Admissible statistics from a latent variable perspective

Zand Scholten, A.

Publication date
2011

Chapter 4

The Guttman-Rasch Paradox in Item Response Theory

For the truth of the conclusions of physical science, observation is the supreme Court of Appeal. It does not follow that every item which we confidently accept as physical knowledge has actually been certified by the Court; our confidence is that it would be certified by the Court if it were submitted. But it does follow that every item of physical knowledge is of a form which might be submitted to the Court.

Sir Arthur Eddington

Abstract

Recently, quantitative measurement in Item Response Theory was questioned because it seems based on a paradox concerning error and precision (Michell, 2008a). The Rasch model ensures interval level measurement. If precision increases in Rasch items, they become Guttman items, losing their quantitative properties. Removing error decreases precision. To address this paradox we consider to what extent both models meet the requirements for interval level measurement. This leads us to conclude that the Guttman model cannot simply be considered an error-free version of the Rasch model. Furthermore, we argue that an increase in precision by adding error is not paradoxical per se, by discussing the well-known phenomenon of stochastic resonance. These arguments together lead us to conclude that the paradox disappears when the crucial aspects of continuity and measurement level, and not error and precision, are considered.

4.1 Measurement Level Difference in IRT Models

There is an unusual difference in measurement level between the closely related Guttman and Rasch models in Item Response Theory (IRT). The Rasch model, a stochastic model that incorporates error, is said to yield measurement at the interval level (e.g., Wright, 1999; Bond & Fox, 2007). A deterministic, error-free version of this model is the Guttman model (Guttman, 1950). Although this model is devoid of error, it provides measurement at the ordinal level only. It has been pointed out by various scholars that this presents us with a paradox (Duncan, 1984; Fischer, 1995; Kyngdon, 2008b). The problem is stated most eloquently by Michell, who asks: *Is it not paradoxical that improving the precision of our observational conditions decreases the precision of our observations?* (Michell, 2008a, p. 15, l. 5-7). In other words, one would hardly expect that removing measurement error from our procedures and instruments would *hurt*, rather than *help* measurement precision.

Previous comments on this paradoxical loss of measurement level were never much more than cursory notes, until Michell (2008a, 2009) recently reintroduced the paradox to argue against the claim that the Rasch model yields interval level measurement. His objection is nicely illustrated by the following analogy. Suppose astronomers discover a new star. They think a planetary system around this star can be observed, although an obscuring haze impedes observation. Now suppose this haze suddenly disappears and observation is entirely unobstructed. If, along with the haze, the planetary system were no longer detectable, what would the astronomers conclude? Obviously the existence of the system would be considered highly doubtful, because the observation depends solely on the presence of error. By analogy, the claim that the Rasch model yields interval level measurement should be doubted, because it too disappears when our view is no longer obstructed by error (Michell, 2008a, p. 15, l. 8-19, paraphrased).

This argument clearly demonstrates a deep concern regarding the merit of measurement claims in psychometrics and psychology in general. We share this concern. The assertion of interval level measurement is anything but trivial. It involves assumptions and restrictions that far too often remain implicit and untested. Objections to such claims, when based on lack of empirical verification or blind acceptance of shaky assumptions, need to be addressed. In this context the Guttman-Rasch paradox clearly demands

our attention. It is important with respect to the status of commonly used measurement models, and by extension to the status of applications of psychological measurement in society at large.

We therefore hope to answer two questions that naturally follow from the renewed interest in the Guttman-Rasch paradox. First, does the paradox indeed have merit as an argument against the claim of interval level measurement by the Rasch model? Second, is the change in measurement level truly paradoxical? Both questions will be answered with a resounding 'no'. To be clear, the goal is not to argue in support of the claim of interval level measurement by the Rasch model. Instead, we show that the paradox is irrelevant to the discussion. The recent characterization of the Guttman-Rasch paradox does not contribute constructively to the debate on the status of measurement pretensions associated with psychological models and measurement procedures. The paradox, as presented above, paints a grossly oversimplified picture of the relation between the Guttman and Rasch models and the relation between measurement level and measurement precision in general. We consider how the argument that employs the paradox is structured and show why this argument is flawed. From this we will see that the paradox itself is an illusion, fostered primarily by ill-chosen formulation. The difference in measurement level between the Guttman and Rasch models is in fact entirely non-paradoxical (see also Sijtsma, 2011 for another approach leading to the same conclusion).

The argument, as we believe it to be generally understood, consists of the following propositions:

1. The Guttman model yields ordinal measurement;
2. The Rasch model yields interval measurement;
3. The only difference between the models is error, and:
4. Removal of error cannot decrease measurement precision.

This last proposition is false when we apply it to the Guttman-Rasch paradox; precision does decrease when we remove error from the Rasch model and are left with the ordinal Guttman model. Therefore scholars like Michell conclude that the second proposition (i.e. the Rasch model yields interval measurement) is false. To show where this argument fails and to better

understand the paradox itself, the validity of each of these propositions will be discussed. Before we can turn to the first two propositions that assert the measurement pretensions of the Guttman and Rasch model however, we need to introduce these two models in more detail.

4.2 Guttman and Rasch

The Guttman and Rasch models are closely related Item Response Theory (IRT) models. IRT is commonly accepted as an improvement over Classical Test Theory (CTT). With IRT we can, for instance, equate tests, investigate item bias, and develop computer adaptive tests. Conceptually, IRT models differ from CTT in two major respects. First, they assume that participant responses to each item of a test can be related to an underlying trait or latent variable. This variable could be a preference, attitude, ability or personality trait, such as paranoia. Both items and participants are assumed to be comparable on the same underlying, one-dimensional continuum reflecting this variable. When a participant 'dominates' an item, we would expect the participant to answer the item correctly or to endorse it. For example, a very paranoid person is expected to endorse the item "I often think people talk about me behind my back". Second, IRT models explicitly specify the relation between the latent variable and the chance of endorsing an item. Different IRT models assume different relational forms. Models can be deterministic or stochastic, discrete or continuous, linear or non-linear.

Table 4.1: Typical probability patterns for the Guttman and Rasch model

(a) Guttman model probabilities

	$a_{(2)}$	$b_{(7)}$	$c_{(12)}$
$x_{(1)}$	0	0	0
$y_{(5)}$	1	0	0
$z_{(9)}$	1	1	0

(b) Rasch model probabilities

	$a_{(2)}$	$b_{(7)}$	$c_{(12)}$
$x_{(1)}$	0.269	0.002	0.000
$y_{(5)}$	0.953	0.119	0.001
$z_{(9)}$	0.999	0.881	0.047

Rows (x,y,z) denote persons, columns (a,b,c) denote items. The probabilities in the cells result from the pairing of persons and items. The subscripted integers indicate the person ability and item difficulty value that determines the probability.

The most restrictive model is the Guttman model. This model specifies that if participants dominate an item, they will always answer this item and all 'easier' items correctly. If an item dominates the participant, the participant will always get this item and all more 'difficult' items wrong (Guttman, 1950). When such data are coded as zeros and ones, and ordered according to persons' ability and item difficulty, the resulting table will show a perfect triangular pattern of zeros and ones. An example of such a Guttman pattern is provided in Table 4.1a.

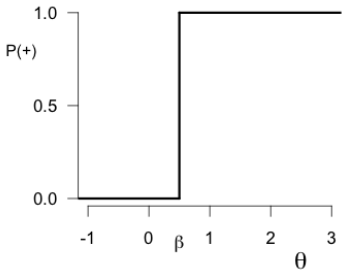
Guttman was concerned with deviations from a perfect pattern and the ability to use the sum score to order persons. He never presented his model in terms of a latent variable or a probability of answering items correctly. However, others have done so (e.g., Andrich, 1985) and the Guttman model is now generally regarded as the most basic IRT model. As such, the Guttman model is a deterministic IRT model: the probability of answering an item correctly is 1 for participants who dominate an item, for the others the probability is 0. The relation between probability of answering an item correctly and person ability is represented by a step function for each item (see Figure 4.1a). The location of the step function on the latent ability scale is determined by the difficulty of the item. Harder items are placed further to the right, requiring a higher ability to answer the item correctly. The model can be formally represented as follows: $P(x+) = 1$ if $\theta > \beta$; else $P(x+) = 0$, where $P(x+)$ denotes probability of answering an item correctly, θ denotes person ability and β denotes item difficulty.

The Rasch model can be viewed as a stochastic extension of the Guttman model. The Rasch model specifies the relation between the latent variable and the probability of endorsing the item as a logistic function (Rasch, 1960; see Figure 4.1c). The Rasch model is a restrictive model, but it does allow the chance of item endorsement to increase as the latent trait increases. This makes sense to most people: for any given math problem, participants who are increasingly good at math will have an increasingly higher probability of getting the problem right than people who are bad at math. Data that fit a Rasch model will result in estimated probabilities like those presented in Table 4.1b. The Rasch model can be represented formally as follows:

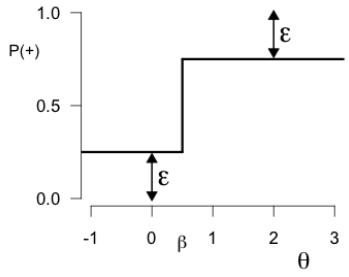
$$P(x+) = \frac{e^{(\theta-\beta)}}{1 + e^{(\theta-\beta)}}.$$

Figure 4.1: Response functions of three closely related IRT models

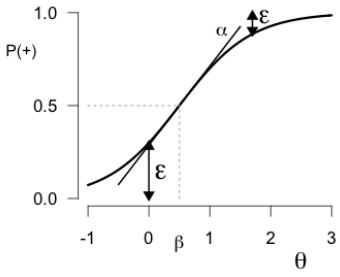
(a) Deterministic Guttman model without error



(b) Proctor constant error rate model with constant error ϵ



(c) Rasch model with error dependent on the latent trait



The functions relate probability of answering an item correctly ($p(+)$), to a latent trait (θ); Item difficulty is denoted by β, item discrimination is denoted by α.

In the Rasch model, the closer a person's ability is to the item difficulty, the greater the probability to get the item wrong when the person in fact dominates the item and vice versa.

4.3 Measurement Pretensions

To understand the seemingly paradoxical difference in measurement level between the Guttman and Rasch models it is essential to know why these models are associated with different measurement levels in the first place. The Guttman model only allows ordering of participants and items on the latent dimension. Differences between participants or items cannot be compared meaningfully. It should be noted that although the Guttman paradigm only allows us to represent the property on an ordinal level, this does not necessarily mean that the latent property lacks quantitative structure. A different measurement procedure could perhaps be used to represent the same latent variable on the interval or even ratio level. When we employ the Guttman procedure we simply do not know whether the proposed latent variable has quantitative structure. Even if we assume that it does, there is no way to determine how distances between persons on the selected test items relate to differences on the latent variable. Perhaps the items were all fairly close together in difficulty, or perhaps the hardest item was disproportionally difficult; we cannot be sure. At best we can say that the sum score on the test is related to the latent variable via some monotonically increasing function. What the exact shape of this function is cannot be determined without further experimentation or extension of the model.

When item responses fit the Rasch model, this allows us not only to order participants and items, but also to interpret differences between them. This is because in the Rasch model differences are considered meaningful. In fact, according to many psychometricians (Rasch, 1960; Brogden, 1977; Perline et al., 1979; Fischer, 1995; Fischer & Molenaar, 1995; Embretson & Reise, 2000; Bond & Fox, 2007) the Rasch model is a probabilistic version of Additive Conjoint Measurement (ACM), a measurement structure described by representational measurement theory (Krantz et al., 1971). For different types of measurement structures this formalized theory specifies a set of axioms that ensure measurement at a certain level is possible. According to this theory the assignment of numerals to objects is considered measurement if the numerical relations between the assigned numerals accurately represent (are an isomorphism of) the empirical relations between the objects. The empirical and numerical relational structure and the axioms that ensure the possibility of representation are specified in a representation theorem. If the axioms hold, we still do not know which subset of the infinite number of possible nu-

merical assignments actually accurately represents the empirical structure. This question is answered in the uniqueness theorem, which specifies what level of measurement is possible by stating the type of transformation that can be used to translate all isomorphic (structure-preserving) numerical assignments into each other (e.g., strictly increasing monotone transformations denote ordinal level, linear transformations denote interval level).

In ACM the empirical relational structure is made up of two disjoint sets of objects and their Cartesian product (Luce & Tukey, 1964). In our case the two sets of objects are items ($I = \{a, b, c, \ldots\}$) and persons ($P = \{x, y, z, \ldots\}$). The Cartesian product is the pairing of each item a with each person x, resulting in the probability (a, x) that the person will endorse the item. The empirical relation between these objects is a weak ordering on the latent property of interest denoted by $\succcurlyeq$. The numerical relational structure that can represent this relation between items, persons and probabilities, consists of the reals, and the numerical relation $\geq$. The ACM uniqueness theorem specifies that items, as well as persons and their pairings (i.e. the item response probabilities) can conjointly be represented on the interval level. We can now address the truly interesting elements of the representation theorem: the axioms. For sake of brevity we only give a very general description. A full treatment is provided in Appendix B.

ACM specifies five axioms, of which the first three are directly empirically testable. The first axiom requires that the ordering of any two probabilities is a weak order and is consistent with the ordering between persons or items. If one person has a higher probability of getting an item right than another person, then the ability of the first person needs to be higher than the ability of the second (idem for items and persons reversed). In the Rasch model the probability is fully determined by the difference between person ability and item difficulty. The only way a probability on a given item can be higher is if the ability of the person is higher. If data fit the mathematical structure of the Rasch model, they will always satisfy this axiom of weak ordering. The same goes for the second axiom of independence. This axiom requires that the ordering between any two persons remains the same, no matter what item you compare them on (idem for items and persons reversed). Since the Rasch model produces non-intersecting, monotonically increasing item response curves this axiom cannot be violated. The third axiom, called double cancellation, requires a more complicated form of consistency, comparing more than two probabilities at once. This axiom will be

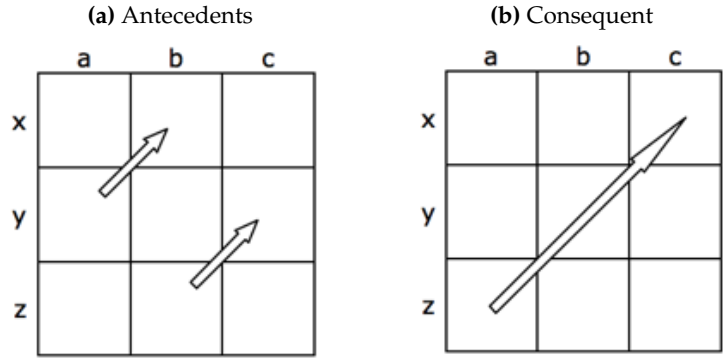
discussed more extensively in the next section. Here it suffices to say that the Rasch model meets this axiom for the same reason it meets the first two: its mathematical form simply incorporates it. The same can be said for the last two axioms. The solvability axiom requires that there is no gap in our set of items, persons or probabilities. If we know the probability (a, x) of persons x answering item a correctly and we have a different person y, then we have to be able to find an item b that will give us the same probability as (a, x) (idem for items and persons reversed). Because both the probability and the latent variable on which the items and persons vary are continuous, this axiom will be met if we imagine we can always find or construct an item of any difficulty or a person of any ability. A similar argument applies to the Archimedean axiom that requires no difference between persons or items is infinitely small. This axiom will also be discussed in more detail in the next section.

We see that the Rasch model meets the requirements for interval measurement. We have also seen that the Guttman model allows for perfect ordering of persons and items. What remains somewhat vague however is why exactly the Guttman model does *not* allow for the representation of quantitative structure. An obvious question therefore presents itself: if the only difference between the Guttman and Rasch model is the presence of error, then why does the Guttman model fail to meet the axioms of ACM? One would expect that the removal of error would result in a model with at least the same measurement level and a higher precision. Examining the answer to this question will provide more insight into the true nature of the paradox. We will therefore evaluate two axioms that prove to be especially problematic for the Guttman model, namely the double cancellation and Archimedean axiom.

4.4 Why Guttman Fails to Yield Interval Measurement

A special case of the double cancellation axiom, the Luce-Tukey condition (Luce & Tukey, 1964; Michell, 1988), will be considered first. The axiom is represented graphically in Figure 4.2, where the Cartesian product of items $\{a, b, c\}$ and persons $\{x, y, z\}$ is displayed. The Luce-Tukey condition requires the resulting probabilities at the tails of the arrows to be compared with the probabilities at the arrowheads in Figure 4.2a. If the probabilities

Figure 4.2: The Luce-Tukey condition



The Luce-Tukey condition is met if and only if it can be shown that for any 3x3 sub-matrix, if the two combinations of pairings in the left panel are both weakly ordered ($\succcurlyeq$), the combination of pairings in the right panel is also weakly ordered in the same direction.

both show the weak order relation $\succcurlyeq$ then this relation should also be observed between the probabilities at the tail of the arrow and arrowhead in Figure 4.2b. More formally, if and only if $(y, a) \succcurlyeq (x, b)$ and $(z, b) \succcurlyeq (y, c)$, then $(z, a) \succcurlyeq (x, c)$. This rather complicated requirement represents the need for consistent ordering. If the two antecedents (Figure 4.2a) show the same relation for the probabilities this means that the rows (persons) and columns (items) are ordered consistently (although not necessarily in the same direction). If this is the case then for the ordering of the probabilities to be consistent, the consequent (Figure 4.2b) must demonstrate the same relation as the antecedents. One may think there are many more combinations of antecedents and consequents required for consistent ordering of the probabilities, but these all logically follow from the axiom of independence (Michell, 1988; see Appendix B for a full treatment of this axiom).

What does the double cancellation condition entail when we pair persons with items for the Rasch model and Guttman model respectively? In Table 4.1 an example of probabilities for each model was given. When data perfectly fit the Rasch model, the probabilities resembling a pattern in Table 4.1b will directly show acceptance of double cancellation for all possible three-by-three sub-matrices. This is because in the Rasch model, the probabilities are a continuous, monotonically increasing function of the difference between the person ability and the item difficulty and because the item re-

Table 4.2: Two problematic Guttman patterns

(a)

	$a_{(4)}$	$b_{(6)}$	$c_{(2)}$
$x_{(5)}$	1	0	1
$y_{(1)}$	0	0	0
$z_{(3)}$	0	0	1

(b)

	$a_{(2)}$	$b_{(4)}$	$c_{(5)}$
$x_{(6)}$	1	1	1
$y_{(3)}$	1	0	0
$z_{(1)}$	0	0	0

Rows (x,y,z) denote persons, columns (a,b,c) denote items. The probabilities in the cells result from the pairing of persons and items. The subscripted integers indicate the ordering of persons and items that determines whether the probability is 1 or 0.

sponse curves are stochastically ordered. This automatically results in the more complicated consistent ordering required by the double cancellation axiom[1]. If Guttman probabilities (Table 4.1a) are put to the test however, double cancellation will be violated in many of three-by-three matrices that are possible.

Two problematic patterns are displayed in Table 4.2. In both cases the antecedents show an equivalence relation ($0 \sim 0$ or $1 \sim 1$) which conforms to $\succcurlyeq$. In contrast, the consequent shows a simple order relation ($1 \prec 0$) that contradicts the weak order found in the antecedents. In the Rasch model, differences between persons and items will always result in differences between the associated probabilities.

In the Guttman model this is not necessarily the case. Two persons differing greatly in their ability still have the same probability of 1 for all the items that the person with the lower ability dominates and the probability of 0 for all the items that dominate the person with the higher ability. The lumping together of persons or items in terms of probabilities leads to the rejection of double cancellation in the Guttman model.

[1]Not everybody accepts this view. Kyngdon (2008a) argues that probabilities in the Rasch model can satisfy the order axioms (see Appendix B) but not necessarily the double cancellation axiom. If this were the case however, the model would not fit. Of course a perfectly fitting model will never occur in practice, but we are interested here not in practical limitations of the model, but in the structural difference between the Rasch and Guttman model, which is what the Guttman-Rasch paradox is about. The fact is that probabilities in a perfectly fitting Rasch model will always conform to double cancellation.

The Archimedean axiom requires that no difference between objects in either set is infinitely small. In other words, it ensures that no item difficulty or person ability is infinitely small or large. This cannot be tested empirically but we can examine it by imagining we have an infinite supply of items and persons. Formally it requires one to order items and persons and denote this order by using natural numbers. Now if you pair person x_i with item a, and you then take the next person x_{i+1}, it should take a different (more difficult or easy) item b to get the same result. In the Rasch model, this is clearly the case. If a person with a certain ability answers an easy item, a person with a slightly higher ability would have to answer a more difficult item to get the same probability of answering this new item correctly. In the Guttman model this is not the case. If a person with a certain ability has a probability of 1 to answer an easy item correctly, a person with slightly higher ability will have exactly the same probability of answering the item correctly, an essential feature of the Guttman model. It does not take a different (more difficult) item to get an equivalence relation between the first person-item pairing and the second. The Guttman model therefore also fails the Archimedean axiom.

The reason Guttman fails to conform to ACM is again due to the lumping together of different persons or items. The Archimedean axiom most clearly shows that the objects of measurement need to vary in a continuous manner. In the case of ACM the objects are not only items and persons but also probabilities. The discrete nature of the Guttman probabilities precludes measurement at the interval level. This can also be understood as follows: the continuous probabilities in the Rasch model that are each associated with a specific ability and difficulty value can only be estimated by using information from other persons and other items. However, once they are estimated the probability of any item contains information on the probabilities on all the other items. Once we know someone's probability on one item, we know the ability. Combined with knowledge of item difficulties, we know this person's probability on all other items we can imagine, provided they conform to the same model. In the Guttman model this is not the case. Knowing the probability of 1 or 0 for a certain item does not give information on the probability on all other items. If the probability is 1, we know that this person will answer all easier items correctly, but we have no idea what this person's probability is on more difficult items (and vice versa for probability of 0 and easier items).

This is not to say it is impossible that the Guttman model can be associated with interval level measurement. One way this can be conceptualized is by considering the sum score on the test instead of the response probabilities separately. The number of items answered correctly might be able to provide more than just the ordering of persons and items if we consider a different representational structure and add some assumptions concerning random selection from an infinite pool of items or persons. However such an approach is so fundamentally different from the ACM framework where items and persons are conjointly measured that we would be comparing apples and oranges. The Guttman model would have to be considered a very different measurement procedure. In that case a difference in measurement level would no longer be considered unexpected or paradoxical.

4.5 Rasch Is Guttman Plus Error

We have seen that the essential property responsible for the difference in measurement level between the Guttman and Rasch models is the discrete versus the continuous nature of the probabilities incorporated in the respective models. This suggests that error is perhaps not the only or at least not the relevant difference between the models. One could argue however that the two are inextricably linked and therefore the distinction is immaterial. We will consider whether this argument has merit. To do so we examine in more detail the proposition that the Rasch model is equivalent to the Guttman model with the addition of error.

In psychometric theory, error is defined as the deviance from the expected response (Lord & Novick, 1968). Concerning the Guttman model we can be brief. In this model, measurement error is absent. The expected response is either one or zero, and the deviance from this score is always zero (see Figure 4.1a). This makes the model very hard to fit in practice. For example, it takes only one very paranoid person who endorses an item indicative of low paranoia for the model to fail. To describe such data more accurately an obvious approach is to allow for some form of measurement error in the model. There are at least two different ways to incorporate error in the Guttman model. In the first approach it is assumed that deviations from the expected response are independent of the ability of the subject. For instance, in the Proctor constant error rate model (Proctor, 1970), the probability of

a deviation from the expected response, ϵ, is constant. It does not matter whether the ability is greatly or only slightly above (or below) the item difficulty. The simplest version of this model is represented in Figure 4.1b. In variants of the Proctor model, ϵ may depend on the item or on the value of the expected score, but it will not depend on the ability, denoted by θ, in a continuous way.

In the second approach, it is assumed that the probability of a deviation from the expected response does depend on the ability level. In these models it is essential that the probability of a deviation decreases monotonically as a function of the difference between ability of the subject and difficulty of the item. Put more formally: $\epsilon = f(|\theta - \beta|)$ where $f(a) > f(b)$ for all $a > b$. Another way to state this assumption is that the item response functions have to be monotonically increasing in θ. The Rasch model, with its item response curve represented in Figure 4.1c, fits into this second category.

We now consider the relation between the Guttman and Rasch models in terms of IRT parameters. The Rasch model contains measurement error that one would prefer to minimize. In IRT models the relevant parameter is the discriminatory power of an item, which is represented by the slope of the item response curve. Items with high discrimination, i.e. steep slopes, differentiate well between persons in a specific range of the latent variable. In the Rasch model the slope of the curve, i.e. the discrimination of an item is not a parameter but a constant. Items are all required to have the same slope or discrimination. However, we could imagine changing the wording of the items or the testing procedure to make all items more discriminating. In this light, reconsider for a moment the curve in Figure 4.1c. For any given ability value, the error becomes smaller when the slope of the curve, denoted by α, becomes steeper. A good test therefore consists of items with high discrimination[2].

Suppose we apply this adage in the extreme. If we keep improving the items by removing error and therefore make them even more discriminating, the paradox emerges. The slope of the response curves become so steep that in the limit our Rasch items turn into Guttman items and we are left with an

[2]It is of course also important that the items vary in difficulty in the range of interest. High ability participants who are presented with items that strongly discriminate, but only on a much lower ability level, will answer all the items correctly, resulting in very low measurement precision (in terms of item information, not retest reliability).

ordinal scale. Thus, by removing error from our items we lose the interval level. Stated in reverse: to achieve interval level measurement, all we have to do is add error to the deterministic Guttman model. This comes across as a Baron von Münchhausen adventure. We pulled ourselves out of the swamp of measurement problems using error as bootstraps. Stated yet another way: we have observed a planetary system where there previously was none to be observed, just by fogging up the lens of our telescope!

Our unease with this paradoxical effect of error relies heavily on the assertion that error is the only difference between the two models. Mathematically it is true that in the limit we arrive at the Guttman model if we let the discrimination of the items approach infinity in the Rasch model. In this sense the difference between the two models is indeed error. However, we saw earlier that the use of discrete versus continuous probabilities is perhaps a more relevant difference between the models. If we consider the different ways error can be added to the Guttman model we see that this discrete versus continuous characteristic is indeed more relevant and not an immaterial by-product of error per se. One cannot achieve a higher measurement level by adding just any error. The Proctor constant error model for instance can also be considered Guttman plus error. However, there is no famous Guttman-Proctor paradox. This is of course because the error added to produce the Proctor constant error rate model results in discrete probabilities. It is not error in general that allows for interval level measurement, but a very specific characteristic of the special type of error incorporated in the Rasch model, namely the continuity of probabilities and their systemic dependence on the latent ability, that allows for the specification of a monotonically increasing relation between items and persons.

4.6 Error that Improves Precision

Although we have shown that error in itself is not responsible for the difference in measurement level, it remains counter-intuitive to introduce continuity by letting it ride piggy-back on the shoulders of a highly specific type of error. One would think error, no matter how special, cannot bear information concerning the property we measure – it is merely a nuisance that restricts our measurement precision. This last proposition plays an important role in perpetuating the Guttman-Rasch paradox. However, the proposition is not

only irrelevant, since it focuses on error instead of continuity, it is also false. Even though untutored intuition has it that measurement error should always decrease the quality of our measurements, there are in fact a number of well-known cases where adding noise to a measurement process is beneficial. This phenomenon is generally referred to as stochastic resonance, a term used to describe the enhancement of a signal by adding noise in certain nonlinear systems. It occurs when a weak signal, often in bistable systems, is amplified by noise. Stochastic resonance was originally proposed to account for the periodicity of ice ages, but has since been shown to occur in many different systems (Gammaitoni, Marchesoni, Menichella-Saetta, & Santucci, 1989; Wiesenfeld & Moss, 1995).

An example that is particularly interesting in the context of the disappearing planets analogy concerns the improved detection of faint stars by adding noise. These stars can be detected better against a slight background fog than against a perfectly black night sky (Dunlop, 1991). Another example of stochastic resonance is the enhancement of audio signals perceived with and without hearing aids by adding white noise (Zeng, Fu, & Morse, 2000; Chatterjee & Robert, 2001; Ries, 2007). In a similar vein, the perception of water movement in the mechanoreceptor cells of crayfish was enhanced by randomly perturbing these receptor cells (Douglass, Wilkens, Pantazelou, & Moss, 1993).

A process related to stochastic resonance is dithering. In digital signal processing repeated quantization of analog signals may lead to correlated errors in the resulting signal, leading to large artifacts. By adding dither (noise), to the quantization process, correlated errors are avoided, and better results are obtained (Schuchman, 1964). Optimization provides another example. In simulated annealing, global solutions are found, and local solutions are avoided, by making many non-optimal steps in the initial phase of the search (Kirkpatrick, Gelatt, & Vecchi, 1983). These examples show that the relation between measurement error and precision is a complicated one. Noise can apparently contain information on the latent property, which means there has to be some systematic relation between the error and the latent variable. To illustrate how such a relationship can be conceptualized, we present a simple example inspired by Lumsden (1976).

Suppose we want to measure the length of a number of people, say A, B, C and D with lengths of 160 cm, 165 cm, 190 cm, and 195 cm. Further-

more, suppose we – being social scientists, lacking the rigorous measurement instruments of physics – have no clue about the actual lengths. We therefore resort to a procedure that is familiar to us, the application of pairwise comparisons. Because we are aware of the high risk of measurement error we repeat our comparisons several times, just to be sure. If we let a young PhD student, of sufficient height and with good eyesight, repeatedly compare the persons in a precise way, D is always larger than C, B and A, C is always larger B and A, and B is always larger than A. This results in an ordinal Guttman scale. Now we ask a somewhat older, farsighted professor to repeat the procedure. Unfortunately this professor forgot her glasses and makes several mistakes. Obviously, she makes more errors when she compares A and B and when she compares C and D, because these are close together and therefore harder for her to judge. For the same reason, errors are less probable when she compares B and C. This gives her quantitative information. Because she makes more errors when comparing A and B, and C and D she can infer that the differences between A and B and between C and D are small relative to the difference between B and C (and A and C and B and D). She can even make a quantitative estimation of these differences, which may be a relatively accurate estimation of the real differences. Here the paradox reappears: the farsighted professor achieves an improvement in precision by making mistakes.

A simulation of this example will illustrate the beneficial effect of error more clearly. Persons A, B, C, and D (respective lengths 160, 165, 190 and 195 centimeters) were compared. For each unique comparison a point was awarded to the longest person, resulting in a number of 'wins' for each person. To simulate measurement error, normally distributed noise was added to the actual difference in length for each comparison. The standard deviation of the distribution providing the noise varied from 0 to 30. The comparisons were repeated 1, 2, 5, 10, 20 or 50 times. Each of these 'trials' that varied in amount of error and number of repetitions, was simulated 500 times. To see whether the added error resulted in a better or worse representation of the length of persons A, B, C and D, the root mean square error (RMSE) of the number of wins and the actual length (both scaled to z-scores for easier interpretation) was calculated for each trial. Since the actual length and number of wins were both scaled to have a mean of zero and a variance of one, the RMSE will be zero if the number of wins perfectly represents the length and will increase as the number of wins becomes a less perfect linear

representation of the actual length. Figure 4.3a shows that adding noise results in a clear decrease in the RMSE, which means that precision improves. If we choose a large number of comparisons, we may even allow for high levels of noise. The optimal level of precision in this simulation is reached when error with a standard deviation between 5 and 10 is added.

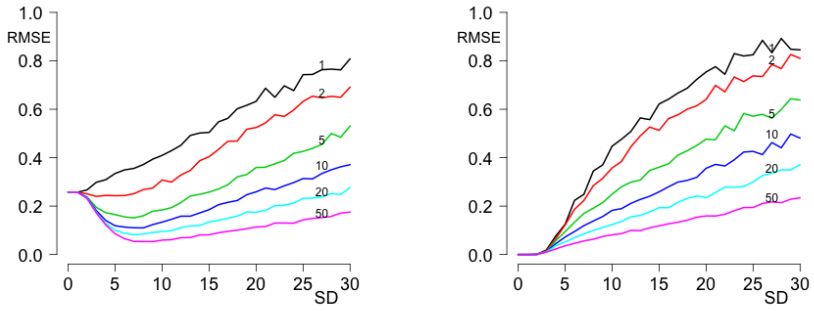
Whether the beneficial effect of error appears depends on many things, among which the measurement procedure and the distribution in the sample. If the simulation is performed with the same scoring-procedure but with lengths of 160, 170, 180, and 190 centimeters, adding error does not improve measurement precision (i.e., does not decrease the RMSE). The results are presented in Figure 4.3b. This result is due to the way we set up our measurement procedure. If, instead of looking at the number of wins per person, we compare each win or loss (scored 0 or 1) per unique comparison to the difference in actual length, the distribution of lengths becomes irrelevant. The results of the simulation using this alternative scoring method are presented in Figures 4.3c and 4.3d for the non-equidistant and equidistant lengths respectively. This time the RMSE decreases (i.e. precision improves) in both the equidistant and the non-equidistant case when error is added, albeit only when we repeat our procedure many times. The positive effect of error is much more dramatic, especially with many repetitions. The overall precision of the alternative scoring method is much lower (RMSE is higher) than that of the original method however, because it is much more indirect. Of course indirect methods like paired comparison are often the only ones available to represent latent properties in psychological research.

The point of these examples is to show that error, when dependent on the latent property, can help to increase measurement precision. Whether it does so in a particular research situation depends on the presence of quantitative structure in the latent property, on what measurement procedure was used, on the amount of error in the procedure, and in some cases on the spacing of objects in the sample. Although the effect of error on precision is probably only favorable in very few cases, the main point here is to establish a proof of principle: a positive effect of error on precision is possible. A favorable effect of error on precision should therefore not necessarily be considered paradoxical. As a result, our untutored intuition that error should not improve precision cannot bolster a general argument against the claim that the Rasch model yields interval level measurement. Since the conditions for a favorable effect of error to occur are very restrictive, and since the example

Figure 4.3: The measurement of length through paired comparison

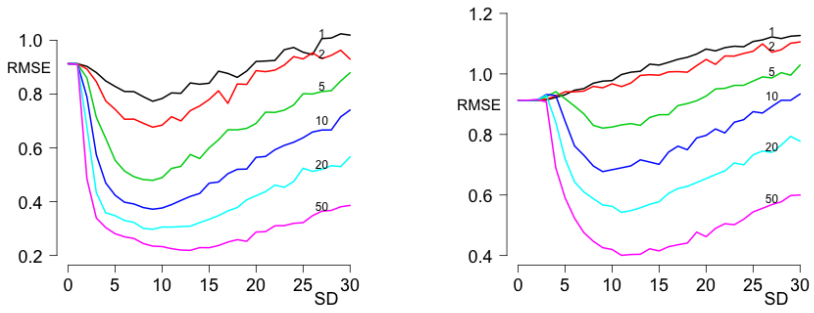
(a) Non-equidistant, scoring of wins per person

(b) Equidistant, scoring of wins per person



(c) Non-equidistant, scoring of wins per comparison

(d) Equidistant, scoring of wins per comparison



The precision of measuring length of four persons using paired comparison is ascertained. Non-equidistant refers to a set of persons with actual lengths of 160, 165, 190 and 195 cm; Equidistant refers to a set of persons with actual lengths of 160, 170, 180 and 190 cm. SD indicates the standard deviation of the normally distributed error added to comparisons. The lines represent different number of repeated comparison. A score of one can either be awarded per person or per comparison. The RMSE of the score and the length (difference in case of scoring per comparison) indicates measurement precision. A lower RMSE value indicates higher precision.

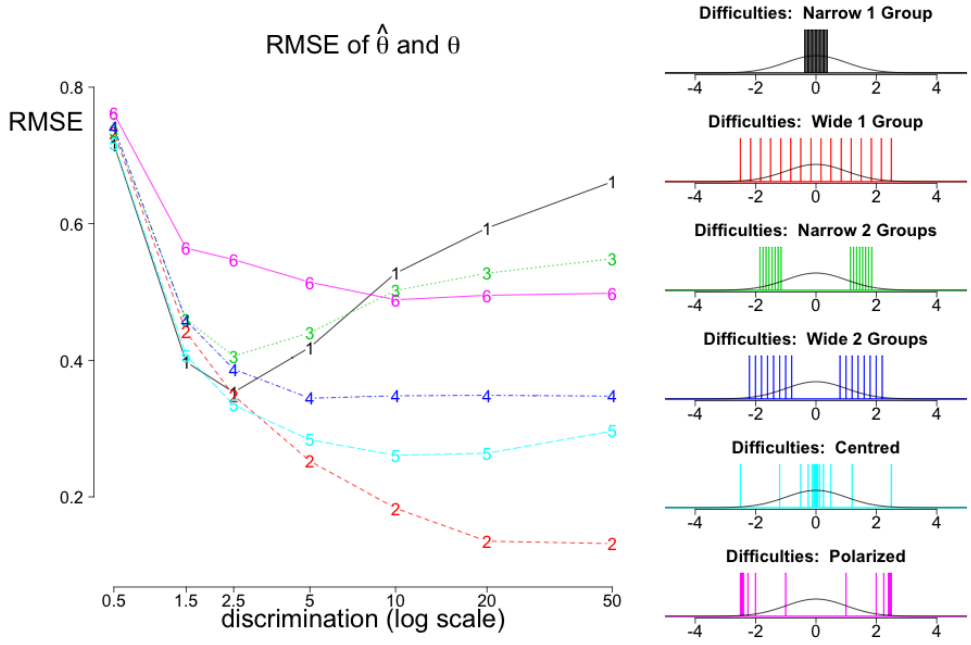
– while informative – does not match typical applications of IRT models we present a third simulation that directly involves the Rasch model.

4.7 Error and Precision in the Rasch Model

The setup of the simulation study is simple. Item responses to 16 items were generated according to a Rasch model for 500 subjects. Person ability values were sampled from a standard normal distribution. The level of discrimination was varied systematically from very low (.1) to very high (50). For each level of discrimination, the simulation was repeated 50 times. Increasing discrimination corresponds to the removal of error, making the item more like a deterministic Guttman item. In terms of the previous example, where a forgetful and farsighted professor made comparisons between the length of persons, increasing the discrimination parameter would correspond to providing her with a borrowed pair of glasses: the closer the prescription is to her own glasses, the fewer errors she will make.

In the previous simulation it turned out to be important whether the persons were spaced equidistantly. This makes sense from the Rasch perspective. It is not just the discrimination that influences precision when pairwise comparison is used. The spacing of the items is also important; when items are highly discriminating but much too easy, precision will be low for persons with higher ability and therefore precision for the entire test will suffer. Therefore the spacing of the item difficulties was investigated by generating six item difficulty distributions. The degree to which a distribution adequately covered the distribution of abilities was varied first of all by generating a very narrow and a very wide distribution (as compared to the ability distribution). Both these distributions were equidistantly spaced and consisted of a single group of items centered around the mean ability. Secondly we were interested to see how precision would be affected if there was a gap in the coverage of the ability scale. Therefore we generated distributions consisting of two separate groups of equidistantly spaced items. Again, a narrow and a wide distribution were generated. Finally we were interested to see how precision in the Rasch model would be affected if the items were not equidistant. To this end we generated non-equidistantly spaced items concentrated at the mean of the ability distribution and non-equidistantly spaced items concentrated at the tails of the ability distribution. The six resulting distributions are displayed in the right panels of Figure 4.4. The coverage of the ability scale can be assessed by comparing the item distribution to the curve in each panel, which represents the ability distribution of the simulated subjects.

Figure 4.4: The effect of error on measurement precision in the Rasch model



The effect of error on measurement precision, in terms of RMSE is displayed for the Rasch model with 16 items, under varying amounts of error and item difficulty spacing. Higher item discrimination indicates lower error. The different item difficulty distributions are shown in the right panels. A lower RMSE value indicates higher measurement precision.

From the simulated data, the maximum a posteriori (MAP) estimate of ability was computed[3]. As before, the root mean square error (RMSE) was calculated, this time of the estimated and original ability values. This indicates how well our ability estimates represent the original ability values. A decrease in RMSE signals an improvement in precision. In Figure 4.4 the mean RMSE of the 50 iterations is reported for each value of the discrimination and item difficulty distribution.

[3]This estimate is relatively easy to compute because the difficulty parameters are known from the simulation setup. That is, we did not need to fit the Rasch model using standard estimation algorithms, which is problematic for data simulated with very high discrimination parameters.

From Figure 4.4 we can derive the following important results. First, in the normal range of discrimination values (below 5), increasing the discriminatory value of items is obviously beneficial to measurement precision, since the RMSE decreases. However, in some cases a further increase in discrimination results in an increase in RMSE (case 1: one narrow group; case 3: two narrow groups; and to a much lesser extent case 5: centered). This means that the removal of error until it is almost entirely absent leads to a decrease in measurement precision. Just as in the previous simulation of pairwise length measurement, precision is optimal when a certain amount of error is *present*, not when error is absent. This allows us to conclude error can also enhance measurement precision in this more realistic Rasch set-up. However, this effect only occurs under special circumstances, when the spread in item difficulties is minimal (relative to the distribution of abilities) or contains gaps. The beneficial effect of error on precision disappears when items are reasonably well spaced over the scale of the latent property.

These simulations have shown that in some cases error can improve precision. In the natural sciences this effect is well-known and sometimes even actively employed to enhance measurement precision. The proposition that a favorable effect of error on precision is paradoxical therefore cannot form a general, conclusive argument against claims of interval level measurement by the Rasch model. However, in the Rasch simulation we found that the beneficial effect of error only appears under very select circumstances and that precision can go both up or down as error increases. This brings attention to another problem with the proposition that error should not improve precision. This problem concerns the focus on precision.

The assertion that we lose 'precision' by removing error from the Rasch model is unhelpful to say the least. We in fact gain precision in the sense that we are now able to observe order perfectly, without fault. If we interpret 'precision' as 'the proximity between the observed and latent scores' then precision is gained or lost, depending on the circumstances. Only if we take 'precision' to mean 'information concerning quantitative structure', is Michell's assertion correct. Interpreted as such, we indeed lose precision, but this is not an unfortunate side-effect of our noble effort to minimize error. We have purposefully discarded useful information; a drop in measurement level is therefore to be expected and can hardly be considered paradoxical.

4.8 Discussion

In our analysis of the Guttman-Rasch paradox, and its use as an argument against the interval level claim by advocates of the Rasch model, we have shown three things. First, we focused on the measurement pretensions of both models in general and the requirements for interval level measurement in particular. We showed that the Guttman model fails these requirements because it employs discrete probabilities where the Rasch model uses continuous probabilities. It is not the removal of error per se, but the transformation of the Rasch probabilities into discrete ones and zeros that prevents us from accessing the quantitative structure that is present in the underlying variable. The assertion that the removal of error from the Rasch model obviously presents an overly simplified picture of the relation between the two models.

Second, to show that our focus on continuity indeed constitutes a meaningful distinction relevant to understanding the paradox, we investigated the difference between the models in terms of error in more detail. The proposition that the Rasch model is equivalent to the Guttman model with the addition of error should also be considered too general a statement. Moving between the Guttman and Rasch model involves more than just the general addition or removal of error. Error can be added to the deterministic Guttman model in many different ways, not all resulting in interval level measurement. Only the special type of error incorporated in the Rasch model, characterized by its continuous nature and relation to the latent variable, produces this change in measurement level.

Third, the proposition that error cannot improve precision was shown to be unfounded. There are indeed cases where error can enhance precision. Using three simulations it was demonstrated how such a beneficial effect of error can arise. Error can provide more information about the latent property when the expected number or size of errors is related to the size of intervals or values on the latent property. In such cases, removing error can non-paradoxically have a negative effect on precision. This result was also found for the Rasch model, but only when the items are spaced very awkwardly. When the items cover the entire range of the latent property with no big gaps in between, removing error generally improves precision. Error and precision are sometimes monotonically related and sometimes not,

depending on the circumstances. This has little to do with measurement level however. Phrasing the Guttman-Rasch paradox in terms of precision deflects our attention from the question that actually requires explanation, namely why removal of error results in a drop in measurement *level*.

We have already seen that the removal, not of error in general, but of continuity is the decisive factor here. We have also seen that error is not as unlikely a candidate to provide information about the latent property as we might have thought. By removing the particular type of error that is embedded in the Rasch model we are throwing away information about the latent property, just as we lose information about the difference in lengths between persons when we provide the nearsighted professor with her glasses. If a person's probability of an error is large, then we know that this person's ability is very close to the item difficulty. By making the probability a discrete 0 or 1, we lose this information. A drop in measurement level due to active dismissal of information, although seeming unusual if formulated in terms of precision, is not paradoxical at all. The Guttman-Rasch paradox should therefore be considered a chimera.

If one wishes to contest the claim of interval measurement by the Rasch model, one needs to present an argument different from the Guttman-Rasch paradox. One avenue is to object to the use of probabilities as objects of measurement on the basis that additive conjoint measurement requires these objects to be spatio-temporally located (Kyngdon, 2008a, 2008b; Borsboom & Zand Scholten, 2008; Michell, 2008a). This objection does not bear directly on the Guttman-Rasch paradox, since both models incorporate probabilities and therefore sin equally against the assumptions of representational measurement theory. However we can see that it is quite a leap to view a probability that can never be observed directly as a form of measurement on par with say, the recording of a person's length. However if we imagine repeatedly administering the test and replacing the probability with a proportion, this leap becomes less extreme. The proportions can now be observed and the requirements placed upon them empirically tested. This does require either a substantive assumption or the effort required to experimentally investigate them. We agree with opponents of the claim of interval level measurement by the Rasch model that such assumptions are made much too easily and experimental testing is sorely lacking. We do believe however that this is a problem that is in principle surmountable, it just has not been addressed yet.

Another objection is that an ACM framework should apply to general sets of objects, not a specific subset of items or persons. The introduction of a new, inconsistent object could easily lead to rejection of the axioms (Kyngdon, 2008a; Michell, 2008a). This argument is supported by the fact that it is not always easy to fit a Rasch model to data. One can imagine that it is tempting to discard items that do negatively affect model fit, even when the substantive reasons for considering the item unidimensionally indicative of the latent property remain unchanged. This goes against the foundations of ACM, which require the specification of a well-defined, and in most cases, infinite set of objects. We agree that claims of interval level measurement are made far too easily for tests that consist of conveniently selected items. This does not mean however that the Rasch model can *never* be associated with interval level measurement.

To conclude our discussion of the paradox we return to the disappearing planets analogy. In this analogy, the observation of a planetary system, or lack thereof, is compared to the assumption of an interval level, or lack thereof, depending on the presence or absence of error. Similar to the focus on error and precision, this analogy can be considered somewhat misleading. The astronomers suddenly stop observing a physical object when error is removed from their measurement instrument. Psychometricians do not suddenly stop observing qualitative relations because error was removed. They are no longer able to detect the quantitative structure present in these relations because they use a fundamentally different measurement procedure. The Rasch procedure requires the psychometrician to assume a latent property and a specific functional relation between this property and the probability of answering an item correctly. The claim of interval measurement applies to probabilities, which is still considered controversial by some (Kyngdon, 2008a; Michell, 2008a).

If we make use of the Guttman model we do not have to consider probabilities or make strong assumptions. We do have to settle for a discrete measurement procedure however, which is intrinsically different from the Rasch model. We also have to accept that we have to purposefully disregard information about the latent property that is contained within the measurement error of the Rasch model. Psychometricians are therefore better compared to astronomers who can all of a sudden no longer observe a planetary system, not because the initial observation was due solely to error, but because they had to change from using an infrared telescope to using binoculars. These as-

tronomers would certainly not find the disappearance of the planets suspect per se. And even if they did, it would be highly unlikely that they would discard the infrared telescope, blaming it for seeing the planetary system in the first place.