Psychological Test and Assessment Modeling, Volume 57, 2015 (3), 377-395

The validity of polytomous items in the Rasch model – The role of statistical evidence of the threshold order

Thomas Salzberger[1]

Abstract

Rating scales involving more than two response categories are a popular response format in measurement in education, health and business sciences. Their primary purpose lies in the increase of information and thus measurement precision. For these objectives to be met, the response scale has to provide valid scores with higher numbers reflecting more of the property to be measured. Thus, the response scale is closely linked to construct validity since any kind of malfunctioning would jeopardize measurement. While tests of fit are not necessarily sensitive to violations of the assumed order of response categories, the order of empirical threshold estimates provides insight into the functionality of the scale. The Rasch model and, specifically, the so-called Rasch-Andrich thresholds are unique in providing this kind of evidence. The conclusion whether thresholds are to be considered truly ordered or disordered can be based on empirical point estimates of thresholds. Alternatively, statistical tests can be carried out taking standard errors of threshold estimates into account. Such tests might either stress the need for evidence of ordered thresholds or the need for a lack of evidence of disordered thresholds. Both approaches are associated with unacceptably high error rates, though. A hybrid approach that accounts for both evidence of ordered and disordered thresholds is suggested as a compromise. While the usefulness of statistical tests for a given data set is still limited, they provide some guidance in terms of a modified response scale in future applications.

Keywords: Polytomous Rasch model, threshold order

[1] *Correspondence concerning this article should be addressed to:* Thomas Salzberger, PhD, WU Wien, Institute for Marketing Management & Institute for Statistics and Mathematics, Welthandelsplatz 1, 1020 Vienna, Austria; email: Thomas.Salzberger@wu.ac.at

Introduction

The extended Rasch model for ordered categories (Andrich 1978, 1988; Masters 1982) is a straightforward generalization of the Rasch model for dichotomous items (Rasch, 1960). The polytomous responses have to be scored using integers starting with zero as derived by Andersen (1977) and Andrich (1978). For each item, thresholds are estimated marking the location where adjacent response categories are equally likely. The score represents a count of thresholds exceeded by the respondent. In the case of dichotomous items, the raw score as the sum of items responded to positively constitutes valid input to measurement only if all items collectively function as a scale. Otherwise no valid measurement can be inferred. The same is true for polytomous items. However, using polytomous response formats does not only require appropriate item wording (i.e. the sentence stem or the question) but also a meaningful and valid design of the response scale. According to Messick (1995, p.141), "[v]alidity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores". Thus, scores, specifically their meaningfulness and justification, play a crucial role as emphasized by Messick who continues "[v]alidity is not a property of the test or assessment as such, but rather of the meaning of the test scores". The fundamental issue of validity is therefore not confined to the items themselves but needs to take the response format into account, too.

When estimating the parameters in the Rasch model for polytomous items, inferences from ordered category formats are "made as if responses at the thresholds were experimentally independent." (Andrich, 2013, p.73) with responses coming from a Guttman subspace that corresponds to the purported order of categories (see Andrich, 2013, for a detailed explication). Empirical threshold estimates can be in any order, as they follow the data and only depend on relative frequencies of two adjacent categories given that the respondent has chosen one of them. Ordered response categories imply ordered thresholds. Therefore, disordered, or reversed, thresholds indicate some kind of problem in the data that needs to be identified and addressed. It should be noted that reversed thresholds do not necessarily imply a problem with the response scale. Rather, the item wording could be unsuitable or multiple dimensions might be addressed by the items thought to represent a unidimensional scale. These alternative causes of disordered thresholds have to be investigated before considering the response scale itself.

In the following it is assumed that all items do relate to the same latent trait and that any deficiency in terms of validity is related to the response format. Then reversed thresholds suggest that the respondents do not use the response categories as if they were properly ordered. Since a violation of category order is closely related to discrimination at the thresholds, reversed thresholds could, and in many cases arguably do, indicate unequal discrimination. The latter implies that the scoring key is not justified and measures inferred from such raw scores are, strictly speaking, invalid. It should be noted that unequal discrimination does not necessarily result in reversed threshold estimates just as unequal item discrimination in the dichotomous model does not necessarily result in

noticeable misfit. In this respect the present contribution aims at shedding more light on how to gauge evidence in the data in terms of proper response scale functioning.

If threshold estimates are reversed or ordered but unexpectedly close, conceptual considerations are advisable. If all response options are verbalized and serious doubts arise as to a clear increase in the verbal labels, the format should be carefully reviewed. Moreover, qualitative interviews with respondents lend themselves as a source of additional evidence. This is particularly important as what appears to be a logical or objective ordering of response categories need not necessarily match interpretations of respondents. Furthermore, the number of categories may exceed the number of distinctions the respondents are capable of distinguishing. Asking respondents about the appropriateness of the number of categories is also useful if only the extremes are verbalized. Nevertheless, in the end such qualitative input helps shaping a hypothesis about the actual functioning of the response scale but does not represent indisputable evidence. What's more, respondents might not be aware of not being able to discriminate between as many categories as they are offered. Thus, researchers should incorporate evidence gathered during data analysis using the Rasch model into their validity judgements.

While ordered thresholds are universally embraced as being desirable, there has been a long tradition of diversity in interpretation as to what reversed thresholds really signify, whether they actually matter, and how they should be dealt with (see Andrich, 2013, Adams, Wu and Wilson, 2012 for recent contributions). Two issues seem to complicate matters. First, reversed thresholds do not necessarily imply misfit based on tests such as those comparing expected and actual item scores. Second, empirical threshold estimates that are reversed do not necessarily imply that true threshold locations are reversed, too. In the following, both issues will be discussed. Particularly, the potential of formal statistical tests of threshold order are explored.

The Rasch model for measurement and the theory of the construct

The philosophy of Rasch measurement, understood comprehensively, rests on two fundamental principles. First, the model takes precedence over the data as the model prescribes a structure that is deemed essential in order to quantify a latent variable (see Karabatsos, 2001). In the end the model relates counts (Wright, 1992) – observed scores – and numbers thought to represent magnitudes. As any other mathematical model, it is void of any content. Thus, the second cornerstone is a construct theory that is independent of the present data set. Misfit of the data rejects both the assumption that a quantitative latent variable has been measured and the construct theory. In contrast, fit of the data to the Rasch model provides evidence that a latent variable has been measured. However, it does not necessarily fully confirm the construct theory. Construct theories vary in terms of their level of detail. In the simplest case, the theory merely claims, often implicitly, that a number of items collectively form a unidimensional scale. Since such a theory does not make any predictions in terms of the item order, the potential to falsify the theory is limited. A more elaborate construct theory suggests a particular order of items

based on conceptions of what means more or less of the property to be assessed. The idea of the construct map (Wilson, 2005) illustrates this level. An even more specific construct theory exposes the mechanism (Stenner, Stone, and Burdick, 2009) that explains item magnitudes. Such a theory allows for concrete, algebraically derived predictions. Empirical confirmation means strong support for the construct theory. In this context, Stenner, Fisher, Stone and Burdick (2013) refer to Causal Rasch models, which consolidate a formal measurement model and a content theory. Regardless of the type of construct theory that is available in a given instance, when polytomous response scales are used, the functioning of the response format has to be integrated into a comprehensive assessment of validity. Thus, validity is not just a question of psychometric fit of data to the Rasch model. This conclusion is important when considering the relationship between formal fit and the requirement of ordered categories.

Beyond psychometric fit

Ideally, data fit the Rasch model providing evidence that the items form a reasonably unidimensional and essentially valid (Kreiner, 2007) scale. In practice, fit of data to the Rasch model is often only achieved by amending the data, for example by deleting items or splitting items because of differential item functioning. Changing the data on a data-driven basis compromises the confirmatory character of a Rasch analysis. On the other hand, it helps identify avenues for future improvement of the construct theory as well as for advancing the instrument. However, even if all items display satisfactory fit, empirical item location estimates may contradict theory-based expectations. The application of the Rasch model provides the basis for statistical tests, such as comparisons of expected and actual responses or checks of local independence. But these tests of fit are insensitive to violations of the construct theory in terms of the order of item locations. In other words, items can be "disordered" compared to the theory despite showing good fit. This issue needs to be resolved by further research resulting in revised items or an adapted theory depending on the nature of the underlying problem.

Polytomous items and the theory of an ordered response scale

The potential advantages of polytomous response formats are manifold, hence their popularity. Most are related to the increase in measurement precision compared to the same number of dichotomous items. When the spread of item locations is limited, polytomous response formats help extend the range for which items provide information about the person locations. In some instances respondents can feel uncomfortable when being forced to make a dichotomous decision. Polytomous scales allow for intermediate responses. Finally, administering fewer polytomous instead of a higher number of dichotomous items can be more economic and reduce response burden.

As mentioned above, these advantages can only be achieved if the item scores are valid and meaningful. Since the thresholds are located on the same continuum as the overall item locations, the thresholds, and by implication the response categories they separate, are also

supposed to represent more or less of the property to be measured. The possibility to grade one's agreement or disagreement in attitudinal measurement implies that the response options quantitatively modify the item content. When designing a response scale, an instrument developer aims at response options that are typically presented in a strictly monotonously increasing order and, in any event, are scored reflecting the proposed order.

The proposed response format can be thought of as a theory on its own – a theory of ordered response categories. As any other scientific theory, it requires being tested empirically. As mentioned above, an objectively ordered response scale, such as never – once a week – twice a week – three times a week – more than three times a week, while being a promising proposal, does not necessarily imply that actual discrimination will be equal at all threshold locations. Offering too many response categories can result in insufficient discrimination between adjacent categories by the respondents.

The situation bears resemblance to the order of items in terms of their overall locations as suggested by the construct theory. Standard fit statistics are incapable of examining this part of the construct theory. In a similar way, the theory of an ordered response scale needs to be addressed separately. The order of threshold estimates is an important symptom in this regard. There is an important difference, though, between the item order and the threshold order. Items being in an order unexpected under the construct theory will not affect the validity of the scoring function, and, by implication, they have no impact on the fit statistics. By contrast, reversed thresholds due to violations of equal discrimination at the thresholds suggest an invalid scoring function for that item. As a consequence, total scores, person estimates, and fit statistics will be compromised as well. The extent to which this is the case presumably depends on the targeting, the total number of items in a scale, the proportion of items with malfunctioning response scales, and the nature of the improper use of the response format. A detailed investigation of these factors is beyond the scope of this contribution, which focuses on the diagnosis of the problem rather than its causes.

While item fit can certainly be helpful in identifying malfunctioning response scales, general tests of fit do not necessarily seem to be sensitive enough to reliably flag such items. The estimation of model parameters is carried out on false premises as far as the meaning of the item score is concerned and the estimated thresholds ensure best-possible item fit. What is more, disordered categories are compatible with fit to the Rasch model as can easily be demonstrated by simulating data based on reversed thresholds. Therefore, a separate investigation of the response scale is suggested with the empirical order of threshold estimates as the focal symptom of a malfunctioning response scale regardless of item fit.

Scoring dichotomous items

In the Rasch model for dichotomous responses (Rasch, 1960), the scoring is straightforward and only requires theoretical considerations as to which response represents more of the latent variable to be measured (scored one) and which less (scored zero). Applying the wrong scoring key, either by mistake or because of a serious misconception, would result in obvious misfit (see Figure 1 for an exemplar, simulated data).

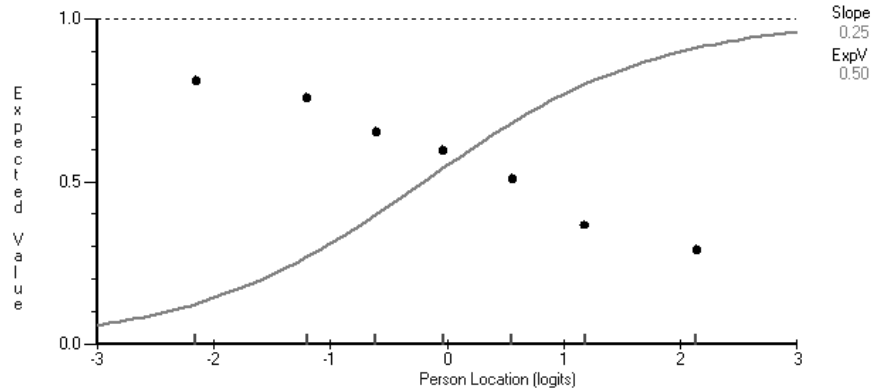


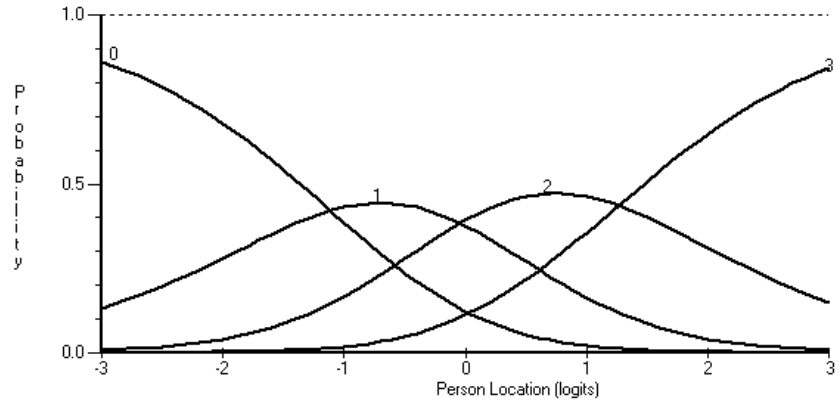
Figure 1:
Theoretical and Empirical Item Characteristic Curve of a Wrongly Scored Dichotomous Item

Scoring polytomous items

Polytomous response scales allow, or require, a grading of the response. Attitudinal measurement typically makes use of rating scales asking the respondent to grade agreement versus disagreement. Assume a four-point response scale offers the response options completely disagree – somewhat disagree – somewhat agree – completely agree. Such scales are scored using successive integers (see Andersen, 1977, and Andrich, 1978, for the derivation of the scoring function in the RM). In this case the scoring key would be 0-1-2-3. In the polytomous Rasch model (Andrich, 1978, 1988), such a four-category item is represented by a set of three threshold parameters, τ_1 to τ_3, marking the boundaries between successive response categories. The overall item difficulty δ_i is the mean of all thresholds and marks the point where the lowest and the highest response category are equally likely. The scoring reflects the assumption that choosing a higher response option implies more of the property compared to choosing a lower category. A respondent is expected to most likely choose the category that corresponds to the respondent's location. If the threshold estimates are ordered, a respondent β_j is – disregarding responses to other items in the instrument – most likely located in one of four sections of the continuum (see Andrich, 2013): $\beta_j < \tau_1$ (indicated by a response of 0), $\tau_1 < \beta_j < \tau_2$ (response of 1), $\tau_2 < \beta_j < \tau_3$ (response of 2), and $\beta_j > \tau_3$ (response of 3). Conversely, a respondent who is located in one of the four sections is most likely to choose the response option that indicates that section (see Figure 2, panel A). If the threshold estimates are reversed, as in Figure 2, panel B, this is no longer the case, as in the center of the scale a response either in the first category, scored 0, or in the fourth category, scored 3, is more likely than a response in either of the intermediate categories 1 or 2. In fact, the latent continuum appears to be split into only two sections: $< \delta_i$, and $> \delta_i$. As a result, when comparing a respondent $\beta_j < \delta_i$ with a respondent $\beta_k > \delta_i$, it is very likely that the latter earns three credits more than the former. As a consequence, the item characteristic curve (ICC) for the item with reversed thresholds is steeper than the ICC for the item

with ordered categories (see Figure 3). If we were to score a dichotomous item 0 versus 3, we would witness a similar, if more extreme effect. Thus, reversed thresholds indicate that the response scale does not work as intended and that the scoring is probably unjustified.

Panel A: Ordered Thresholds



Panel B: Disordered Thresholds

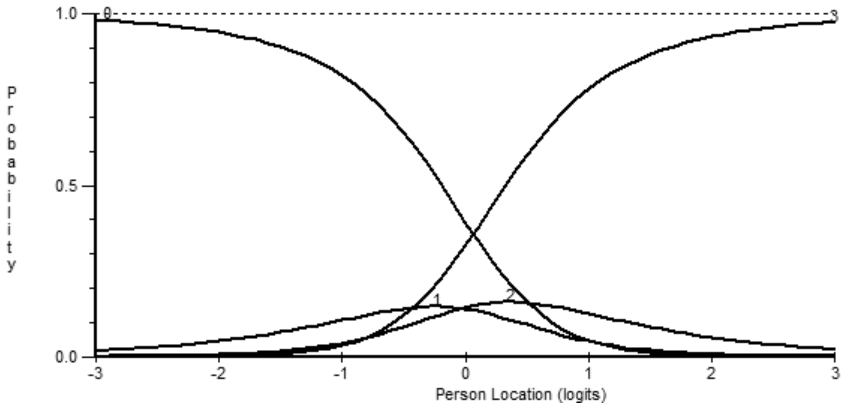


Figure 2:
Category Characteristic Curves of a Four-Category Item with Ordered and Disordered Thresholds

Figure 3:
Item Characteristic Curves (ICCs) of Two Four-Category Items with Ordered (flatter ICC) and Disordered Thresholds (Steeper ICC)

Improper scoring of polytomous items

The scoring key for polytomous items can be inappropriate for different reasons. First, in the trivial case, the scoring can be exactly reversed, for example due to ignoring true item reversal when disagreement signals more of the property rather than less. The empirical item response function would then be decreasing (see Figure 4, simulated data where responses were scored 3-2-1-0). The complete reversal could be seen as the most extreme example of misrepresenting the actual order, which is discussed next.

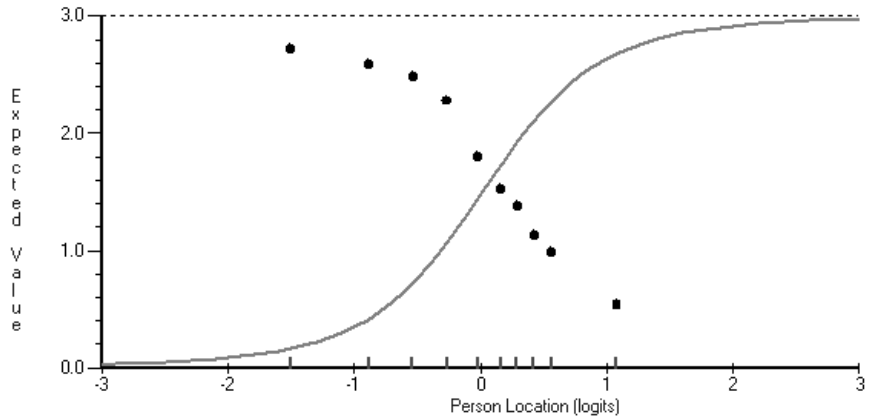


Figure 4:
Item Response Function of a Wrongly Reversed Scored Polytomous Item

Second, response categories can be ordered but the scoring does not reflect their order correctly. For example, the third category can be harder than the fourth. This situation can be simulated by scoring the responses 0-1-3-2.The empirical item response function would not be monotonously increasing (see Figure 5, simulated data).

Third, the respondents may fail to discriminate between adjacent categories, i.e. decide between two categories on a random basis, or disregard one or even more categories and rather decide between the next lower and the next higher category.

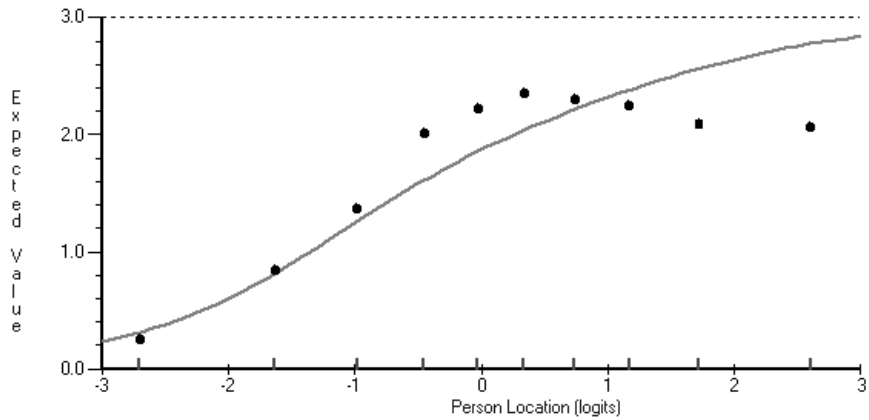


Figure 5:
Item Response Function of a Wrongly Scored Polytomous Item

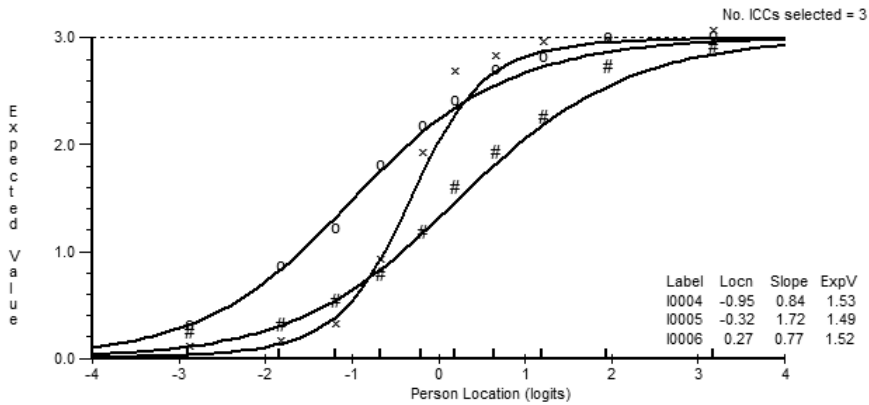


Figure 6:
Item Response Function of a Polytomous Item with Reversed Thresholds

Identifying malfunctioning polytomous items

The first two cases mentioned above are rather trivial. Fit statistics will alert the analyst, probably even when using small samples. In addition, conceptual considerations should easily identify the nature of the problem, which can be resolved by reversing the scoring key in the first case, and amending the scoring key in the second case.

The third case is more problematic, though. The response categories might conceptually appear to be meaningful and to represent increasing amounts of the property to be measured. But if, for example, the respondents only differentiate between agreement and disagreement while the response scale provides four options (as outlined above), only one threshold will properly discriminate. The item essentially provides as much information as a dichotomous item. However, agreement would be scored 3 rather than 1 as in a dichotomous item. Such a polytomous item would enhance the difference between two respondents, one agreeing, the other disagreeing, with the same total score on all other items. Thus, the item would appear to properly discriminate and in all likelihood even exhibit a bigger slope of the item response function. Figure 6 shows the steeper item response function for item 5 with reversed thresholds in comparison to the item response functions for items 4 and 6, which both have ordered thresholds (simulated data). In other words, the empirical item response function would be increasing and look perfectly acceptable. Fit statistics that are based on a comparison between observed and expected responses (e.g. the chi-square test of fit in RUMM 2030, Andrich et al., 2009) or fit statistics that assess actual discrimination at the item level (e.g. the fit residual statistic in RUMM 2030) would not necessarily detect serious problems.

Reversed threshold estimates

Disordered threshold estimates represent a symptom for category disorder that is independent of fit, i.e. they might occur regardless of the degree of item fit. Since estimates of thresholds are, as all parameter estimates, imperfect, it seems worth considering their standard error and making use of statistical inference. The estimates of truly ordered thresholds may be empirically disordered because of random variation in the responses. However, the reverse may also be true, i.e. truly disordered thresholds may be accidentally ordered.

Inference of the threshold order

For the sake of simplicity, in the following we consider the simplest polytomous item with just three response categories *i-1* (scored 0), *i* (scored 1) and *i+1* (scored 2). τ_i is the threshold between response categories *i-1* and *i*, while τ_{i+1} is the threshold between response categories *i* and *i+1*. $\hat{\tau}_i$ and $\hat{\tau}_{i+1}$, respectively, represent the empirically estimated thresholds. A properly functioning response scale requires that the true threshold locations are ordered: $\tau_{i+1} > \tau_i$. If $\tau_{i+1} = \tau_i$, then the responses to categories *i-1*, *i* and *i+1* would be equally likely at that point on the continuum. In this case, there would be no

interval where the response category *i* would be the single most likely option. Different approaches lend themselves to the judgement of the empirical threshold order.

Simple order of threshold estimates

In the following, four approaches for the empirical assessment of the threshold order will be discussed. First, the evaluation of the empirical threshold order can rest upon the actual order of threshold estimates. Consequently, if $\hat{\tau}_{i+1} > \hat{\tau}_i$, the thresholds are assumed to be properly ordered. By contrast, the thresholds are considered disordered, if $\hat{\tau}_{i+1} \leq \hat{\tau}_i$. Since this decision rule does not require standard errors for the threshold estimates, it can be applied when standard errors are not available. However, as a matter of principle, this approach ignores the fact that threshold estimates can be accidentally disordered but also accidentally ordered.

Statistical tests for the order of threshold estimates

When standard errors of the threshold estimates (SE_τ) are available, statistical tests can be carried out informing the assessment of the response scale functioning. At first, the test requires a null hypothesis and an alternative hypothesis. The formulation of hypotheses is not completely trivial as the equality of thresholds $\tau_{i+1} = \tau_i$ may, in principle, be part of the null hypothesis or the alternative hypothesis.

Therefore, the second approach views an ordered response scale as requiring positive empirical support. Then $\tau_{i+1} > \tau_i$ represents the alternative hypothesis H_A, while $\tau_{i+1} \leq \tau_i$ would be the null hypothesis H_0. Consequently, the estimated thresholds $\hat{\tau}_{i+1}$ and $\hat{\tau}_i$ are required to be "significantly" ordered calling for a one-tailed test of the difference $\tau_{i+1} - \tau_i$. Assuming a 5% type-I-error rate, $\hat{\tau}_{i+1}$ would need to be bigger than $\hat{\tau}_i$ by 1.65 times the joint standard error of the two threshold estimates $\sqrt{SE_{\hat{\tau}_{i+1}} + SE_{\hat{\tau}_i}}$. While significantly ordered thresholds would provide strong justification of the scoring of response categories, type-II-error (the probability of retaining the null hypothesis implying disordered thresholds) could be as high as 95% (assuming a type-I-error rate of 5%). This rate applies to marginally ordered thresholds. In practice, type-II-error would be smaller but can still be quite substantial enough to frequently reject the hypothesis of ordered categories when they are truly ordered.

The third approach counters the excessive false diagnosis of reversed thresholds in the second approach by reversing the specification of empirical evidence. Now, ordered response categories are the default assumption. Empirical evidence needs to refute the meaningfulness of the scoring. In other words, the hypothesis of ordered categories persists in the absence of evidence to the contrary, which makes this approach very conservative. Then H_0 states that $\tau_{i+1} \geq \tau_i$, while H_A implies $\tau_{i+1} < \tau_i$. Maintaining the 5% type-I-error rate, $\hat{\tau}_{i+1}$ would need to be smaller than $\hat{\tau}_i$ by 1.65 times the joint standard

error of the two threshold estimates in order to infer that the thresholds are actually disordered. Like in the previous approach, type-II-error could be very large and approach 95%. However, here it would mean accepting the hypothesis of ordered categories even though they are in fact not properly ordered. From a pragmatic point of view, the most important difference between the second and the third approach are the vastly different probabilities of failing to identify disordered thresholds and of wrongly identifying items as problematic.

Simple order of threshold estimates reframed as a statistical test

Although the first approach does not carry out any statistical test, the decision rule can be framed as a statistical test of the null hypothesis of disordered thresholds stating $\tau_{i+1} \leq \tau_i$. The formal type-I-error rate would be 50% (truly disordered thresholds remain undetected), while type-II-error could be 50% at most. The virtue of the first approach is twofold. On the one hand, the decision rule is simple and does not even require any computations. On the other hand, type-I-error and type-II-error rates are balanced. The latter appears to be justifiable in a situation where both types of errors (applying an unjustified scoring key versus wrongly changing a response scale format) equally matter. Nevertheless, the error rates can still be quite high. One way to accommodate this problem would be to increase sample size. This would decrease the standard error of the thresholds and imply that a wrong diagnosis occurs predominantly in cases where the true thresholds are in relatively close proximity. At any rate, the first approach neglects any statistical information on the thresholds. Whether this really implies a shortcoming is controversial. Andrich (2011) questions the value of statistical tests in this regard pointing out that "the significance of such tests is substantially a function of sample size and therefore it can be contrived" (p.581). We will discuss this general shortcoming of statistical tests later.

A hybrid approach

In the following, as a fourth approach, a hybrid procedure is suggested that is a compromise between the second and the third approach. The null hypothesis is $\tau_{i+1} = \tau_i$, while there are two alternative hypotheses. H_1 states that $\tau_{i+1} > \tau_i$ and, thus, represents ordered categories. H_2 states that $\tau_{i+1} < \tau_i$ representing disordered categories. There are three different outcomes: The thresholds might be considered ordered (H_1), disordered (H_2), or H_0 might be retained. In the latter case, no decision would be made as to the true order of thresholds, even though, like in the first approach, $\hat{\tau}_{i+1} > \hat{\tau}_i$ would rather point at ordered thresholds and $\hat{\tau}_{i+1} \leq \hat{\tau}_i$ would hint at disordered thresholds. In these cases, additional information on the threshold order could be based on replications. If an instrument uses the same response scale format for a series of items, the responses to these items could be interpreted as replications, even though the response scale might work slightly differently for different items. If the empirical estimates of particular thresholds are always or almost always properly but not significantly ordered, evidence builds up in favour of ordered thresholds. By contrast, if a good part of the empirical estimates of thresholds between the same categories are disordered, the category ordering should be questioned

seriously even though no single item shows significantly disordered thresholds according to the third approach.

Table 1 summarises the approaches in terms of their formulated hypotheses, decision rules and error rates. Figure 7 shows the error rates associated with the decision rules. Table 2 shows the consequences for the given data set and future applications of the instrument.

Table 1:
Approaches to Evaluate Threshold Order

Approach	H_0	H_A	Decision rule	Type-I-error	Type-II-error
1 simple decision rule	$[\tau_{i+1} \leq \tau_i]$	$[\tau_{i+1} > \tau_i]$	ordered: $\hat{\tau}_{i+1} > \hat{\tau}_i$; disordered: $\hat{\tau}_{i+1} \leq \hat{\tau}_i$	50% (actually disordered)	up to 50% (actually ordered)
2 seeking evidence for ordered thresholds	$\tau_{i+1} \leq \tau_i$	$\tau_{i+1} > \tau_i$	ordered: $\hat{\tau}_{i+1} > \hat{\tau}_i + 1.65 \bullet S.E.$; disordered otherwise	5% (actually disordered)	up to 95% (actually ordered)
3 seeking evidence for disordered thresholds	$\tau_{i+1} \geq \tau_i$	$\tau_{i+1} < \tau_i$	disordered: $\hat{\tau}_{i+1} + 1.65 \bullet S.E. < \hat{\tau}_i$; ordered otherwise	5% (actually ordered)	up to 95% (actually disordered)
4 hybrid, seeking evidence for ordered as well as for disordered thresholds	$\tau_{i+1} = \tau_i$	H_1: $\tau_{i+1} > \tau_i$ H_2: $\tau_{i+1} < \tau_i$	ordered (H_1): $\hat{\tau}_{i+1} > \hat{\tau}_i + 1.65 \bullet S.E.$; disordered ($H_2$): $\hat{\tau}_{i+1} + 1.65 \bullet S.E. < \hat{\tau}_i$; undecided ($H_0$) otherwise	5% (actually ordered/ disordered)	[up to 90% (actually ordered or disordered)]

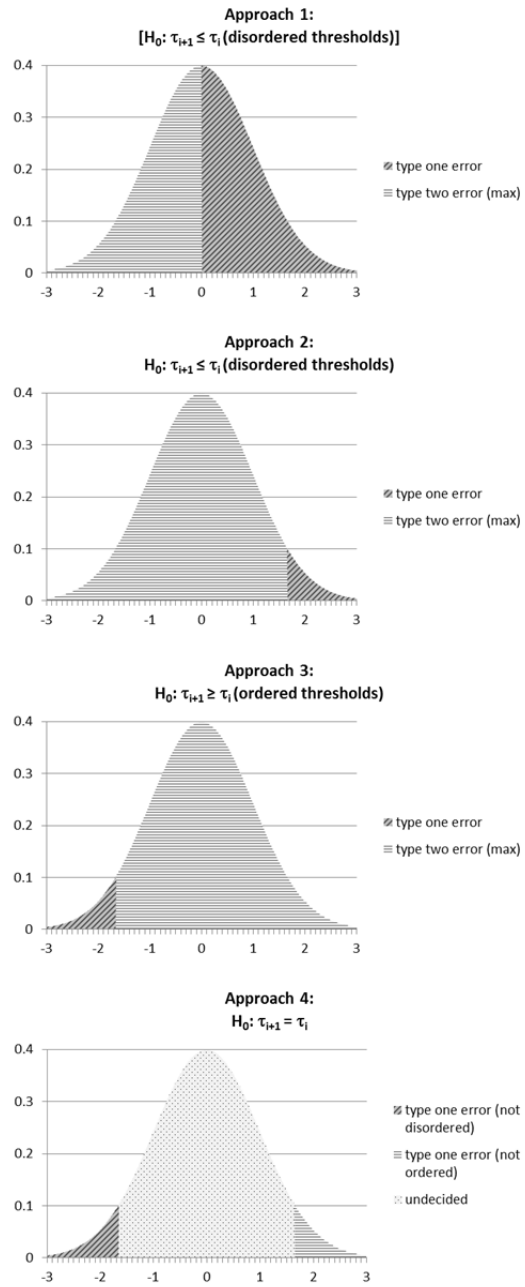


Figure 7:
Error Rates Involved in the Assessment of Threshold Order

Table 2:
Consequences of the Assessment of Threshold Order

Approach	Conclusion about thresholds τ_{i+1}, τ_i	Threshold estimates $\hat{\tau}_{i+1}, \hat{\tau}_i$,	Corrective action for given data set	Adaptation in future application
1 simple decision rule	-	$\hat{\tau}_{i+1} > \hat{\tau}_i$ *ordered*	none	none[a]
	-	$\hat{\tau}_{i+1} \leq \hat{\tau}_i$ *disordered*	collapse categories	revise response format
2 seeking evidence for ordered thresholds	$\tau_{i+1} > \tau_i$ *ordered*	$\hat{\tau}_{i+1} > \hat{\tau}_i$ *ordered*	none	none
	$\tau_{i+1} \leq \tau_i$ *disordered*	$\hat{\tau}_{i+1} > \hat{\tau}_i$ *ordered*	none	revise response format, or replicate given format
	$\tau_{i+1} \leq \tau_i$ *disordered*	$\hat{\tau}_{i+1} \leq \hat{\tau}_i$ *disordered*	collapse categories	revise response format
3 seeking evidence for disordered threshold	$\tau_{i+1} \geq \tau_i$ *ordered*	$\hat{\tau}_{i+1} > \hat{\tau}_i$ *ordered*	none	none[a]
	$\tau_{i+1} \geq \tau_i$ *ordered*	$\hat{\tau}_{i+1} \leq \hat{\tau}_i$ *disordered*	consider collapsing categories	revise response format, or replicate given format
	$\tau_{i+1} < \tau_i$ *disordered*	$\hat{\tau}_{i+1} < \hat{\tau}_i$ *disordered*	collapse categories	revise response format
4 hybrid, seeking evidence for ordered as well as for disordered threshold	$\tau_{i+1} > \tau_i$ *ordered*	$\hat{\tau}_{i+1} > \hat{\tau}_i$ *ordered*	none	none
	$\tau_{i+1} < \tau_i$ *disordered*	$\hat{\tau}_{i+1} \leq \hat{\tau}_i$ *disordered*	collapse categories	revise response format
	$\tau_{i+1} = \tau_i$ *undecided*	$\hat{\tau}_{i+1} > \hat{\tau}_i$ *ordered*	none	revise response format, or replicate given format
	$\tau_{i+1} = \tau_i$ *undecided*	$\hat{\tau}_{i+1} \leq \hat{\tau}_i$ *disordered*	consider collapsing categories	revise response format, or replicate given format

[a] Consider revised response format if $\hat{\tau}_{i+1} \cong \hat{\tau}_i$

Consequences of disordered thresholds

The diagnosis of disordered thresholds is important as it affects the justification of the scoring of polytomous items and, thus, represents a very important element of measurement. If the input to the Rasch analysis is questionable, inferring valid measurements is equally dubious. In the end, a malfunctioning response scale should be modified before

new data are collected. While future administrations of an instrument can make use of a revised response format, in a given data set modifications are confined to the scoring of the response options. As a post-hoc remedy, adjacent categories can be collapsed, that is scored equally, when threshold are disordered.

Such a rescoring is recommended in approaches 1 (simple decision rule), 3 ("conservative") and 4 ("hybrid") when true thresholds are considered disordered as the estimated thresholds would then be reversed in any case (see Table 2). In approach 2 ("demanding"), thresholds might not be considered properly ordered even though the actual estimates are ordered. Collapsing categories seems implausible, particularly if it adversely affects item fit. Impaired item fit after collapsing categories in the presence of ordered threshold estimates would suggest that the response format might actually function properly.

In approach 4 (hybrid), the conclusion "undecided" raises questions. If the empirical threshold estimates are disordered, collapsing categories should be considered, especially if item fit improves. The same applies to the conclusion of ordered thresholds in approach 3 when actual estimates are disordered. It follows that, generally speaking, the simple approach 1 seems to be sufficient as far as modifications of the scoring in a given data set are concerned.

While collapsing response categories by scoring them equally may avoid implausible conclusions in a given data set, presenting, for example, three response options and scoring two of them equally is not the same as presenting only two options in the first place. Collapsing categories is a transitory remedy. Future studies should cross-validate the proposed new format based on data that are collected actually using a modified format. Reversed threshold estimates suggest considering a modification of the response scale, even though, following approach 3, would not imply an urgent need to do so when the thresholds are not significantly disordered. By contrast, when adhering to approach 2, modifications might be envisaged even in the presence of ordered threshold estimates when order cannot be generalised. In other words, a judgment by the researcher is called for. This becomes explicit when, in approach 4, the conclusion is undecided. Then the researcher has to decide whether the original format or a modified response scale should be used in future applications based on other evidence such as qualitative interviews with respondents.

In the end, the consequences for a given data set vary only slightly due to the approach chosen to evaluate the threshold order. Whenever the estimates are disordered, collapsing categories should be envisaged. An exception is the conservative approach 3, which seeks evidence of disordered thresholds. If threshold estimates are disordered but not significantly, the disordering might be ignored. However, even then, collapsing categories might be considered, particularly if item fit improves.

If collapsing response categories results in properly ordered remaining thresholds and improved item fit, it is certainly a promising indication that the amended scoring key is more reasonable than the original. However, if the original scoring key results in reversed thresholds but good fit, any rescoring is very likely to impair item fit. The reason is that if a particular scoring of the responses is associated with good fit, it is almost

mathematically impossible that a revised scoring fits equally good or even better (see Andrich, 1995, on the non-dichotomization of polytomous responses). If the thresholds are only marginally disordered, one might ignore the reversal, specifically if one adheres to the conservative approach 3. Alternatively, the item can be deleted since an item showing misfit but ordered categories can hardly be considered an improvement over a fitting item with reversed thresholds. Nevertheless, in future applications, a revised response format should be envisaged.

Summary, conclusions and further research

A comprehensive psychometric data analysis needs to address several questions. First, formal requirements of measurement, as prescribed by the Rasch model, have to be tested. Fit statistics provide useful guidance in this respect. Second, beyond fit assessment, the analysis has to investigate to what extent the predictions made by the construct theory, if any, are confirmed. Third, if polytomous items are used, the analysis has to deal with the proposed order of response categories representing a theory on its own. On the one hand, malfunctioning response scales compromise the meaning of the total score and, therefore, interfere with fit assessment. On the other hand, disordered response categories may or may not lead to item misfit. For this reason, the inspection of threshold estimates provides a useful diagnostic tool. However, the fact that threshold estimates may accidentally be disordered raises the question whether formal statistical tests taking standard errors into account could be helpful. At first, it has to be decided whether threshold estimates need to be significantly ordered or merely not significantly disordered. While the former, more rigorous demand is likely to result in many incorrect rejections of properly working response scales, the latter, more conservative approach will retain response scales with truly disordered categories. A suggested hybrid approach combines both perspectives, but comes at the expense of possibly inconclusive results where subjective judgements have to be made. Nevertheless, as far as future improvements of a scale are concerned, the hybrid approach seems to provide reasonable guidance. When a common response scale is used across items, the tests can be interpreted as replications helping decide whether a threshold reversal is likely to be incidental or generalizable. By contrast, in terms of collapsing categories in a given data set as an exploration of alternative scoring schemes, the simple decision rule based on point estimates of threshold still seems to be useful.

Since rescoring responses implies changing the data, it should under all circumstances be seen as a preliminary measure. While amending the model for the sake of better fit runs contrary to the philosophy to Rasch measurement, changing the data should be viewed with extreme caution, too. After all, observations are reinterpreted after the fact in order to improve the match of the data and the measurement model.

While the possible inconclusive outcome of the hybrid approach appears to be a disadvantage, it can also be reframed as a virtue. Particularly with small sample sizes, the power to detect truly reversed thresholds or identify truly ordered thresholds is very limited. In these cases, the outcome generally depends on the perspective the researcher

takes in terms of whether thresholds are required to be significantly ordered or merely not significantly disordered. In the former case, relying on statistical tests would raise mostly false alarms (truly ordered thresholds are not significantly ordered because of insufficient power), while in the latter may lull the researcher into a false sense of security (thresholds disordered but not strongly enough to be significant). While statistical tests should never be carried out mechanically, this is particularly true in the case of threshold order. In the end, statistical evidence is of little help when it comes to immediate post-hoc remedies, such as collapsing categories or item deletion. Rather, it may be used as guidance for shaping future amendments of the response scale. Nevertheless, changes to the response format should not be based on statistical evidence alone. After all, statistics merely indicate problems in the response scale but do not necessarily, if at all, point out how the scale should be changed. Qualitative interviews with respondents are more likely to reveal avenues to more suitable response formats. If statistical evidence is to be used, the hybrid approach seems to be the most cautious approach. In any case, the sample dependence of statistical tests has to be kept in mind. With very small sample sizes, difference tests of thresholds make little sense, while very large samples would hardly imply additional insight beyond the mere order of threshold estimates.

What exactly constitutes a reasonable sample size is subject to future research. Further applied research should investigate whether the hybrid approach, or possibly one of the two other statistical decision rules, improves the correct identification of malfunctioning response scales in real data. Another starting point for further research is the question of how accurately the empirical order of threshold estimates actually reveals disordered response categories. Reversed thresholds imply malfunctioning of the response scale once other reasons, such as multidimensionality or other violations of the Rasch model have been ruled out. But is the reverse necessarily true, as well? Does malfunctioning of the response scale lead to reversed thresholds? If there are conditions under which threshold estimates are likely to be ordered even though successive categories do not indicate an increasing amount of the property, the implications would be twofold. First, the conservative approach of looking for significantly disordered thresholds would be hard to defend. Second, the investigation of the response scale functioning would have to take other information explicitly into account, such as threshold characteristic curves and discrimination at the threshold. The investigation of threshold estimates alone is possibly too indirect.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement, 72*(4), 547-573.

Andersen, E.B. (1977). Sufficient Statistics and Latent Trait Models. *Psychometrika, 42*, 69-81.

Andrich, D. (1978). Application of a Psychometric Rating Model to Ordered Categories Which Are Scored with Successive Integers. *Applied Psychological Measurement, 2* (4), 581-594.

Andrich, D. (1988). A General Form of Rasch's Extended Logistic Model for Partial Credit Scoring. *Applied Measurement in Education, 1*(4), 363-378.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*, 571-585.

Andrich, D. (2013). An Expanded Derivation of the Threshold Structure of the Polytomous Rasch Model That Dispels Any ''Threshold Disorder Controversy''. *Educational and Psychological Measurement*, *73*(1) 78–124.

Andrich, D., Sheridan, B.S., &Luo, G. (2009). Rumm 2030: Rasch Unidimensional Measurement Models [computer software]. RUMM Laboratory Perth, Western Australia.

Karabatsos, G. (2001). The Rasch Model, Additive Conjoint Measurement, and New Models of Probabilistic Measurement Theory. *Journal of Applied Measurement, 2* (4), 389-423.

Kreiner, S. (2007). Validity and objectivity: Reflections on the role and nature of Rasch Models. *Nordic Psychology, 59* (3), 268-298.

Masters, G.N. (1982). A Rasch Model for Partial Credit Scoring, *Psychometrika, 47* (2), 149-174.

Messick, S. (1995). Validity of Psychological Assessment. Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning, *American Psychologist, 50* (9), 741-749.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests, Copenhagen: Danish Institute for Educational Research (reprint 1980, expanded edition, with foreword and afterword by B. D. Wright). Chicago: The University of Chicago Press.

Stenner, A.J., Stone, M.H., & Burdick, D.S. (2009). The concept of a measurement mechanism. *Rasch Measurement Transactions, 23*, 1204–1206.

Stenner, A.J., Fisher, W., Stone, M.H., Burdick, D.S. (2013). Causal Rasch Models. *Frontiers in Psychology, 4*, Article 536, 1-14.

Wilson, M. (2005). Constructing Measures, An Item Response Modeling Approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B.D. (1992). Rasch Model derived from Ratio-Scale Counts. *Rasch Measurement Transactions, 6*:2, 219.