Admissible statistics from a latent variable perspective

Zand Scholten, A.

Publication date
2011

Citation for published version (APA):
Zand Scholten, A. (2011). *Admissible statistics from a latent variable perspective*.

Chapter 7

Discussion

The purpose of models is not to fit the data but to sharpen the question.

Samuel Karlin

7.1 Summary

The contribution of latent variable models to the determination of measurement level and the legitimacy of inferences is obviously controversial. In this thesis I have tried to argue that latent variable models can be useful in this respect but perhaps in a more modest way than some propose. In this final chapter first the conclusions and results of the preceding chapters will be summarized, followed by a more general conclusion about the usefulness of latent variables in addressing measurement level issues in psychology.

7.1.1 Measurement vs. Statistics

Chapter 2 of this thesis focused on Lord's influential contribution to the measurement-statistics debate (1953), which supplied most of the ammunition in this fierce conflict. The theory of measurement levels and admissible statistics (Stevens, 1946) states that the choice of statistical test should depend on the measurement level associated with the data. This is because a statistic should remain invariant, resulting in the same inference about the measured property, under all structure-preserving transformations. Lord presented a fictional example where an inadmissible t-test was performed on nominal football numbers, in order to conclude whether a sample of these numbers was taken randomly from the population. The implicit message is that inadmissible tests can be useful and should not be prohibited.

Lord's argument was shown to be flawed however. The assumption that the numbers are nominal representations can be questioned. A reinterpretation of the thought experiment showed that the test can be considered an admissible test on data representing the interval property of bias towards low numbers in the machine issuing the numbers. A conclusion Lord himself would probably have agreed with, given his follow-up publication a year later (Lord, 1954). In this short rejoinder he indicates that when the numbers represent a property at the ordinal level and the goal is to infer something about this underlying property, then a t-test is ill-advised. The initial, intended point was ostensibly that each research question requires a careful analysis of what property the data represent, what level is associated with this representation and what inference is to be made based on the statistical analysis.

It is unfortunate that this rejoinder did not receive more attention, since the debate was rekindled using Lord's original thought experiment several times (Gaito, 1980; Velleman & Wilkinson, 1993). Although the debate lost some of its fire with the advent of Representational Measurement Theory (RMT), it was never fully resolved. Perhaps this is because RMT, which fully formalized Stevens' concept of isomorphism between the empirical objects and numbers, changed the nature of the debate by also recasting the problem of admissible statistics into the concept of meaningfulness. This focused our attention on what was at the center of the issue all along, namely the legitimacy of the inference, not the admissibility of the statistic per se.

Although it was welcomed by experts in the field of measurement theory and psychometrics (Hand, 2004; Embretson & Reise, 2000; Bond & Fox, 2007), RMT was never incorporated by the broader research community. Textbooks to this day present measurement levels and the associated restriction on statistical manipulation in Stevens' terms, without even explaining his informal concepts of representation and meaningfulness, let alone the improved versions of these concepts, axiomatized in RMT (Tabachnick & Fidell, 2007; Agresti & Franklin, 2008). Glossing over the rationale behind Stevens' theory, students are encouraged to simply assume data based on Likert-scale items provide interval level measurements, thereby ignoring the need to show empirically what type of isomorphism exists between the objects and the numbers.

On the other hand, one can ask what the practical consequences of such misapplication of Stevens' theory are. In Lord's thought experiment an inadmissible test resulted in a useful conclusion; it was found to be useful because it actually concerned an interval level property. The alternative moral of Lord's story could therefore be that we should not worry so much about what test we should or should not perform, but instead we should direct our attention to results of inadmissible tests that consistently turn up under replication. Although they do not provide direct evidence that the property that was intended to be measured has a quantitative structure, such results could very well indicate that there is an interesting property lurking in our data, waiting to be identified, not unlike the property of bias in Lord's thought experiment.

Whether such an interpretation seems appealing or not, and whether one decides to adhere to the prescription of admissible tests or not, the general conclusion that follows from both approaches is that we should put more effort into actually testing assumptions about the structure of the property we are measuring. Improving our understanding of the structure of psychological variables seems the only way to increase our ability to conclude something interesting about them.

7.1.2 Rasch Model as Additive Conjoint Measurement

In chapter 3 the usefulness of a particular latent variable model in assessing the measurement level of psychological variables was discussed. The Rasch model according to some is an instantiation of Additive Conjoint Measurement (ACM). ACM is a measurement structure defined in RMT that allows additive representation of an empirical structure indirectly for two properties that conjointly determine a third. In the Rasch model item difficulty and person ability jointly affect a third variable, namely the probability of answering an item correctly. In order to decide whether the Rasch model can help us to identify quantitative properties, several objections and problems with this claim of interval level measurement by the Rasch model were discussed.

To do so it is important to first ensure it is clear what we mean with quantitative measurement. According to RMT, the goal is to show that certain qualitative relations hold for a set of objects, and that these can be validly represented by the numerical relational system of the real numbers. More explicitly, a property is quantitative if the numerical mapping is unique up to affine transformation or multiplication by a positive constant.

This is not the only theory on measurement however. Michell (1997) advocates a return to what he calls the classical theory of measurement, going back to Euclid. In this theory measurement is the discovery of ratios between quantities. Numbers are not mathematical constructs but rather they are instantiations of the ratios and thereby empirical entities. Whichever theory one subscribes to does not have pragmatic consequences as far as the present discussion goes however, since both theories require the empirical demonstration of adherence to the axioms formulated in RMT.

Therefore the question that remains is whether the axioms of ACM are met by data generated by the Rasch model and whether this guarantees interval level measurement. The first part of the question seems uncontroversial. No critic has raised objection to the claim that the Rasch model is structurally equivalent to ACM. The aim in developing the model was to achieve specific objectivity, which corresponds to the single cancellation axiom of ACM. Specific objectivity refers to the fact that the determination of person abilities is independent of the items, and vice versa. Adherence to the other important axiom of double cancellation is ensured by the choice of the logistic function that translates the difference between person ability and item difficulty into the probability of answering an item correctly. The objections against the claim of interval level lie not with the structural equivalence with ACM, but with differences in the interpretation of the concepts 'objects' and 'empirical demonstration' in RMT.

In RMT the set of objects that the measurement structure applies to needs to be clearly defined. Unfortunately it is impossible for most psychological properties that are amenable to measurement using test items, to specify beforehand what the characteristics of a valid item are. A fitting Rasch model is often achieved by sequential item deletion, where items that do not fit the model are discarded while the cause of misfit is unclear. This results in ad hoc, arbitrary tests that show an artificially created quantitative structure by careful selection. Obviously if the focus is on producing such measurement instruments without investigating the reason that items do not fit, no progress will be made.

Another problem concerns the intrinsic impossibility for many psychological properties to find increasingly precise item difficulties. If properties are truly quantitative it has to be possible to always find an object in between any other two objects even if they are very close together. How this can be done for items that assess traits such as extraversion or spatial ability, is unclear.

The empirical status of probabilities is also considered problematic when regarding the Rasch model as a form of ACM. RMT requires the empirical demonstration of adherence to its axioms. This means the objects and their relations should be somehow observable. Kyngdon (2008a) objects that the probability of answering an item correct, which figures as an object in the RMT formulation of the Rasch model, is not spatio-temporally located. It remains to be seen whether the representationalists would not accept the Rasch probabilities as empirically testable or observable, since they are not very explicit about what is required exactly. For example, the probability can be replaced by a more tangible property, the proportion with which the item is answered correctly after repeated administration. Alternatively items of the same difficulty can be administered once to determine the proportion for an equivalence class of items, or the proportion can be obtained by administering an item to persons with the same ability level. This objection to the use of probabilities in the empirical relational structure seems overshadowed by problems associated with the discrete structure of many psychological properties and item selection however.

7.1.3 Guttman-Rasch Paradox

Chapter 3 discussed objections against the claim that the Rasch model is a form of ACM. In chapter 4, a more general critique of this claim was considered. The critique is based on an apparently paradoxical difference in measurement level between the Rasch model and what is considered an error free version of this model, namely the Guttman model. If precision is increased in the Rasch model, the probabilistic model transforms into the deterministic Guttman model, which is associated with an ordinal level of measurement. The claim of interval level measurement linked to the Rasch model should be doubted because of this paradox, since the addition of error to a deterministic structure resulting in quantitative representation is contradictory, at least according to Michell (2008a, 2009).

Our analysis of the paradox showed however, that the addition of error per se is not critical for producing a higher measurement level. For both models the axioms of ACM and the manner in which error is incorporated in the model were considered. There are several ways to add error to the Guttman model that do not change the measurement level. It is the continuous versus the discrete nature of the error and the direct relation between the size of the error and the underlying property, that leads to a higher measurement level for the Rasch model.

The fact remains that, even if the error in the Rasch model is of a very special continuous kind, it is still error that produces better measurement characteristics. The argument that error should not increase measurement precision remains persuasive. This argument is somewhat misleading however, since it is phrased in terms of precision, not measurement level and phrased as such it does not apply in all circumstances. The addition of error to a measurement procedure does not necessarily have to lead to a decrease in precision. In fact, in some cases error can improve the quality of measurement or ability to detect a signal. This phenomenon is known as stochastic resonance in physics and biology (Gammaitoni et al., 1989; Chatterjee & Robert, 2001; Ries, 2007).

With a simple simulation of a thought experiment where length is measured by paired comparison, we showed how error that is somehow related to the property being measured, can lead to higher precision. Another simulation showed that stochastic resonance can also occur in the Rasch model, but that it is only apparent when items are awkwardly spaced in terms of difficulty. Since the size of the error is directly related to the difference in item difficulty and person ability, removing it is equivalent to throwing away valuable information.

The difference in measurement level between the Guttman and Rasch model should not be considered paradoxical. The Rasch model uses continuous probabilities that are related to the underlying trait; the Guttman model uses discrete probabilities of zero and one, which cannot be distinguished from the raw scores and therefore contain no extra information beyond these raw scores. To call a difference in precision – or measurement level – between these models a paradox is like saying the difference in ability to detect far-off plantery systems between a pair of binoculars and an infrared telescope is paradoxical. The two models should be viewed as modeling two different measurement procedures, which differ in many more respects than simply the addition of random error.

7.1.4 Assessing Inferential Error with IRT

Even if one accepts that the Rasch model is an instantiation of ACM and therefore provides interval level measurement, even if one is willing to accept probabilities as measurements and even if one is willing to assume that an endless supply of people and items is available for a particular trait or ability, it is unlikely that the Rasch model can be used to establish interval level measurement for very many psychological properties because of its stringent nature. Although other latent variable IRT models cannot be used directly to ascertain interval level measurement, these types of models can be used to answer other questions pertaining to measurement level and legitimacy of inferences. In most cases we assume an underlying property is quantitative, but we are only able to obtain data at the ordinal level. We can only assume that these ordinal data are related to the latent variable via some monotonically increasing function. In such a situation, latent variable models can be used to assess the risk of making an inferential error based on an inadmissible test if these assumptions hold.

IRT models specify parameters that refer to substantive characteristics of items and persons and can therefore guide our assumptions about the specific form of the relation between the latent and observed level. This allows us to narrow down the circumstances that put us at risk of making inferential errors. In the final two chapters 5 and 6 the Two-Parameter Logistic model (2PL) and the Graded Response Model (GRM) were used to transform latent effects into observed scores to asses the risk of inference errors for factorial research design concerned with interaction effects. These effects are potentially vulnerable to misinference due to arbitrary choice of scale for ordinal data. For standard two-group mean comparisons, inferential error is impossible if the assumptions of the statistical test are met – samples are independent and normally distributed with equal variances (Davison & Sharma, 1988). For simultaneous comparison of groups on two or more factors, the normality and homogeneity of variances assumptions do not protect against inferential error however (Davison & Sharma, 1990, 1994).

Simulation research using the Rasch and 2PL model showed that inappropriate test difficulty, increased item discrimination, test length and effect size had a detrimental effect on the number of significant interactions detected in data that showed no interaction on the latent level (Embretson, 1996; Kang & Waller, 2005). These simulations were performed using a Moderated Multiple Regression (MMR) approach, with random factors and a large sample size. Under such circumstances latent ability estimates obtained from IRT modeling can be used to limit the risk of inferential error. For more typical experimental settings with small sample sizes, it is not possible to reliably fit an IRT model. To assess the risk of inferential error under typical experimental settings we simulated a two-by-two fixed effects design, with small sample size and small to moderate effect sizes with different types of main and interaction effects at the latent level.

The effect of inappropriate test difficulty, item discrimination and sample size was investigated for latent values transformed according to the 2PL model. The 2PL model extends the Rasch model by allowing items to discriminate differentially between persons on the latent ability. Results showed that an inappropriately difficult test results in a floor effect on the observed scale, distorting the original interactions, or lack thereof, thereby increasing the number of inference errors. High item discrimination exacerbates this risk. The risk is elevated only for certain types of effects however. When there are two main effects and no interaction or there is an order-independent interaction present at the latent level, risk of inferential error is high. Due to sampling error a risk is also present, albeit to a much lesser degree, for partial and minimal interactions. Sample size distorted results more when the 'true' effect showed no interaction, but less when this effect did consist of an interaction.

Tests of the assumptions of normality and homogeneity of variances both showed higher rejection rates due to the floor effect caused by the nonlinear transformation specified by the 2PL model. Unfortunately these rates increased equally for effects that were distorted and effects that remained unchanged. Although rare, the latent effects could be distorted even when normality and homogeneity of variances assumption held. For situations where IRT modeling is not an option, such as experimental settings with small sample size, checking of these assumptions does filter out a large part of inferential errors but not all of them; it can also easily lead one to discard a sound inferences.

A better way to safeguard against inferential errors is to employ a test that is appropriate for the sample in terms of difficulty. If this is not possible a normalizing transformation of the data can at least mitigate the risk of inferential error. Such a transformation can be considered an informal attempt at approximating the latent ability values, without modeling the latent ability.

These and previous results all pertain to binary scored items. Of course, it is interesting to see whether the risk of inferential error is also present for items that allow a person to choose a response from more than two response options. Since these types of Likert-scale items are supposed to provide measurement that is close enough to assume the interval level, we would expect a smaller risk of inferential error here. Results from a simulation study using the GRM to transform latent values into observed scores confirmed this expectation. The same setup as before was chosen with the difference that only a small sample size was used and that different numbers of response categories were simulated.

Results were similar to, but less extreme than those found for the dichotomous item models. As the number of response categories increased, the number of inference errors in high-risk circumstances decreased. A normalizing transformation seems to provide a good way to mitigate the remaining risk of inferential error, removing the risk almost entirely for higher numbers of response categories.

Of course these results all hinge on the assumption that a quantitative latent variable is present, that the observed scores are ordinal and that the data were generated by a process that produces results that are accurately described by the 2PL model and GRM respectively. Whether these models are plausible forms the subject for another thesis. If we provisionally accept them however, these models enable us at least to identify situations that put us at an elevated risk of making inferential errors when we perform strictly inadmissible tests.

7.2 Conclusion

Legitimate inference is essential to any scientific endeavor. In most physical sciences the legitimacy of inferences is not an issue because measurement is straightforward and the quantitativeness of properties is self-evident, rendering almost all tests admissible and the corresponding inferences legitimate. In the social sciences the situation is quite different. It is highly doubtful whether many variables intrinsically have a quantitative structure, and even if they do, it remains to be seen whether measurement procedures will ever be found that allow representation of such structure at an interval or ratio level.

One would expect that in the face of so much uncertainty, social scientists would take up the challenge and put a fair amount of effort into establishing at least an ordinal level of measurement for their measurement procedures, into singling out and investigating the structure of those properties that are the most likely candidates for showing quantitativeness, and into actively assessing or avoiding the risk of inferential error. Except for in some areas of psychophysics however, most researchers in experimental and applied psychology seem happy to simply assume an interval level of measurement and to leave the task of verifying this assumption to psychometrics.

7.2.1 Measurement in psychometrics

Unfortunately, psychometricians do not seem to put verification of measurement level assumptions very high on their list of priorities. Psychometrics is concerned with the formulation of latent variable models that at first glance do not seem to say much about the quantitative structure of psychological properties. In factor analysis continuous latent and observed variables, requiring an interval level, are assumed; this assumption is never directly tested in the model however. The same applies to IRT models that specify the underlying variable as continuous and the observed variable as categorical. The 'measurement' models in latent variable modeling apparently address the question whether certain observed scores are good indicators of some underlying latent variable, not whether the structure of this variable is quantitative.

This kind of question *is* tackled in psychometrics when the fit of factor or IRT models is compared to the fit of latent class or latent profile models, which assume a categorical latent variable. These types of comparisons are used to establish, for example, whether the development of certain cognitive abilities should be seen as a gradual increase on a single continuum (assuming a continuous and quantitative structure) or as the progressive acquisition of more elaborate solving strategies. Of course such a strategy helps to identify latent properties with a strictly nominal structure, but if such investigations show preference for a model specifying a continuous latent variable, this does not mean we can assume the latent variable is quantitative, since ordinal variables are generally well approximated by continuous linear and nonlinear models.

Other even more basic questions that psychometricians are engaged with concern the dimensionality of properties and the identification of coherent sets of indicators of a latent variable. An appeal to test the assumption of quantitativeness seems very ambitious when in many cases the prerequisites for determining nominal and ordinal, let alone quantitative structure have not even been met.

7.2.2 Role of Latent Variable Models

One model that might allow verification of measurement assumptions is the Rasch model, which is said to be structurally equivalent to ACM. Although objections to this claim that concern the use of probabilities as empirical observations and the counter-intuitive role of error have been assuaged, it is unlikely that the model will allow for the indisputable claim of interval measurement for many psychological variables. It is hard to conceive of psychological properties for which a clearly defined set of items can be specified in advance that will allow the endless production of items that will all fit the model. In practice the number of items that will result in acceptable fit is small; a good explanation of why some items show misfit cannot always be found. Even for relatively simple properties it is often hard to conceptualize increasingly – let alone infinitely – more fine-grained items between two existing items in terms of difficulty. It must be noted that the limited usefulness of the Rasch model is due to the nature of psychological properties, not the (statistical) characteristics of the model itself.

Another, more modest but much more useful application of latent variable models, at least for now, lies in the assessment of the risk of inferential error due to inadmissible tests, combined with the assessment of research and test characteristics that could be associated with the risk of inferential error. Of course such investigation starts from the assumption that the latent variable actually has a quantitative structure, but as long as this assumption is clear and explicit, the assessment of this risk seems very welcome information. Another way that latent variable models can be of use, is the continued development of models and tests for ordinal structures, so that more alternatives to inadmissible tests become available. This issue was not addressed in this thesis, since it does not bear directly on (il)legitimate inference.

7.2.3 Admissible Statistics, Yes or No

Should inferences be made based on inadmissible tests? The legitimacy of inferences should be warranted, but requiring the empirical demonstration of quantitativeness before allowing any type of parametric test would bring the social sciences to a grinding halt. The conclusion is therefore that inadmissible tests can be performed, but that a disclaimer should accompany the inference based upon it. As long as the conclusion one draws is an informed one, in which the risk of inference error due to arbitrary choice of scale is not only recognized but assessed, it should not be prohibited.

In many cases this risk, or lack thereof, is known or at least something can be said about it. For many simple two-group comparison tests the assumptions of the statistical test eliminate this risk. For comparisons where groups are compared simultaneously on two or more factors, the risk is often negligible and when it is not it can be mitigated by optimizing research design and choice of measurement instruments. For more complicated tests and models, more research is obviously needed. Another interesting avenue of research is to investigate, in situations where the risk of inference errors is present, what type of transformation produces erroneous inferences. It is conceivable that although a distorting transformation can be found, in many cases the transformation will be so outrageous that it seems very unlikely to represent the true relation with the underlying property (again assuming this has quantitative structure).

7.2.4 RMT in Psychometrics

This is not to say that social science is off the hook as far as investigating the (quantitative) structure of its properties is concerned. The question is whether we should actively try to fit our existing measurement procedures into the molds of RMT measurement structures. Given that we are still struggling to come to grips with the dimensionality and fundamental structure (nominal vs. ordinal) of our properties, adherence to the unforgiving axioms of RMT seems far beyond our reach for quite some time to come. Perhaps critics such as Michell have a point when they argue that the social sciences should not present themselves as a quantitative field when they are not. However, even if the word measurement was no longer used for the act of categorizing or ordering on some property starting tomorrow, these critics would probably argue that the object of any science is to express the world in terms of quantitative laws and urge us to actively seek out ratios of quantities. Such admonitions will not bring psychology closer to being on par with physics however, because the subject matter in psychology is much more complex and we have only barely started to study it. We will either slowly plough our way through the mess that is psychological measurement or, if we fail to make any progress, be required to reinvent ourselves as a qualitative science, as large parts of medicine, biology and genetics are. In either case, the critics demanding quantitativeness will get it later, rather than sooner.