Department of Statistics, USBE, Umeå University, 901 87 Umeå, Sweden
e-mail: marie.wiberg@umu.se

J. O. Ramsay
Department of Psychology, McGill University, Montreal, Canada
e-mail: james.ramsay@mcgill.ca; ramsay@psych.mcgill.ca

J. Li
Department of Mathematics and Statistics, McGill University, Montreal, Canada
e-mail: juan.li3@mcgill.ca

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_1

since test takers usually find it hard to understand the meaning of the parametric IRT scale scores, which may take any value on the real line. Test takers tend not to be convinced that a score of zero represents average performance. A further problem is that commonly not all items are satisfactorily modeled with parametric IRT models, even in large-scale tests that have been carefully developed.

A choice other than using parametric IRT models is to use nonparametric methods to estimate test takers' ability and the item characteristic curves (ICC). Nonparametric IRT has been used in several studies in the past. Mokken (1997) examined nonparametric estimation and how it worked in connection to monotonicity. Ramsay (1991, 1997) proposed ICC estimation using kernel smoothing over quantiles of the Gaussian distribution. This technique gave fast and reasonably accurate ICC estimation, and was implemented in the computer program TestGraf. Rossi et al. (2002) and Ramsay and Silverman (2002) used the expectation-maximization (EM) algorithm to optimize the penalized marginal likelihood, and the estimates came close to the three-parameter logistic IRT model as the smoothing penalty was increased. Ramsay and Silverman (2005) proposed a nonparametric method for not strictly monotonic curve estimates. Woods and Thissen (2006) and Woods (2006) proposed a method for simultaneously estimating item parameters using a spline-based approximation to the ability distribution. Lee (2007) made a comparison of a number of nonparametric approaches.

As yet another alternative approach to test scoring, this paper will focus on optimal scoring. This method was proposed by Ramsay and Wiberg (2017a) and practical concerns were discussed in Ramsay and Wiberg (2017b). The basic idea behind optimal scoring is to use the interaction between the test takers' responses on each item and the impact of the corresponding items on the estimate of their performance by letting high-slope items be more influential than low-slope items when calculating the test scores. Optimal scoring differs substantially from previous nonparametric approaches in several important ways. First, it uses a faster and more sophisticated approach than the EM algorithm. Second, it uses spline basis expansions over non-negative closed intervals to facilitate the interpretation of the test scores for the test takers. A featured shared with the other nonparametric methods is that it succeeds to get well-fitting ICC's when parametric IRT models fail to give a good fit. The overall aim of this paper is to discuss the nonparametric IRT based optimal scores as a good alternative to sum scores and expected sum scores and to illustrate this with real and simulated test data. This paper also differs from Ramsay and Wiberg (2017b) by extending the comparison to include expected sum scores.

The next section describes the quantitative skill test used as an illustration, followed by a third section where three different test scores are defined. The fourth section contains a description on how to estimate the ICC's with optimal scoring. In the fifth section a comparison between sum scores, expected sum scores and optimal scores are given. The paper ends with a short discussion, which includes some concluding remarks.

2 A College Admission Test and Its Empirical Test Distribution

The data used in this paper come from an administration of the Swedish Scholastic Assessment Test (SweSAT), which is a binary scored multiple-choice college admissions test. The SweSAT contains a verbal and a quantitative parts, each containing 80 items. Sum scores are routinely used in the SweSAT, although the obtained scores are equated to scaled scores, which are comparable over test administrations and these scaled scores are used by test takers in their college applications. A sample of 30,000 test takers who took the quantitative part of the SweSAT is used throughout the paper and the empirical distribution of the sum scores is displayed in Fig. 1. From this figure, we can draw the conclusion that a majority of the test takers found the SweSAT difficult, with a median score of 35, a lowest score of 4 and no test taker with a perfect score. In Fig. 1 we have added a smooth function of the distribution, which was constructed from a B-spline expansion of the log density (Ramsay et al. 2009), since the empirical distribution of the sum scores did not resemble any of the common parametric densities.

Note that in general the distribution of the θ estimates can be transformed whether or not a parametric or nonparametric IRT is used. Suppose we have a one-to-one increasing and smooth transformation $\varphi = h(\theta)$, then there exists an alternative item response function $P_i^*(\varphi)$, so that $P_i^*(\varphi) = P_i(\theta)$. Thus, we can transform any specified distribution of θ into an alternative distribution of φ. For example, we transform from the whole real line into a closed interval such as $[0, n]$ by defining $\varphi = n/(1 + e^{-\theta})$.

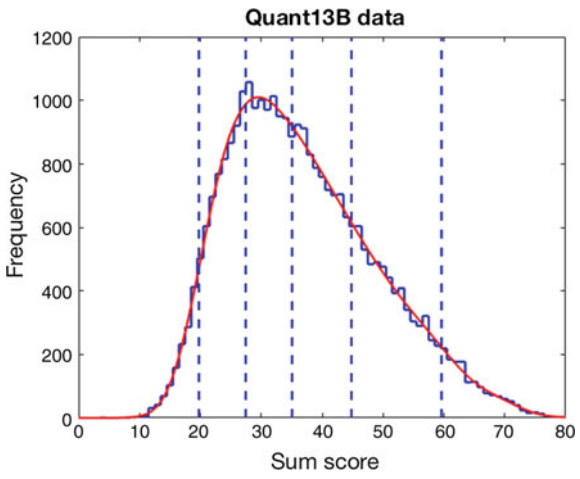

Fig. 1 The empirical distribution of sum scores. The blue histogram indicates the number of test takers within each score range, the red line indicates the smooth density function, and the blue dotted lines are 5, 25, 50, 75 and 95% quintile lines respectively

3 Three Test Scores

Let S_j denote the sum score of test taker j $(j = 1, \ldots, N)$ and define it as the number of correctly answered binary items. Let $P_i(\theta_j)$ be the probability that a test taker with ability level θ_j answered item i $(i = 1, \ldots, n)$ correctly. The expected sum scores are defined as

$$E_j = \sum_i^n P_i(\theta_j). \tag{1}$$

Note, a commonly used expected score uses parametric IRT to model $P_i(\theta_j)$.

To estimate optimal scores O_j (Ramsay and Wiberg 2017a) we focus on estimating the more convenient log-odds function

$$W_i(\theta) = \log\left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right). \tag{2}$$

To estimate $W_i(\theta)$ we can use B-spline basis function expansions

$$W_i(\theta) = \sum_k^K \gamma_{ik}\psi_{ik}(\theta), \tag{3}$$

where for each item i, γ_{ik} is the coefficient of the basis function, $\psi_{ik}(\theta) = B_k(\theta|\xi, M)$ is the B-spline basis function, ξ is a knot sequence, K is the number of spline functions and M is the order of the spline. The advantage of this approach is that B-spline basis functions are easily expanded in dimensionality and they give stable and fast computations.

The left panel of Fig. 2 contains the P_i estimates of the 80 item response functions and the right panel of Fig. 2 shows the W_i estimates of the SweSAT data. From Fig. 2 we learn that items vary in shape of their ICC and their corresponding log-odds functions W_i. Some items are very difficult, other items have low discrimination. If U_{ij} is test taker j's response (0/1) to item i and if either $P_i(\theta)$ or its counterpart $W_i(\theta)$ are either known or we can condition on estimates on them, then the left hand side of

$$\sum_i^n \left[U_{ij} - P_i(\theta)\right]\frac{\mathrm{d}W_i}{\mathrm{d}\theta} = 0 \tag{4}$$

is the derivative of the negative log likelihood

$$-\log L(\theta_j) = -\sum_i^n \left[U_{ji}W_i(\theta_j) - \log(1 + \exp(W_i(\theta_j)))\right].$$

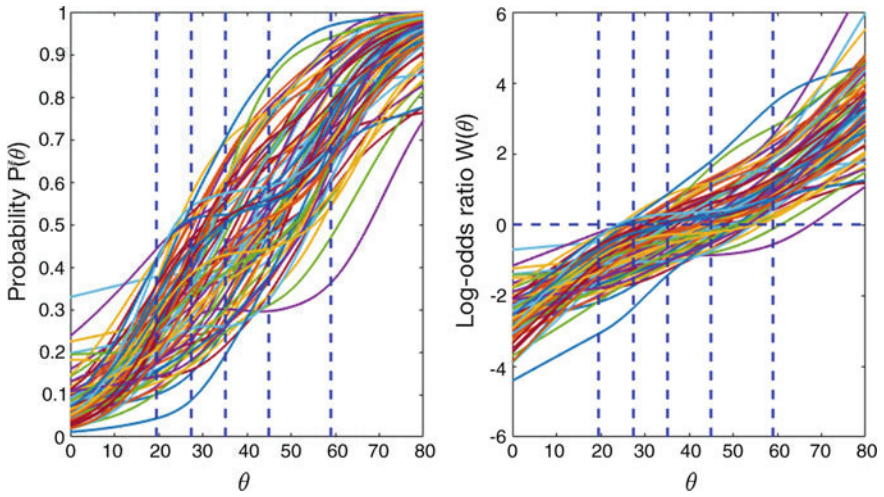

Fig. 2 The left panel displays the $P_i(\theta)$ curves for each item i estimated over the closed interval [0, 80] and the right panel displays the estimated log-odds functions W_i for the SweSAT. The vertical dashed lines are the 5, 25, 50, 75 and 95% quintiles of the empirical distribution of the sum scores

with respect to θ, and the right hand side is zero for its optimal value. Equation 4 is interesting in several aspects. The slopes of the log-odds functions $W_i(\theta)$ at the optimal θ weight the residuals $U_{ij} - P_i(\theta)$. The optimal scores thus correspond to the ability that minimizes the difference between the answers and their probabilities in which each item is weighted by its impact (or sensitivity) value. In practice, this means that high-slope items are mainly influencing the differences in scores among the test takers. The most useful items for assessing test takers at level θ have higher slopes of W_i at that location, while items having nearly flat W_i are down-weighted, which would be the case for easy items being given to high-level θ test takers. We will refer to the interaction between item weights and item performance in the weighting as the *item impact function*. The item impact curves $(dW_i/d\theta)$, corresponding to the curves in Fig. 2, are shown in Fig. 3. From Fig. 3, it is obvious that items have various weights or performances for a certain ability level θ, and one particular item's performance will change at different θ. Summing up, the optimal scoring algorithm is focused on the items that are most informative as reflected by the size of the item impact function $dW_i/d\theta$, which yields the amount of information provided by answers to item i.

$$dW_i/d\theta$$

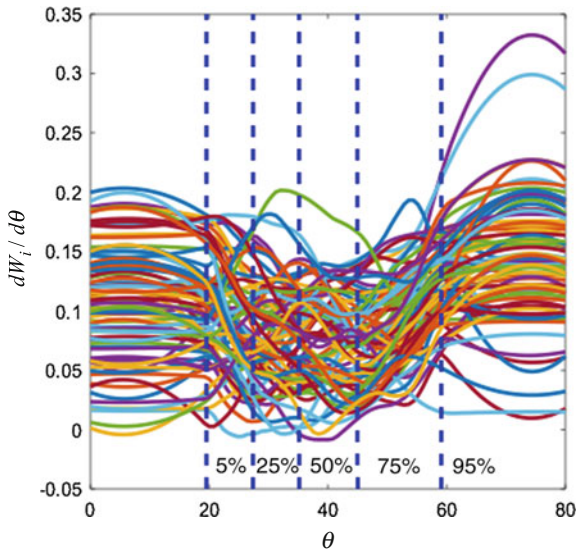

Fig. 3 The item impact curves, $dW_i/d\theta$, that provide the optimal weighting of item scores. The vertical dashed lines are the 5, 25, 50, 75 and 95% quintiles of the empirical distribution of the sum scores

4 Estimating Nonparametric ICC's

An efficient nonparametric procedure for joint estimation of the n functions W_i and the knowledge states θ_j was described in Ramsay and Wiberg (2017a). In their procedure they use parameter cascading (PC), which is a generalization of profiling that is computationally faster than marginalization over θ. Let θ_j be represented by smooth functions $\theta_j(W_1, \ldots, W_n)$. The PC optimizations performed are initialized by a fast data smoothing approach to estimate the W_i as described in Ramsay (1991). PC is a compound optimization procedure in which an inner optimization $(H(\theta|\gamma))$ of a penalized log likelihood function with respect to the θ_j is updated, each time an *outer* optimization $(F(\gamma))$ adjusts the coefficients of the B-spline basis function expansions of the W_i. In PC, the gradient plays a crucial role in the outer optimization through the implicit theorem such that an efficient search is made possible. Details of how to perform PC are provided in Ramsay and Wiberg (2017a). We emphasize that PC is different from using alternating optimization (AO) as for example the EM-algorithm. Instead of a compound optimization as in PC, AO switches between optimizing one criterion F with respect to some γ keeping θ fixed, and optimizing another criterion H with respect to θ keeping γ fixed.

5 Optimal Scores in Comparison with Sum Scores and Expected Sum Scores

5.1 Simulation Study

As a first step, the difference between optimal scores and sum scores as well as optimal scores and expected sum scores were calculated for the SweSAT data. In order to further examine the difference between sum scores, expected sum scores and optimal scores we used simulations from the populations defined by the W_i curves and the θ_j's estimated from the data. The first obstacle was how to handle the problem of identifying the distribution of θ. To make a fair comparison with the sum scores we simulated test data using a smooth estimate of the density of the sum scores based on the SweSAT empirical distribution shown in Fig. 1. As we had access to a sample of 30,000 test takers the W_i have been pre-calibrated and were considered to be known (and can be seen in Fig. 2) and thus we only simulated the test takers' responses. Root mean squared error (RMSE) of θ was used to assess recovery. The analysis was performed using PC for optimization. The 81 sum score values were used as fixed values of θ and we simulated 1000 test takers responses. Sum scores, expected sum scores and optimal scores were averaged across 1000 simulated samples for each value of θ. The average bias of θ for each test score was also used to evaluate the different test scores.

5.2 Results of the Simulation Study

The difference between optimal scores and sum scores as well as optimal scores and expected sum scores are displayed in Fig. 4 for the SweSAT data. The left panel in Fig. 4 shows a large increase in test scores for high-performing test takers if they would get an optimal score instead of a sum score. The expected sum score in the right panel is overall more similar to the sum scores than optimal scores, but the really top achievers among the test takers get penalized with an expected sum score. The sum score/optimal score and sum score/expected score differences can be as large as the size of 20% for some of the scores (for sum scores around 40, the difference can be ± 8).

In Fig. 5 the empirical distributions are displayed in the left panel and the average RMSE and bias are shown in the right panel for each value of θ. The empirical distributions for the three different scores only differ slightly. For the mid 90% of the test takers the bias is close to zero regardless of the test scoring method. But low-performing test takers get higher sum scores than the corresponding θ values used to generate the data, while at the same time, high-performing test takers lose about five items using sum scores. For the 5% top- and bottom-performing test takers the bias and RMSE for sum scoring is substantial. For the mid 90% of the test takers the RMSE is larger for the sum scores than for optimal or expected sum

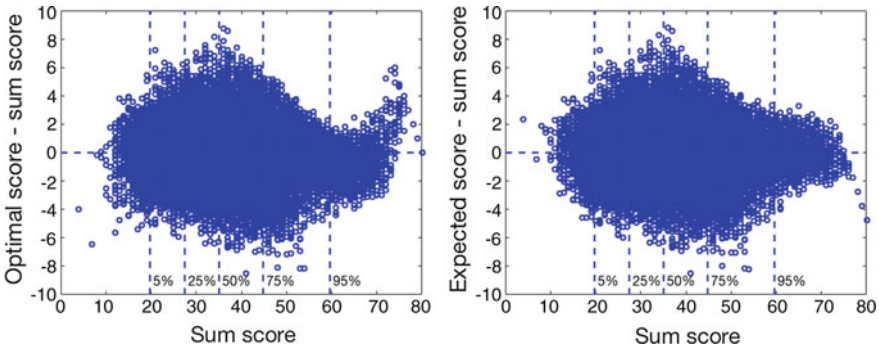

Fig. 4 The left panel displays optimal scores minus the sum scores plotted against sum scores and the right panel displays the expected sum scores minus the sum scores for the SweSAT

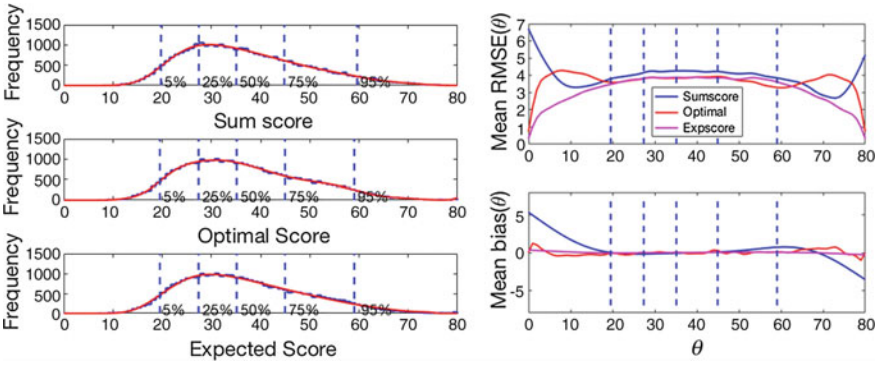

Fig. 5 The left panel displays the empirical distribution of sum scores, optimal scores and expected sum scores and the right panel displays the average RMSE of θ and average bias of θ for the three test scores. The vertical dashed lines are the quintiles of the empirical distribution of the maximum likelihood estimates

scores. From the simulations, the optimal score RMSE was on average 6.8% lower than the sum score RMSE, which corresponds to a mean squared error (MSE) of 14%. Because the MSE declines in proportion to $1/n$, we see that the sum-scored SweSAT would have to be 11 items longer than an optimally scored test in order to achieve the same average accuracy. Note that the expected sum scores have the lowest RMSE and bias at each score values, but they are expected scores and are thus not built on the observed scores as sum scores and optimal scores are. The results from the simulations for optimal scores in comparison to sum scores are in line with the results in Ramsay and Wiberg (2017a), who used simulations based on three different tests and compared optimal scores and sum scores.

6 Discussion

This paper used a large sample from a college admissions test in order to discuss optimal scores in comparison to sum scores and expected sum scores. A closed interval in terms of the range of the sum scores was used in order to model student performance differences. This choice facilitates comparisons with the sum scores and the expected sum scores in terms of bias and RMSE, and also allows for understandable interpretations for the test takers.

The simulation study indicated that the expected sum scores and optimal scores should be preferred over sum scores as their average bias and average RMSE were lower than for the corresponding sum scores. The improvement in terms of RMSE was about 6% for 90% of the test takers. Even though the expected sum scores had the lowest bias and RMSE we cannot recommend it in general as it measures something else than the optimal scores, i.e. it is an expected score instead of an observed score. It was mainly included here for sake of completeness and as expected sum scores are sometimes used in test analysis. The largest problem with sum scores is the substantial negative bias for high-performing test takers and the positive bias for low-performing test takers. The substantial improvement is important, especially in high-stakes test as the SweSAT. To get an improvement of 6% could be the difference of being accepted into the university program of one's choice or not. The improvement found in the well-designed SweSAT lead us to expect a larger benefit if we have less well-designed tests, as for example those given in classrooms. We are not stating that sum scores should never be used as they might be useful in some situations. However if we put some effort into explaining how optimal scores work it may be beneficial for both test constructors and test takers as they contain more information.

In the future it is important to continue examining the performance of optimal scoring, especially against parametric IRT as that is used all over the world by test constructors. As additional information about test takers in terms of covariates are regularly gathered when large-scale tests are given it should be interesting to examine optimal scoring with covariates as it has been used successfully in other test areas as for example test equating (Bränberg and Wiberg 2011; Wallin and Wiberg 2017; Wiberg and Bränberg 2015). Other interesting future directions include the use of optimal scoring with polytomous scored items and multidimensional tests. In order to spread the usage of optimal scoring it is crucial to develop an easy to use software. Currently the authors are developing a new version of TestGraf (Ramsay 2000) which will incorporate all the important features of optimal scoring. In summary, optimal scoring provides a number of interesting opportunities as it is built on efficient and advanced statistical methodology and technology. We need to stop the waste of valuable information and give our top-performing test takers the score they earn.

Acknowledgements This research was funded by the Swedish Research Council grant 2014-578.

References

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 tests* (pp. 395–479). Reading, MA: Addison-Wesley.

Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48,* 419–440.

Lee, Y.-S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 37,* 121–134.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York, NY: Springer.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Ramsay, J. O. (1997). A functional approach to modeling test data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York, NY: Springer.

Ramsay, J. O. (2000). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*. [Computer software and manual]. Department of Psychology, McGill University, Montreal Canada.

Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis in R and Matlab*. New York, NY: Springer.

Ramsay, J. O., & Silverman, B. W. (2002). Functional models for test items. In *Applied functional data analysis*. New York, NY: Springer.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer.

Ramsay, J. O., & Wiberg, M. (2017a). A strategy for replacing sum scores. *Journal of Educational and Behavioral Statistics, 42,* 282–307.

Ramsay, J. O., & Wiberg, M. (2017b). Breaking through the sum score barrier. In *Paper presented at the International Meeting of the Psychometric Society, July 11–15, Asheville, NC* (pp. 151–158).

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of the Behavioral and Educational Sciences, 27,* 291–317.

Wallin, G., & Wiberg, M. (2017). Non-equivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang. (Eds.), *Quantitative psychology—81st annual meeting of the psychometric society, Asheville, North Carolina, 2016* (pp. 309–320). New York: Springer.

Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39,* 349–361.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods, 11,* 253–270.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71,* 281–301.

Disentangling Treatment and Placebo Effects in Randomized Experiments Using Principal Stratification—An Introduction

Reagan Mozer, Rob Kessels and Donald B. Rubin

Abstract Although randomized controlled trials (RCTs) are generally considered the gold standard for estimating causal effects, for example of pharmaceutical treatments, the valid analysis of RCTs is more complicated with human units than with plants and other such objects. One potential complication that arises with human subjects is the possible existence of placebo effects in RCTs with placebo controls, where a treatment, suppose a new drug, is compared to a placebo, and for approval, the treatment must demonstrate better outcomes than the placebo. In such trials, the causal estimand of interest is the medical effect of the drug compared to placebo. But in practice, when a drug is prescribed by a doctor and the patient is aware of the prescription received, the patient can be expected to receive both a placebo effect and the active effect of the drug. An important issue for practice concerns how to disentangle the medical effect of the drug from the placebo effect of being treated using data arising in a placebo-controlled RCT. Our proposal uses principal stratification as the key statistical tool. The method is applied to initial data from an actual experiment to illustrate important ideas.

Keywords Causal inference · Placebo effects · Principal stratification

1 Introduction

Placebo-controlled, blinded randomized controlled trials (RCTs) are the standard for approving pharmaceuticals to be given to human beings in the United States, European Union, and much of the world. In fact, agencies such as the U.S. Food and

R. Mozer (✉) · D. B. Rubin
Harvard University Department of Statistics, Cambridge, MA 02138, USA
e-mail: rrose@g.harvard.edu

D. B. Rubin
e-mail: rubin@stat.harvard.edu

R. Kessels
Emotional Brain, B.V., Almere, The Netherlands
e-mail: r.kessels@emotionalbrain.nl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_2

Drug Administration (FDA) and the European Medicines Agency (EMA) usually require evidence from such trials that the drugs being proposed are safe and effective. It is a widely accepted stance in the world of drug development that if a drug is "snake oil", meaning it is ineffective and only appears to work because of presumed expectancy effects, then the producer of the drug should not profit from its sale. It was because of this attitude that placebo controlled, double-blind randomized trials became essentially necessary for the approval of new drugs in the 1960s. That is, for a drug to be considered effective, the active drug (treatment) must be compared to an inactive drug (a placebo), which (to a user) is indistinguishable from the active drug, where assignment to the treatment versus control is random and unknown to the experimental units until the completion of the experiment; here the units are said to be "blinded" to the actual assignment. If assignment is unknown to both the experimental units and the experimenter, the experiment is considered "double-blind".

Although randomized experiments have been used for nearly a century, for decades they were only used with unconscious units, such as plants, animals, or industrial objects, none of which presumably could be influenced by the knowledge that they were objects of experimentation. Historically, it has been recognized that humans are different and can be influenced by the knowledge that they are part of an active experiment. In some cases, that knowledge alone has been shown to influence participants behavior, as with the well-known "Hawthorne effect" (Landsberger 1958), where awareness of participation in a study influences outcomes. In other examples, the knowledge that some individuals would receive an active drug with a particular anticipated effect creates the expectation among all experimental units that this anticipated effect will be achieved among all participants, a version of so-called "expectancy effects (Rosenthal and Fode 1963; Rosenthal and Jacobson 1966). Thus, a number of complications may arise when analyzing data from randomized experiments with human subjects when the conduct of the experiment itself influences participants' outcomes.

2 Motivation

Emotional Brain (EB) is a research company based in the Netherlands that is developing a therapy for improving sexual functioning in women, which they call Lybrido. Lybrido is designated for the treatment of a medical condition in women called Female Sexual Interest/Arousal Disorder (FSIAD). The increase in "satisfying sexual events" (SSEs) per week from baseline (before any drugs, active or placebo, have been received) is the accepted primary outcome of interest, and for approval of Lybrido, by either the FDA or EMA, there must be evidence that the drug is superior to placebo with respect to increase in SSEs from baseline, ΔSSE. As with other psychopathologies, experiments on therapies for treating this condition are believed to suffer from large placebo effects because the anticipation of effects of the drug can have obvious effects on the self-reported number of SSEs.

A variety of small, but expensive, randomized placebo-controlled double-blind trials have been conducted to study the effectiveness of Lybrido (Van Der Made et al.

2009a, b; Poels et al. 2013). In these trials, simple analyses comparing the randomized groups with each other (intention to treat analyses) generally show significant positive effects for Lybrido relative to placebo, but the large placebo effects (that is, large increases in SSEs observed in all groups) complicate the interpretation and implication of the results.

The desire to disentangle the active effects of Lybrido from its related placebo effects is important for several reasons. First, assume Lybrido has a true effect for some subset of women, but this true effect is masked by highly variable placebo effects; how do we eliminate the noise and so identify that subset of women? This is related to the current hot-topic issue of "personalized medicine", which describes selecting treatments that are tuned to patients characteristics. Another important question concerns what outcomes should be anticipated in actual medical practice, when doctors prescribe a treatment and patients are aware of the prescription they receive. In this setting, patients' outcomes will reflect both placebo effects as well as the medical effects of the active drug. Considering both types of effects may allow prescribing physicians to anticipate better the benefits a patient can expect when using the drug outside of the setting of an RCT.

The objective of this work is to disentangle active drug and placebo effects in RCTs, such as those with Lybrido. Previous attempts to address this issue using existing methods are summarized in Kessels et al. (2017), and, though some have interesting ideas, none are statistically fully satisfactory. Here we use the statistical tool called Principal Stratification (Frangakis and Rubin 2002) to estimate jointly treatment and placebo effects within the framework of causal inference based on potential outcomes, commonly called the Rubin Causal Model (Holland 1986) for a body of work done in the 1970s (Rubin 1974, 1975, 1978, 1980); a short summary of this perspective is in Imbens and Rubin (2008) and a book on it is Imbens and Rubin (2015).

In principle, we consider the administration of placebo as an intervention, just as the administration of an active drug. The placebo effect is then defined by comparing potential outcomes under assignment to placebo to potential outcomes under no treatment at all. Just as active treatment effects can vary across units, so can placebo effects, which can also vary as a function of patients' individual characteristics. Further, the effects of the active treatment can also vary with respect to characteristics of patients, including their individual placebo effects, which further complicates statistical inference.

3 The Principal Stratification Framework for Joint Estimation of Treatment and Placebo Effects

3.1 Notation

Consider an RCT with N subjects, indexed by $i = 1, \ldots, N$. Subject i is assigned treatment Z_i, which equals 1 for subjects assigned and receiving active treatment

and equals 0 for subjects assigned and receiving placebo. Throughout, we assume full compliance with assignment. Interest focuses on the effect of treatment ($Z_i = 1$) compared to placebo ($Z_i = 0$) on an outcome variable, defined in terms of change from a baseline measurement Y_{i0}. For each subject we may also observe a vector of p pre-treatment covariates $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ where X is the $N \times p$ matrix of covariates for all subjects. The outcome variable takes the value $Y_i(1)$ if subject i is assigned treatment and $Y_i(0)$ if subject i is assigned placebo. The "fundamental problem facing causal inference" (Rubin 1978) is that we cannot observe both potential outcomes $Y_i(0)$ and $Y_i(1)$, but rather $Y_i^{obs} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$, the observed outcome for subject i. Additionally, we consider a third potential outcome, $Y_i(-1)$, which is defined but never observed for any unit in the situation we consider and represents the outcome that would be observed if unit i is neither assigned nor receives either treatment or placebo and is aware of this. We then define the causal effects of interest by differences in potential outcomes, where $Y_i(0) - Y_i(-1)$ is the "placebo effect" for unit i and $Y_i(1) - Y_i(0)$ is the "medical effect" of active treatment for unit i, or for descriptive simplicity, the treatment effect.

3.2 General Modeling Strategy

Because we believe that effect of the active treatment can depend on both individual characteristics of the patient (i.e., covariate values X_i) and the magnitude of the patient's response to placebo, $Y_i(0)$, our approach is a version of the one used in Jin and Rubin (2008), which deals with "extended partial compliance", a special case of principal stratification that defines principal strata based on continuous measures of how each patient would comply with their assignment under both treatment and control.

Here, we view patients' response to placebo as roughly analagous to compliance status under active treatment, and following Jin and Rubin (2008), we define continuous principal strata according to this potential outcome, which is only partially revealed (i.e., revealed for those patients assigned placebo), but is missing for those patients assigned the active treatment. Causal effects of the active treatment versus placebo are then defined conditional on the observed covariates and the potential outcomes under placebo. Regression models (typically not linear) are used for the joint conditional distribution of potential outcomes given covariates, specified by the distribution of the placebo potential outcome (given covariates) and the conditional distribution of the potential outcome under treatment given the potential outcome under placebo (and, of course, the covariates). This is explained in greater detail in Sect. 4. For analysis, we use Bayesian models with proper prior distributions and employ Markov Chain Monte Carlo (MCMC) methods, which are only outlined in this paper. Under this framework, missing potential outcomes are multiply imputed to obtain a large number of completed data sets, from each of which, all causal estimands, including individual causal effects, can be computed. Aggregates of the

estimated individual effects across the multiply imputed data sets then approximate the posterior distributions of interest.

3.3 Assumptions

Throughout this article, we assume the Stable Unit Treatment Value Assumption (SUTVA) (Rubin 1980), which requires that there is no interference between units (that is, treatment assignment for an individual unit has no effect on the potential outcomes of other units) and that there are no hidden versions of treatments. We also assume ignorable treatment assignment (Rubin 1978), which requires that the treatment assignment is known to be a probabilistic function of observed values and is true by design in randomized experiments. Next, we assume that the potential outcomes under no treatment, $Y_i(-1)$, defined as the change in the outcome from its measurement at baseline, is zero for all units (i.e., $Y_i(-1) = 0$ for all $i = 1, \ldots, N$). This important assumption implies that the outcome that would be observed for each unit if they were given neither the active treatment nor placebo, and are aware that they are receiving neither, will be exactly equal to the value of that unit's outcome at baseline; assessing this assumption would require a design with such an assignment (i.e., an assignment with instructions to take nothing and continue to be followed up with measurements as if the patient had been assigned either active treatment or placebo).

All other assumptions are extensions of the classical assumptions utilized in problems involving principal stratification. In particular, we assume positive side-effect monotonicity on the primary outcome for both treatment and placebo, that is, $Y_i(1) \geq 0$ and $Y_i(0) \geq 0$ for all i, which implies that neither the treatment nor the placebo are harmful to any units, in the sense that an individual will not experience a decline in their outcome (measured as change from baseline) as a result of either intervention.

We also assume additivity of the treatment and placebo effects on some scale. This is analagous to the perfect blind assumption commonly made in causal inference, which requires that, upon receipt, the active drug is indistinguishable from the placebo except for its active effect. Under this assumption, for a unit assigned to treatment, the portion (on some scale) of the observed outcome that is attributable to the placebo effect is exactly equal to the placebo effect that would be observed if that unit had been assigned placebo. Thus, the potential outcome when assigned treatment can be viewed as the sum of the "placebo effect" and some "extra" effect achieved under treatment that is attributable to the active drug, which we call the treatment effect.

Together, these assumptions also imply that for every patient, the total response that would be observed when assigned treatment is greater than or equal to the response that would be observed when assigned placebo (i.e., $Y_i(1) \geq Y_i(0)$ for all i).

4 Model and Computation

4.1 The General Model with No Covariates

We begin by considering the simplest case of an RCT with no covariates. We first specify a distribution for the potential outcomes under control, $Y(0)$, conditional on some global parameter θ:

$$Y_i(0)|\theta \sim \mathcal{L}^{(0)}, \quad Y_i(0) \geq 0 \ \text{ for all } \ i, \theta \tag{1}$$

where $\mathcal{L}^{(0)}$ denotes the probability law for $Y_i(0)$, governed by some parameters, which are functions of the global parameter θ. We then specify a distribution for the potential outcomes under treatment, $Y_i(1)$, conditional on the potential outcomes under control, $Y_i(0)$ and θ as

$$Y_i(1)|Y_i(0), \theta \sim \mathcal{L}^{(1)}, \quad Y_i(1) \geq 0 \ \text{ for all } \ i, \theta \tag{2}$$

where

$$E[Y_i(1)|Y_i(0), \theta] = Y_i(0) + f(Y_i(0)). \tag{3}$$

Here, $\mathcal{L}^{(1)}$ is another probability law, and f is an arbitrary function that generally defines heterogeneous treatment effects across units as a function of potential outcomes under placebo. Under this formulation, $f(Y_i(0))$ is the treatment effect for unit i. By the assumptions stated in Sect. 3.3, $f(\cdot)$ must be chosen such that $f(Y_i(0)) \geq 0$ for all i and $Y_i(0) + f(Y_i(0))$ is monotonically non-decreasing in $Y_i(0)$, which defines a positive, monotonically non-decreasing curve, analagous to a dose-response curve, which captures the expected effect of assignment to treatment versus assignment to placebo for each possible value of placebo response.

For example, consider the specification of $f(\cdot)$ as the polynomial $f(x) = a_0 + a_1 x + a_2 x^2$, where a_0, a_1 and a_2 are constrained such that $f(x) \geq 0$, and $1 + f'(x) = 1 + a_1 + 2a_2 x \geq 0$ for all x (thereby satisfying the monotonicity constraint). Under this specification, the parameters of interest are (a_0, a_1, a_2), where the intercept parameter a_0 is a common treatment effect across all subjects, including those who have zero response to placebo, and the parameters a_1 and a_2 capture how treatment effects vary linearly and quadratically, respectively, with the magnitude of placebo response. Figure 1 illustrates such a specification, where the left plot displays the expected medical effect of the active drug as a function of placebo response, which is relevant for drug approval, and the right plot displays the overall expected response to being assigned the active drug and taking it as a function of placebo response, which is relevant for anticipating the benefits a patient can expect when using the drug as prescribed by a doctor.

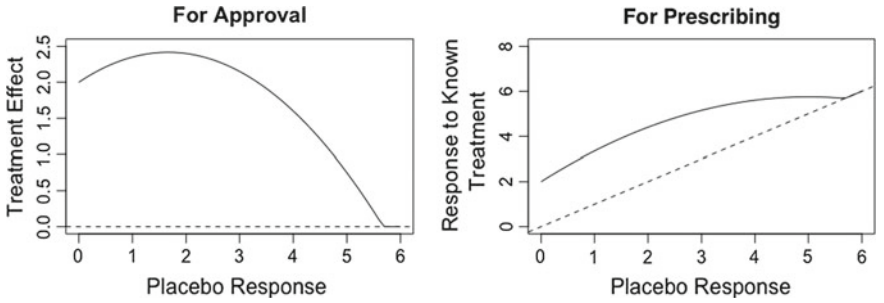

Fig. 1 Two illustrations of a possible quadratic specification of f

4.2 Computation

Under the general formulation above, the complete-data likelihood for the data $Y = (Y(0), Y(1))$ (meaning the likelihood if both $Y_i(1)$ and $Y_i(0)$ were observed for all units) is:

$$p(Y|\theta, Z) = \prod_i p(Y_i(1), Y_i(0)|\theta) = \prod_i p(Y_i(1)|Y_i(0), \theta)p(Y_i(0)|\theta). \tag{4}$$

For Bayesian inference, with prior distribution $p(\theta)$ on θ, the posterior distribution of θ given the complete data Y is then:

$$p(\theta|Y, Z) \propto p(\theta)p(Y, Z|\theta) = p(\theta)p(Y|\theta, Z), \tag{5}$$

where the equality follows from the randomization of Z. Posterior inference on θ can then be done using straightforward application of MCMC techniques, such as the Gibbs sampler (Geman and Geman 1984; Gelman et al. 2014). For example, in each iteration of the Gibbs sampler, we draw the missing potential outcomes Y^{mis} given the observed data Y^{obs} and the current draw of the parameter θ:

$$\begin{aligned} p(Y^{mis}|Y^{obs}, \theta, Z) &= \prod_{i\in\{Z_i=0\}} p(Y_i(1)|Y_i(0) = Y_i^{obs}, \theta) \\ &\quad \times \prod_{i\in\{Z_i=1\}} p(Y_i(0)|Y_i(1) = Y_i^{obs}, \theta) \\ &= \prod_{i\in\{Z_i=0\}} p(Y_i(1)|Y_i(0) = Y_i^{obs}, \theta) \\ &\quad \times \prod_{i\in\{Z_i=1\}} p(Y_i(1)|Y_i(0) = Y_i^{obs}, \theta)p(Y_i(0) = Y_i^{obs}|\theta) \end{aligned} \tag{6}$$

where the second equality follows from Bayes Rule. We then draw θ given the completed data $Y = (Y^{obs}, Y^{mis})$ using Eqs. 4 and 5, and we continue this process until

convergence in distribution. Depending on the specifications of $\mathcal{L}^{(0)}$ and $\mathcal{L}^{(1)}$, the conditional distribution of Y^{mis} given Y^{obs} and θ, and the conditional distribution of θ given the complete data Y, may not have closed-form solutions that allow us to sample directly values of Y^{mis} or θ. In such situations, Metropolis-Hastings steps can be used to draw approximate samples from the desired conditional distributions in each iteration of the Gibbs sampler. For posterior inference on causal effects of interest, we continue this sampling procedure after approximate convergence, in each iteration drawing the missing potential outcome, $Y_i(0)$ or $Y_i(1)$, for each patient. Thus, in each iteration, we construct a completed dataset consisting of all observed potential outcomes and the imputed missing potential outcomes, and then use this completed data to calculate the implied placebo and treatment effects. Repeating this process over many such simulated datasets produces the approximate posterior distribution for all causal effects of interest. In the same way, posterior samples of θ can provide posterior estimates of the parameters of the function f, which characterizes the relationship between expected response to treatment and expected response to placebo.

Depending on the exact specification of $f(\cdot)$, the likelihood may suffer from problems with multimodality, as is common with many specifications of mixture models, such as this one. In such situations, initialization of the MCMC procedure can have an impact on convergence, and first finding regions of high posterior density (e.g., maximum likelihood estimates—MLEs) for model parameters using a method such as a variant of Expectation Maximization (EM) (Dempster et al. 1977) to inform initial values in the MCMC procedure can help. In cases of extreme multi-modality of the likelihood, one can also specify more restrictive prior distributions on the parameters governing $f(\cdot)$.

4.3 Incorporating Covariates

The model presented in Sect. 4.1 considers a patient's response to placebo as an underlying, psychological, characteristic that exists prior to treatment assignment. By defining heterogeneous treatment effects as a function of this characteristic, we can estimate both the expected effect of assignment to treatment versus assignment to placebo (the medical effect of the active drug) and the expected effect of assignment to placebo versus assignment to neither treatment nor placebo (the placebo effect) for each type of patient, at least under specific assumptions.

When covariates, $X_i = (X_{i1}, \ldots, X_{ip})$, are observed for patients, we can specify the distribution for potential outcomes under control, $Y_i(0)$, conditional on X_i and the global parameter θ as:

$$Y_i(0)|X_i, \theta \sim \mathcal{L}^{(0)} \;\; Y_i(0) \geq 0 \text{ for all } \; i, \theta. \tag{7}$$

We then model the potential outcomes under treatment, $Y_i(1)$, conditional on $Y_i(0), X_i$, and θ as

$$Y_i(1)|Y_i(0), X_i, \theta \sim \mathcal{L}^{(1)} \;\; Y_i(1) \geq 0 \text{ for all } \; i, \theta, \tag{8}$$

where $\mathcal{L}^{(1)}$ is such that

$$E[Y_i(1)|Y_i(0), X_i, \theta] = E[Y_i(0) + f(Y_i(0))|X_i]. \tag{9}$$

In general, we assume that covariate effects on $Y_i(0)$ are conditionally independent of effects on $Y_i(1)$. For example, continuing the example where f is specified using the polynomial $f(x) = a_0 + a_1 x + a_2 x^2$, we might consider linear regression models for covariate effects on both $Y_i(0)$ and $Y_i(1)$:

$$\begin{aligned} Y_i(0)|X_i, \theta &= \beta_0 + X_i\beta + \epsilon_i \\ Y_i(1)|Y_i(0), X_i, \theta &= Y_i(0) + X_i\gamma + a_1 Y_i(0) + a_2 Y_i(0)^2 + \eta_i, \end{aligned} \tag{10}$$

where ϵ_i and η_i are independent residual terms and $\beta, \gamma \in \mathcal{R}^p$ govern covariate effects. Here, we may include an intercept term for the distribution of potential outcomes under control but not for the distribution of potential outcomes under treatment. In this example, posterior inference for θ comprises two standard Bayesian regressions (Gelman et al. 2014).

5 Evaluating Treatment and Placebo Effects of Lybrido on Sexual Function

To illustrate our proposed approach, we return to our motivating example of Lybrido. Data for this example were pooled from two double-blind, placebo-controlled RCTs conducted by EB to investigate the efficacy of Lybrido among patients for whom FSIAD was believed to be caused by insensitivities in the brain to sexual cues. Because the actual results of both studies are under peer review process with an implied embargo, a subset of 67 patients was sampled from these data to be used for illustrative purposes here, 34 randomized to treatment (Lybrido) and 33 randomized to control (placebo).

The primary outcome of interest in this example is the increase from baseline in number of SSEs within a four week period during the study. In this example, no baseline measurements for SSEs are directly observed for any participants in the sample, but implicitly these values are all equal to zero, because the patients in these experiments have FSIAD and therefore suffer from low sexual desire. This likely leads to infrequent SSEs among these patients, which makes the assumed value of zero SSEs at baseline realistic. In addition to the outcome, we observe the age and body mass index (BMI) for each patient at the time of enrollment, as well as 40 other covariates collected via self-report using the Sexual Motivation Questionnaire (SMQ), as described in detail in a subsequent publication.

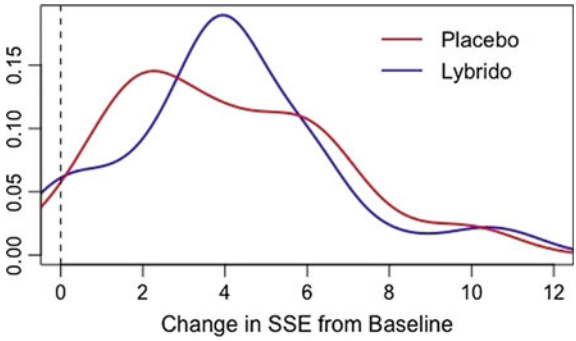

Fig. 2 Kernel density estimates for the distributions of observed potential outcomes in treatment (Lybrido) and control (placebo)

We observe an average of 4.00 SSEs over the four week study period for patients randomized to receive treatment, with a standard deviation of 2.58, and an average of 4.06 SSEs over the four week period for patients randomized to receive placebo, with a standard deviation of 2.58. Kernel density estimates of the distributions of observed potential outcomes in the treatment and control groups are shown in Fig. 2.

Using simple intention to treat (ITT) analysis (Sheiner and Rubin 1995), which compares the means of observed potential outcomes among treated units to those in control, we estimate the ITT effect of assignment to Lybrido to be $4.00 - 4.06 = -0.06$. At first glance, this result suggests that Lybrido has essentially zero effect compared to placebo and might lead to the conclusion that the drug is ineffective as a treatment for FSIAD. However, because both the placebo and treatment groups are observed to have large and highly variable responses (with standard deviations of approximately 2.58 in each group), this finding may instead suggest that any effect of the active drug is simply being masked by large placebo effects and varying treatment effects, which more sophisticated statistical analyses might be able to detect.

5.1 Model Specification

To illustrate our proposed approach on these data, we consider models both with and without the observed covariates. For both models, we specify the function $f(\cdot)$, which relates each patients' treatment effect to their expected potential outcomes under assignment to placebo, using the simple quadratic form $f(x) = a_0 + a_1 x + a_2 x^2$. In the model for $Y_i(1)$ that includes covariates, however, no intercept term is included because $Y_i(1)$ is already centered at $Y_i(0)$. For both models, we assume a truncated normal distribution for placebo response, $Y_i(0)$. With no covariates, this is:

$$Y_i(0)|\theta \sim \mathcal{N}_+(\mu_0, \sigma_0^2), \tag{11}$$

where $\mathcal{N}_+(\mu, \sigma^2)$ denotes a normal density with mean μ and variance σ^2 truncated to the interval $[0, \infty]$. Similarly, we specify a truncated normal distribution for treat-

ment response, $Y_i(1)$, given $Y_i(0)$ as:

$$Y_i(1)|Y_i(0), \theta \sim \mathcal{N}_+(Y_i(0) + f(Y_i(0)), \sigma_1^2). \tag{12}$$

In this illustrative example, we use truncated normal distributions for both placebo response, $Y_i(0)$, and treatment response, $Y_i(1)$, to satisfy the assumption of positive side-effect monotonicity, which requires that both $Y_i(0)$ and $Y_i(1)$ be strictly non-negative for all i. However, other distributions that satisfy this constraint (e.g., Poisson) could be specified for one or both of these variables. In general, we advise researchers implementing this approach in practice to choose appropriate distributions based on domain knowledge about the treatment and population under investigation.

When including covariates, we model $Y_i(0)$ conditional on X_i as:

$$Y_i(0)|X_i, \theta \sim \mathcal{N}_+(\beta_0 + X_i\beta, \sigma_0^2), \tag{13}$$

and model $Y_i(1)$ given $Y_i(0)$ and X_i as:

$$Y_i(1)|Y_i(0), X_i, \theta \sim \mathcal{N}_+(Y_i(0) + f(Y_i(0)) + X_i\gamma, \sigma_1^2). \tag{14}$$

In the model without covariates, the global parameter is $\theta = (a_0, a_1, a_2, \sigma_0^2, \sigma_1^2)$, and with covariates we have $\theta = (a_0, a_1, a_2, \beta_0, \beta, \gamma, \sigma_0^2, \sigma_1^2)$, where β_0 is an intercept term for the regression of response to placebo, $Y_i(0)$, on the covariates X_i, and β and γ are p-dimensional vectors with components for coefficients for the covariate effects on $Y_i(0)$ and $Y_i(1)$. In both models, we use weakly informative prior distributions on all parameters, where each prior distribution is proper and fully specified.

5.2 Results

Results from the models with and without covariates are displayed in Fig. 3. When using the model without covariates, we estimate the function f as $\hat{f}(Y_i(0)) = 0.288 - 0.035Y_i(0) - 0.481Y_i(0)^2$, which suggests that Lybrido has the largest effects on patients that do not respond to placebo ($E[Y_i(1)|Y_i(0) = 0] \approx 0.288$). Further, we see that estimated treatment effects decrease with response to placebo, such that patients who have a placebo response of approximately one or more post-assignment SSEs are expected to have essentially zero treatment effects. That is, big placebo responders do not benefit from receiving the active treatment. The findings are similar when employing the model that incorporates covariates. Using this model, we obtain $\hat{f}(Y_i(0)) = 0.321 + 0.016Y_i(0) - 0.323Y_i(0)^2$ with $E[Y_i(1)|Y_i(0) = 0] = 0.321$. Among covariates considered, none were identified as significant predictors of