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 1

Introduction

I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be.

Lord Kelvin

All science is either physics or stamp collecting.

Ernest Rutherford

1.1 Measurement in Psychology

Measurement has always come naturally to the physical sciences. The subject matter in these fields lets itself be expressed in numbers relatively easily. Construction of measurement instruments is – again relatively – straightforward, since our grasp of the structure of the property that is to be measured is often clear. The famed statements quoted above, emphasizing the importance of quantification in science were obviously made from a position of privilege. Unfortunately fields like psychology have to make do under less desirable circumstances. The ontological status of psychological variables is in most cases uncertain. Concepts are fuzzy and vague. The structure of psychological traits or abilities such as extraversion, intelligence or quality of life is poorly understood. Whether such traits and abilities form variables appearing consistently in nature, effecting change in other variables in a systematic manner, is questionable for most variables. Even if such structure exists at all, psychological variables are messy things, ridden with systematic and random error. It should come as no surprise that measurement in psychology has been surrounded with controversy from its inception.

Whether psychological variables can be measured of course depends on what one means by 'measurement'. Around the time that psychology came on the scientific scene – at the close of the nineteenth century – the best definition representing the state of affairs in physics was given by Campbell (1920). He maintained that the basis for measurement is the assignment of numbers to objects. But this definition is not complete. The reason for expressing some property in numbers is to enable the application of mathematics to these numbers and the formulation of physical laws. In doing so, the additive structure of the real numbers is used. In order for the results to be representative of the property being measured, this property needs to exhibit the same additive structure. To show additive structure requires direct or indirect concatenation of objects and the adherence to several axioms that ensure additivity. For example, to show that length is an additive property one can lay different rods end to end and compare the resulting concatenation to other rods. If any comparison of singular or concatenated rods can be shown to be a weak order and other axioms are also satisfied, then we can conclude length is additive.

Not many psychological variables are even remotely likely to conform to this definition of measurement. This led physicists like Norman Campbell to object to claims of measurement in psychology (Campbell, 1940). These objections resulted in the formation of a committee (Ferguson et al., 1940) that was to investigate whether measurement of psychological variables was at all possible. The committee was unable to reach consensus, even after eight years of deliberation and the matter remained unsettled for several years.

Stevens (1946) broke this impasse by suggesting we let the numbers conform to the structure of the property, not the other way around. Instead of demanding additive properties because we want to use the additivity of numbers, we can consider what structure is determinable and use only the corresponding characteristics of numbers. This redefinition of measurement resulted in the well-known levels of measurement. The nominal level refers to assignment of numbers where only the inequality of the numbers is used to denote different types or classes of objects. Similarly the ordinal level concerns the representation of ordering of the objects on the property of interest; the interval level concerns the representation of quantitative comparison of differences between objects; and finally, the ratio level concerns representation of direct quantitative comparison of objects.

1.2 Admissible Statistics

Letting the numbers conform to the structure of the property of interest does have some drawbacks. It means that the additive structure of numbers for data that represent a property on the ordinal level, for example, does not have an empirical counterpart. Since there is no additive structure in the property, we should not use this structure in the numbers. This idea was reformulated into the theory of admissible statistics, in which level of measurement is identified by the type of transformation that leaves the structure intact. Numbers that represent a property at the nominal level can be transformed according to any one-to-one function; for the ordinal level the corresponding transformation can be any monotonically increasing function; for the interval level an affine transformation is needed; and for the ratio level multiplication by a positive constant is the only allowed transformation.

When an inadmissible transformation is performed, the structure that was originally represented is lost. This applies not only to transformations but also to other numerical manipulations of the data, such as the computation of test statistics. A test statistic, or at least the conclusion that is based on it, should not be determined by structure in the numbers that is not present in the underlying property. More importantly, the conclusion based on a test statistic should remain invariant under any permissible transformation of the data. If it is not invariant this means that the conclusion depends on some arbitrary characteristic of the chosen scale.

The theory of measurement levels and the accompanying theory of admissible statistics was eagerly accepted by psychologists, since it uncomplicated the problem of measurement in psychology. The question whether psychological measurement was at all possible transformed into the question of what level of measurement could be achieved. Statisticians however, where not charmed with the notion that level of measurement should limit the use of statistics. They argued that a statistical test tells us something about the probability that a sample was drawn from a specific distribution of numbers The assumptions involved in statistical testing contain no reference to what the numbers actually represent and a test can therefore always be performed on any data that satisfy the assumptions (Lord, 1953; Burke, 1953; Gaito, 1980; Velleman & Wilkinson, 1993). A fierce debate ensued. This debate, discussed in Chapter 2, fell silent however with the advent of Representational Measurement Theory.

1.3 Representational Measurement Theory

RMT was built on the ideas of Stevens' levels of measurement but far exceeds it in scope and rigor. In RMT representation of the structure of a property is formalized in terms of a homomorphism between an empirical relational structure and a numerical relational structure. The empirical structure consists of a clearly defined set of objects that exhibit the property (rods), and a set of empirical relations (comparing the extension of adjacent rods) and operations that define the property (laying rods end-to-end). The goal in measurement is to find a homomorphism – a structure-preserving mapping – into a numerical structure that consists of a set of numbers (the reals) and a set of numerical relations ('larger than') and operations (addition). Whether such a mapping exists is axiomatized in a representation theorem. The associated measurement level is provided in a uniqueness theorem. Several types of measurement structures exist. Besides the additive structure that best suits the property of length there are, for example, also multiplicative structures and conjoint structures that are indirectly additive.

RMT is a very elegant and fully formalized theory of measurement and is considered the accepted view by most people who write on the subject of measurement in psychology (Hand, 2004; Roberts, 1979). Not all subscribe to this view however. The most noted critic of RMT is Michell (1986, 1990, 1993, 1994, 1997, 2003) who advocates a classical theory of measurement that via Hölder (Hölder & Michell, 1997) goes back to Euclid. According to this theory, measurement consists of the determination of ratios of quantities. These ratios are discovered rather than constructed, as is the case in RMT. The divide between empirical objects and numbers, the latter of which exist separately in a non-empirical realm, is the basis for RMT does not exist in the classical theory. Numbers *are* the ratios between quantities that are found in nature. Whether any property can be expressed as a ratio between quantities is a question that needs to be decided empirically. Although the two theories have very different philosophical underpinnings, the axioms of RMT, at least for interval and ratio structures, can be used to provide empirical support that a property is measurable or quantitative. Note that in the classical theory measurement and quantitativeness can be used interchangeably. In RMT measurement is defined as structure-preserving representation, and although the emphasis is on quantitative structures, measurement can also refer to ordering or classification.

As to the admissibility of statistical test, both theories agree that an inference based on a test statistic should not change under structure-preserving transformations of the data. In RMT the concept of invariant statistics first proposed by Stevens (1946) was formalized and reformulated into the concept of meaningfulness. Meaningfulness refers to the invariance of the truth-value of a statement based on a test statistic, not the invariance of the statistic itself. Unfortunately the formalization of the concept resulted in necessary but not sufficient conditions for meaningfulness (Narens, 1988, 2002). In the classical theory of measurement the concept of meaningfulness is accepted but often referred to as legitimate inference, which is perhaps a more befitting term.

It is interesting that in most introductory psychology textbooks on research methods and statistics (Tabachnick & Fidell, 2007; Agresti & Franklin, 2008) neither RMT nor the classical theory is discussed. The measurement levels first introduced by Stevens are rehashed and rules of thumb concerning the inadmissibility of certain statistical tests are provided, mostly without any justification. Unfortunately this has led to the widespread practice of simply assuming an interval measurement level, without testing this assumption. The lack of familiarity with more rigorous theories of measurement has certainly not helped advance psychological measurement.

1.4 Latent Variable Models

It remains questionable whether measurement in psychology would improve if these theories were well-known among experimental psychologists. Both theories are highly prescriptive and deterministic, leaving little room for the considerable amount of error that is present in most psychological measurements. This is where latent variable models could come in handy. Very generally, latent variable models are psychometric models that assume the observed scores we collect as measurements are caused or explained by one or more underlying variables, or latent properties. We will consider only simple, unidimensional measurement models consisting of one latent variable. The underlying property we are measuring can be very straightforward, such as length, but in psychology it is most often only indirectly available to us. Instead of taking one measurement of someone's length, we administer multiple items that assess a trait or ability. Latent variable models are further characterized by assumptions about whether the observed and latent variable are continuous or categorical, see Table 1.1.

Table 1.1: Latent variable models categorized according to type of observed/latent variables

	Observed	
Latent	Continuous	Categorical
Continuous	factor analysis/SEM	latent trait (IRT)
Categorical	latent profile	latent class

We will consider only models with a continuous latent variable, which we will assume is quantitative, and categorical observed scores, which in most cases will be ordinal. This combination best represents the problematic situation faced in psychology; the property of interest is assumed to be quantitative, but the best representation available to us is of the ordinal level. Of course this does not immediately get us anywhere. So far the use of a latent variable model has only made some implicit assumptions explicit. However, the latent trait or Item Response Theory (IRT) models we will consider, add the assumption of a specific type of relation between the latent property and observed scores. Different IRT models specify different relations.

The general question that formed the motivation for this thesis is how we should deal with the problem of admissible statistics and legitimate inference from a latent variable perspective. The concept of admissible statistics entails a very strict limitation on the tests that we can perform on data that were modeled with a latent variable model. Whether this prescription should be given heed or whether assumptions or other characteristics of latent variable models protect us from making illegitimate inferences is an important topic, especially since psychometrics has been under increased attack recently from critics who maintain that psychometrics is shirking its responsibility to actively inquire into the quantitative status of psychological measurement (Michell, 1986, 1997, 2008a, 2008b, 2009). Perhaps latent variable models can be used to lay bare this elusive quantitative structure in more psychological properties or assist in the answering of measurement level questions in other, more indirect ways.

1.5 Overview

In this thesis the feasibility of legitimate inference and the assessment of measurement level for psychological properties is discussed from a latent variable perspective. The three chapters following this introduction are devoted to several theoretical problems associated with admissible statistics and legitimate inference. The first chapter deals with arguments against the restrictions that admissible statistics pose. In the second and third chapter a latent variable model is addressed that is claimed to ensure legitimate inference. There are several conceptual problems that surround this claim. The final two chapters concern a more practical approach. In both chapters a simulation study is presented that assesses the risk of inferential error. Different latent variable models were used to generate the simulated data.

In **Chapter 2** the origins of the measurement-statistics debate are discussed. The argument that statistics should not be governed by measurement level considerations is critically evaluated. The focus of the chapter is on a thought experiment by Lord (1953) that for a large part set off and perpetuated this debate. The thought experiment concerns a fictional professor who feels guilty performing inadmissible tests on his students' ordinal exam scores. He suffers a breakdown and retires early. In his new job, selling jersey numbers to the football teams, an altercation between two teams arises.

One team complains that they received lower numbers due to tampering by the other team. A statistician settles the matter by performing a t-test on the mean of the numbers. Although this test is inadmissible, since the numbers are of the nominal level, only distinguishing players on the field, the conclusion seems useful and somehow meaningful. A critical evaluation of this thought experiment shows however, that the parable can be reinterpreted so that it is perfectly in line with the rationale behind the concept of legitimate inference. A latent variable perspective is used to identify the underlying property of interest and a representational approach is used to show that this property can be represented at the interval level. This analysis shows not only that the most influential argument against inadmissible statistics is flawed and that application of only admissible statistics seems justified, but also that measurement level considerations are more complex than we generally think.

The property of bias in a number-issuing machine in Lord's thought experiment is not a very interesting one from a psychological point of view. For most properties that are of interest to psychology, interval level measurement is unattainable. There are circumstances however where such measurement is possibly within reach. These circumstances are captured by the assumptions of the well-known Rasch Model. This model, according to many, is an instantiation of an additive conjoint representational measurement structure. This is a structure for which it was shown (Luce & Tukey, 1964) that if its axioms hold, the representation is of the interval level. This is a potentially important foothold for psychology. If the Rasch model can provide us with a method to attain interval level measurement then for at least some properties legitimate inference will no longer be a problem. The claim that the Rasch model ensures interval level measurement is not uniformly accepted however. In **Chapter 3** general problems of combining the latent variable modeling perspective with representational measurement in general and the Rasch model in particular are examined concerning the interpretation of probabilities, the spatio-temporal status of model parameters and the finite and inherently discrete nature of many psychological properties.

Unfortunately these are not the only problems with the interval level claim associated with the Rasch model. Another issue concerns a paradox we call the Guttman-Rasch paradox. This paradox consists of a counterintuitive difference in measurement level between the Guttman and Rasch model. Both models are IRT models that describe the probability of answering an item correctly as a function of the difference of the person ability and item difficulty. The Guttman model is deterministic and states that this probability is 0 if the ability is lower than the item difficulty and 1 if it is higher. The Rasch model is probabilistic and specifies the probability as a logistic function of the difference between person and item. The Rasch model can be considered an extension of the Guttman model that arises from the addition of error by assuming a monotonically increasing relation between probability correct and the latent trait. Now, the Guttman model, which contains no error, allows ordering of items and persons. In contrast, the Rasch model, which does contain error, allows interval level measurement. Introduction of error is therefore associated with an increase in measurement level. This paradoxical situation is discussed in **Chapter 4**, where it is shown that although an increase due to error is counter-intuitive, it is not paradoxical per se. Since the error in the Rasch model is dependent on the difference between item and person, it is in fact informative about the latent trait. In physics and biology this phenomenon is known as stochastic resonance.

Latent variable models can also be incorporated to investigate admissible statistics in a very different way. They are employed to assess the actual risk of obtaining ambiguous or distorted results when an inadmissible test is performed. One type of research effect for which inadmissible tests can lead to distorted results, or inference errors, is the interaction effect. Latent variable modeling can be used to make specific assumptions about the relation between ordinal observed scores and the assumed underlying quantitative variable. In **Chapter 5** several types of interaction effects were simulated according to the two-parameter logistic IRT model (2PL). An experimental setting was simulated by using a fixed-effects, small sample two-by-two design. Results showed that inferential errors occur when the test is ill-matched in difficulty to the ability of the sample. If a test is too hard, floor effects occur that result in unequal stretching of the observed scale, due to the nonlinear relation that the 2PL model specifies. Inference errors only occur however when test discrimination is high and only for certain types of interaction effects. The risk of inferential error therefore seems limited to very special circumstances. Moreover, in these specific circumstances, inferential error can be mitigated to some extent by performing a normalizing transformation on the observed scores.

In **Chapter 6** an extension of this simulation study is presented. The same simulation was performed, but this time the Graded Response Model (GRM) was used to transform latent values into observed scores. The GRM models item responses for items that have multiple ordered response categories. Besides the match between test difficulty and item ability, other factors such as item discrimination, effect size, type of effect and number of response categories was investigated. Results were comparable but much less severe for the data that was generated using the polytomous GRM model compared to the binary 2PL model. The risk of inferential error clearly decreased as the number of response categories increased. A normalizing transformation was also more effective in mitigating the risk of inferential error. These results support the conclusion that inadmissible statistics pose a real threat to our inferences. Only if a test is extremely hard or easy, and only then if the items also discriminate strongly and if certain interaction effects occur.

A summary and discussion of both the theoretical (Chapters 2, 3 and 4) and more practical (Chapters 5 and 6) treatment of admissible statistics is provided in **Chapter 7**. Here it will be argued that the prescriptive requirement of axiom testing advanced by representationalists as well as their fiercest critics, the traditionalists, is not a fruitful strategy to further scientific knowledge of the structure of psychological properties. The properties that are even remotely amenable to such rigid methods are scarce and the available methods themselves are ill-equipped to deal with psychology's error-laden properties. Psychological researchers would do well to be a bit more conservative in their inferences and use of the word measurement when the measurement status of their observed data is dodgy at best. This is not to say that psychology should no longer be considered a science. The fact that psychology deals with a much more elusive subject matter than physics does not mean we should give up at the outset. It is important to realize however, that this is where psychology as a quantitative science still finds itself. We are a nascent quantitative field, not an established one.