Admissible statistics from a latent variable perspective

Zand Scholten, A.

Publication date
2011

Preface

The idea for this thesis came about while teaching a sophomore psychology research course. I had to teach students that they could not perform t-tests and ANOVAs on ordinal data and I thought it would be good to explain the rationale behind this rule of thumb. But when I asked myself why such tests are a bad idea I drew a blank. One thing led to another, and what started out as a small research course assignment turned into a PhD project. I have been occupied with admissible statistics and measurement levels for quite some time now and have encountered very different schools of thought on the issue, including approaches that say these tests should not be considered bad practice at all. If you ask me which of these approaches is the right one I can give you two answers. The short answer will still vary according to which day you ask me. The long answer will take a very long time to produce, since it contains a lot of ifs and buts, and of course it will turn out to be a compromise that tries to do justice to all sensible viewpoints. You can find it in the concluding chapter of this book.

Hopefully you will find the chapters that precede it more interesting. They illustrate, pretty much in chronological order, how my thoughts on measurement levels and admissible statistics evolved. In Chapter 2 the argument that statistics do nothing more than answer the question whether it is likely that a sample originated from a certain distribution and therefore have nothing to do with what the numbers refer to. This position seemed

very attractive for a while. But then researchers are generally in the business of inferring conclusions about what the numbers refer to, not the numbers themselves. A more in depth reading of the representational and meaningfulness literature convinced me that even with a pragmatic outlook, the legitimacy of inferences needs to be warranted.

A similar change of heart happened while writing Chapter 3. I started out thinking the correspondence between the Rasch model and Additive Conjoint Measurement (ACM) was just a happy coincidence and the claims of interval level measurement should be taken with a large grain of salt. When I found out that Rasch developed his model with the express purpose of obtaining interval level measurement, this made me reconsider its measurement pretensions. The use of probabilities in an emprical structure is critized by some, but this does not form a substantial problem for the interval level claim in my view. I do find it very hard to conceive that it is possible to keep finding items that can be placed between existing items in terms of difficulty ad infinitum. Only a few, if any abilities will allow their structure to be captured by items for which this is possible. Given a set of items that have desirable test qualities and are deemed good indicators by experts, I find it hard to claim the interval level of measurement for a test that consists of a subset of these items, where some have been eliminated due to model misfit while experts cannot point out any flaw in the item. The Rasch model may form an instantiation of ACM, but for all but some properties it will never hold for all conceivable items that tap into the property of interest.

When my promotor Han told me about an interesting little paradox concerning the Guttman and Rasch model I was once again confused about the interval level measurement associated with the Rasch model. Could probabilities be a problem for the Rasch model after all? Perhaps their lack of spatio-temporal presence could be accommodated but the fact that probabilities that represent and introduce error produce a higher measurement level had me stumped. After a long process involving many frustrating discussions with Denny and Gunter on this issue I feel I now have a more firm grasp on the role of error, precision, information and probability in the Guttman and Rasch model. I am comfortable with our conclusion that there is in fact nothing paradoxical about the difference in measurement level, but our analysis has left me uncomfortable with the fact that the logistic function is so pivotal in obtaining additivity. The normal ogive model, which seems entirely equivalent in substantive terms at least, cannot boast interval

level measurement. The lack of a substantive argument to prefer one over the other remains troublesome.

With the Rasch model as the only viable option to approach interval level measurement as specified in RMT, the chances of interval level measurement in psychology seem limited. Not very many properties can be scores using test or questionnaire items, and even if they can only very few will result in acceptable fit. A more practical approach seems to be to promote awareness of the risk of illegitimate inference and assess the amount of risk under relevant circumstances. Although focused on a very specific set of circumstances – small sample, fixed-effect two-by-two design – and very specific assumptions – a two parameter logistic or graded response model –, it was a nice change to be able to contribute to the debate on legitimate inference in a more practical and positive way.

We may not have many opportunities to demonstrate interval level measurement and latent variable models may have a limited use in this respect, but in a more indirect way they can be a very important tool to ensure that at least we are not setting ourselves up for illegitimate inference, in the case our interval level assumption are valid. This is the message I hope comes across in the chapters that follow.

Finally it should be noted that, although this thesis can be read as a book, the chapters are based on research papers that have been published or are currently under review. These papers have been altered slightly in some parts, considerably in others to form more cohesive chapters. There remains some overlap between the chapters however in terms of explanation of background and basic concepts to ensure that each chapter can be read on its own. The research papers that the chapters are based on are listed below.

Chapter 2: Zand Scholten, A. & Borsboom, B. (2009). A reanalysis of Lord's statistical treatment of football numbers. *Journal of Mathematical Psychology, 53*, 69-75.

Chapter 3: Borsboom, B. & Zand Scholten, A. (2008). The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory & Psychology, 18*, 111-117.

Chapter 4: Zand Scholten, A., Borsboom, D. van der Maas, H.L.J., Maris, G.K.J. & Iverson, G. (under review). The Guttman-Rasch paradox in

Item Response Theory.

Chapter 5: Zand Scholten, A. & Borsboom, D. (under review). How to avoid the misinterpretation of interaction effects.

Chapter 6: Zand Scholten, A. & Borsboom, D. (under review). Using the Graded Response Model to assess the risk of misinterpreted interaction effects in fixed-effects ANOVA settings: Polytomous items mitigate inferential error.