PSYCHOMETRIKA—VOL. 60, NO. 1, 37–46
MARCH 1995

FURTHER REMARKS ON NONDICHOTOMIZATION OF GRADED RESPONSES

DAVID ANDRICH
SCHOOL OF EDUCATION
MURDOCH UNIVERSITY

1. Background

In addition to addressing the points raised in Roskam (1995) as a rejoinder to Andrich (1995), these further remarks on the nondichotomization of graded responses attempt to clarify the difference in perspective between the two articles that gives rise to the different appreciation of what appear to be the same details. Therefore, these remarks take the following form: first, a brief comment is made on the development of the argument and perspective against routine dichotomization of graded responses (which is often seen as no more than commonsense); second, some points raised in the rejoinder are reexamined and clarified from this perspective; third, the perspective is consolidated with examples to show the way in which nondichotomization of graded responses is consistent with intuition. Roskam (1995) and Andrich (1995) are referred to, respectively, as *the rejoinder* and *the paper*.

I agree with Roskam that the property of the Rasch model that categories may not be grouped in general, and dichotomized in particular, is surprising, even perhaps counterintuitive, in the first instance. Rasch himself realized this when he wrote "If the basic model holds with, say, five categories, it is mathematically almost impossible for the three-category model also to hold. Thus the grouping, tempting as it may be, will usually tend to slur the specific objectivity" (1966, p. 107). Subsequently, in series of papers, Jansen and Roskam (1986; referred to as JR) formalized the grouping of categories into their *joining assumption*.

Apart from Rasch's (1966) remarks, it seemed that up to that time, there were no other comments in the literature concerning the grouping of graded responses, with the standard model (Samejima, 1969) having the property that they could be grouped by definition. As part of this background, I acknowledge stimulating discussions with Roskam, beginning in 1984 (Roskam, personal communication, 1984), and the importance of the work of Jansen and Roskam in appreciating, and not ignoring, the significance of the result of nondichotomization in the unidimensional Rasch model (URM). Their work gave focus to the paper under discussion.

Although appearing sure that nongrouping feature of the model was significant, Rasch seemed not to have a complementary intuitive explanation for it (Rasch, personal communication, 1974, 1977). Andersen (1977) took the understanding of this result a step further by showing analytically that if the unidimensional model were to have sufficient statistics, then the scoring functions had to increase by a constant and that two adjacent categories could be grouped only if their scoring functions were identical. The first step in developing an intuitive appreciation, in conjunction with an

Requests for reprints should be sent to David Andrich, School of Education, Murdoch University, Murdoch WA 6150, AUSTRALIA.

0033-3123/95/0300-94125$00.75/0

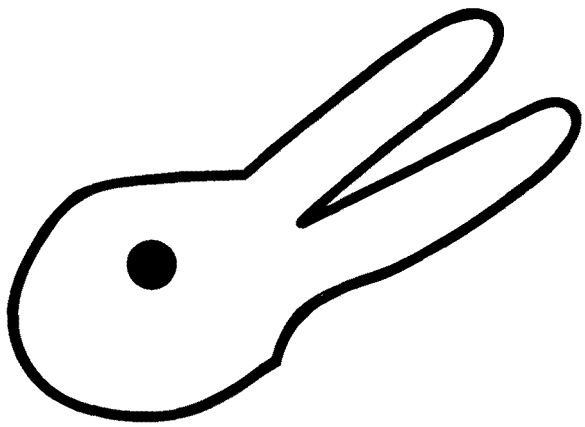

FIGURE 1.
Is this a rabbit or a duck?

analytical one, appeared in Andrich (1978), which showed that the successive scoring functions were sums of discriminations at successive thresholds partitioning the continuum, and that these discriminations also appeared as products of thresholds and discriminations in the model. In order to consolidate intuitively, as well as analytically, why the unidimensional Rasch model (URM) does not permit grouping, I will have to invoke the full ramifications of these discriminations: therefore, their role is now consolidated.

It followed from Andersen (1977) and Andrich (1978) that if the discriminations at thresholds were identical, the increment in the successive scoring functions for the successive categories was constant and that the scoring functions became integers expressing the count of the number of thresholds exceeded. Importantly, these relationships also explained that if a threshold between two categories had zero discrimination, then their scoring functions were identical and the outcome between the two categories was a random allocation. It then followed that if the outcome between the categories was no more than a random allocation, then the categories could and should be grouped. Complementary to this, it implied that if a threshold between two categories did discriminate, then it affected the distribution of responses and the categories could not be grouped. Although nongrouping when a threshold did discriminate could be defended on the grounds that it was complementary to grouping when a threshold did not discriminate, it still did not make clear *intuitively* why the categories could not, in general, be grouped.

The step consolidating an intuitive appreciation for non-grouping of categories in the URM arose when, for convenience (and unaware of the consequences), I expressed the model in the multiplicative metric with equal sized categories, and then seeing that the resultant model would be the multiplicative Poisson model (MPM) if the number of categories were permitted to be unlimited. The MPM, with its elegant properties, which are exploited in the paper, related the distribution of the responses among categories to the precision of the estimate of the relevant location in a simple and transparent way.

As in many features of the Rasch models for measurement where they conflict initially with intuition, understanding the implications is not only a matter of analyzing the details, but is in large part a matter of a Gestalt switch, a matter of seeing the whole from a different perspective, more or less consistent with the paradigm shift of Kuhn (1970, p. 85), which I have discussed in general in Andrich (1988). To illustrate this point using a standard metaphor, it is the difference between seeing a duck or a rabbit in the picture in Figure 1.

I make this point partly because it might appear that Roskam and I are "talking through each other," to take another of Kuhn's ideas (1970, p. 109), but also to explain why I will concentrate on the Gestalt as much as on the detailed argument. To summarize, the first aim of the paper and these further remarks, is to provide an intuitive understanding, complementary to an analytic one, as to why the joining assumption does not hold in the URM. However, a function of models is to generate logical deductions which give new insights which might be difficult to gain otherwise. Therefore, the second aim is to show the circumstances where one would generally *not want* the joining assumption to hold in graded response data. I begin with the rejoinder's summary points in which Roskam considers that I (a) do not treat the joining assumption as defined by JR; (b) confuse rescaling of the response continuum with the replacing of response categories; and (c) have the response distribution tied to the item parameter.

2. ξ Invariance, the Joining Assumption, Rescaling and the Item Parameter

The joining assumption as defined in JR. In the paper I assume that JR asserted the equivalence of ξ-invariance and the joining assumption. Roskam reacts that this is not the case. I interpreted this equivalence from many statements and implications in JR including those quoted in the paper. However, even granting that I have overstated their implied relationship between the joining assumption and ξ-invariance, I believe I am not overstating the case that JR implied that because the Rasch model violated the joining assumption, it also violated ξ-invariance in a substantive way. It seems to me that JR rejected a special case of the URM, which they termed the rating Rasch model (RRM), because they argued that in violating the joining assumption, it also violated ξ-invariance.

They write:

> In so far as the joining assumption . . . applies to partitioning of the response continuum before the fact, the construction of graded responses interferes with the measurement outcome in the RRM . . . . Therefore, the validity of the RRM as an instrument for measuring subjects is questionable. Since the property of ξ-invariance is seen to be essential for rating data, [the RRM] can not be considered as a measurement model for rating data. (Jansen & Roskam, 1986, p. 81).

Rescaling of the latent continuum. There is no confusion in the paper between any rescaling of the response continuum and the replacement of categories, because the location parameter of the person and the continuum are never rescaled. The only change is to partition the same continuum into different sized categories and, specifically for illustration in the MPM, halving the size of the previous categories before the data are collected. If there is discrimination at each threshold, the responses have to be redistributed among twice as many categories which reflect an increase in precision attributed to the extra discrimination gained and, in the process, the distribution is narrower and the joining assumption does not hold. There *appears* to be a rescaling when the counts are reexpressed in terms of the original units, but this is not a rescaling or transformation of the continuum, it is simply similar to reexpressing counts in 1/2 yards to counts in yards, or counts in feet to counts in yards. The continuum itself is not shrunk or stretched—ξ_p itself is never transformed. I return to this feature in section 5 where I comment on the association of successive integers with successive categories.

The response distribution is tied to the item parameter. I take it that Roskam's comment that the response distribution is tied to the item parameter refers to the set of threshold parameters, because the single item parameter, such as difficulty, is absorbed into the threshold parameters for convenience without loss of generality. And this is exactly the case—the response distribution is tied to the number and locations of the *discriminating* thresholds—it is not a confusion or an error, but the key result: I claim it is consistent with intuition and requirements of measurement that increasing the number of discriminating thresholds (that is, meaningful categories) should increase the precision of the estimate of the location ξ by making the response distribution on the same continuum narrower. Thus here is an intuitive, in addition to an analytic, understanding, of why dichotomization is not acceptable in the URM. This, of course, is a property of the model, and it is, as Roskam and I agree, an empirical question whether or not data do conform to the model. I seem to disagree with Roskam in that I conclude that, because it would indicate that more categories give more precise information, graded response data *should* violate the joining assumption.

3. Semantic Partitioning, ξ-Invariance, and Parameter Estimation

Roskam considers that the "joining assumption is meaningless when the categories bear no semantic labels" (p. 5), and he proceeds to give examples in which he would expect the joining assumption to hold. To bring evidence against this position, I first draw on some writing of the linguist Saussure who explained that categories which partition some semantic space are not just defined *operationally*, but also *relatively*. Thus Saussure articulated that linguistic entities did not have meanings only in terms of some object or concept that they represented, but were also defined *in terms of each other*. The former Saussure referred to as the *signification* of the linguistic entity, and the latter he referred to as its *value* (Saussure, 1959, pp. 114–116).

> Within the same language, all words used to express related ideas limit each other reciprocally; synonyms like French *redouter* "dread," *criandre* "*fear*" and *avoir peur* "be afraid" have value only through their opposition: if *redouter* did not exist, all its content would go to its competitors . . . . The value of just any term is accordingly defined by its environment; it is impossible to fix even the word signifying "sun" without first considering its surroundings: in some languages it is not possible to say "sit in the sun" . . . (Saussure, 1959, p. 116).

From this linguistic analysis, I take it that if a set of categories were located on a semantic continuum of some kind and defined operationally, that they would operate not only as they had been defined operationally, but also relatively in terms of each other, and that if one category were eliminated or added, it is likely to change the relative meaning of all of the categories in the space, even if the remaining categories retained their original operational definition. If relative meanings change when a category is eliminated, objects may be reclassified in such a way that the joining assumption does not hold. Thus, here we have evidence that even in semantic spaces, there seems no reason to expect that the joining assumption should hold.

Roskam indicates that he knows of no examples where the joining assumption has been shown to fail. I suspect that this is in part because people have taken the joining assumption for granted and not put it to the test. Taking a combination of the result from the URM and Saussure's analysis as a motivation, I have reported two empirical studies recently which illustrate the violation of the joining assumption (Andrich, 1993a, 1993b). In the former (Harris, 1991), a set of essays was graded by the same

graders according to two formats: in one there were three graded categories and the other there were four where the fourth category was explicitly made a subset of the third. The first three categories were defined operationally in exactly the same way in the two sets. This enhanced the possibility that the joining assumption would hold. The result of this exercise was that the distribution of the responses in the first two categories was different in the two formats, and the sum of the probabilities of the responses in the third and fourth categories was not equivalent to the probability of the responses in the third category in the three category set. The inclusion of the extra category had an impact on the meaning of all of the other categories and the joining assumption did not hold.

In the second study, semantic labels with three and four ordered categories were again defined operationally where the fourth category in the four-category set was a subset of the third category in the three-category set, again enhancing the possibility that the joining assumption would hold. Persons were then asked to partition a continuum to match their understanding of the relative length of these categories. Again, the partitioning of the continuum for the first two categories was different, thereby violating the joining assumption, even though their operational definition was the same in the three and four-category sets.

These results are consistent with the case volunteered in the rejoinder, where Roskam tentatively seemed to take the perspective of the URM, and acknowledged that adding a category *strongly agree* to the set *strongly disagree*, *disagree*, and *mildly agree* in the *stochastic subject* interpretation, may change the distribution. Roskam goes on to say that this might violate the joining assumption and might also violate the URM. However, the point is that it is likely to violate the joining assumption but that it might not violate the URM. If this were the case, then, because the estimate of ξ would be the same in the two sets of categories and because the precision would be greater with the four category set, the URM would be ideal to use to locate the person on the continuum. Thus the perspective in the rejoinder that "Semantic (re)partitioning of the response continuum is considered arbitrary and the response probabilities should follow the joining assumption . . . " (p. 30) is challenged by both a comparison with a linguistic analysis and empirical evidence.

Although the empirical examples summarised above are complementary to the argument against the joining assumption, they are strictly beside the point in understanding analytically and intuitively the reason that the URM violates the joining assumption. To consolidate the difference in perspective, and different interpretations of what appear the same results, I consider here three further comments in Roskam's rejoinder.

(i). "The joining assumption thus is based *in the semantic of response categories*, and is *independent of any measurement model.*" (p. 29, italics in original). I argue that the joining assumption is not independent of the measurement model—it is clearly violated in the URM, and a close analysis of the reason it is violated has something to say about the semantics of the response categories and, in Saussure's terms, their "values."

(ii). "The crucial question for empirical science is: what do subjects do when confronted with different sets of response categories, and is the measurement model compatible with that?" (p. 31). It seems to me that this is putting the question back to front. It seems to me important to know what they are *required* to do if their responses in a greater number of categories are to provide greater precision in estimating the same location parameter ξ. The MPM gives this insight. Then knowing what they *might* do is relevant in setting up the categories or providing, as in performance assessment, rele-

vant training as to how to use the categories. It may be the case that with many categories the joining assumption does not hold, but in that case it means many categories are redundant and the persons are not using the categories as intended. In this context, and apparently from Roskam's perspective, if the joining assumption held with graded responses, he would not employ the URM; from the Rasch (and my) perspective, if the joining assumption held, it would be evidence that the data are suspect. I take this point further in section 4.

(iii). "Behavior is not explained by statistics" (p. 34). I totally agree, but neither, in the physical sciences, is the behavior of objects explained by mathematics. Nevertheless, mathematical models have proved powerful in gaining insights into the behavior of physical objects. Similarly, measurement models can give insights into actual human behavior and required human behavior when it is to serve certain functions (as in this case which reveals how graded responses should function if more categories are to give more precise information). The URM, derived purely from an abstract definition of measurement as the invariance of comparisons, does lead to such an insight. Furthermore, the sensitivity of the model to the addition or elimination of categories, which appears consistent with the linguistic analysis described briefly above, opens up a range of potential empirical studies that integrate semantic categorisation and mathematical modeling.

4. Effect of Instrumentation

There are some further points where Roskam and I agree on details, but seem to have a different interpretation, although here our differences either appear more subtle or our perspectives might be closer. The issue is concerned with the effect of the instrument itself on the manifest outcome, which is the effect characterised by the URM—the outcome arises from the contact between the object of measurement and the instrument, and properties of both govern the outcome.

For example, Roskam considers the case of asking persons whether or not they can carry an object over ranges of 0 and 1 inches, and so on. Rather than these small distances where the classification system is unlikely to evoke variation because of easy targeting (creating a floor or ceiling effect), consider that we were to test the ability of persons to carry objects of different mass, without taking a rest, by having them carry objects up the steps of one of the large pyramids of Mexico, where, at a given rate of walking it would not be expected that persons could get much further than about half way, perhaps between a quarter and three quarters of the way. Here the locations of the thresholds, whose successive differences give the size of the steps to be taken, would have a big bearing on how far each person would be able to carry the object and their perception of their ability to carry the object. To be more explicit, suppose that in one case a person were required to take one step at a time, and in the second to take two steps at a time, that is, to have half as many thresholds and therefore have categories of twice the length (assuming the steps are of equal size for this purpose). It would be expected that the same person could carry the same object further when the thresholds were closer than when further apart, which is consistent with the well known principles of gear construction. The effect of the instrument (in this case the location of the thresholds) is exactly the kind of effect that the URM reflects. The probability of any particular outcome depends on the location of all the thresholds, making it unlikely that the joining assumption could hold when some of the thresholds are eliminated. All other factors held constant, the closer the thresholds, the greater the height reached for the same ability and the greater the precision of the estimate of the ability. Even looking up the pyramid and seeing how many further steps the object is to be carried will have an

effect on whether the person will take the next step. For example, for a given level of tiredness, if there is only one step to go, a person may take it, while if there are 100 steps to go, a person may simply stop there. Thus the probability of being in any category x is a function of the location of all thresholds, as implied in the URM, and not just the first $x - 1$ thresholds.

Roskam goes on to describe that my fundamental error is that I ". . . ignore that we are not dealing with the reading of a ruler, but with a probabilistic response process." and that "The variance of this process does not vanish (in the limit) by observing it more precisely" (p. 33). First, it is exactly what Roskam says I ignore that is central to my case, once again illustrating the difference in perspective rather than the details. Thus Roskam and I agree that the outcome is a matter of the response process, and not just reading a ruler. To me, and from the model, this implies that if the response process of the instrument locates objects one yard apart then it cannot locate them 1.5 yards apart, for then the response process is in terms of 1/2 yards, and not yards. In the case of carrying an object up the Mexican pyramid taking one step at a time, one could not, for example, carry the object 155.5 steps say, but only 155 or 156 steps. Taking two steps at a time, one could not, for example, take the equivalent of 161 single steps, but $160/2 = 80$ double steps or $162/2 = 81$ double steps. However, after the fact, one may reexpress lengths in terms of a standard unit, and say that 162 single steps is equivalent in length to 80 double steps. I return to this point in the next section.

It is important to appreciate that when the units are halved in the MPM, it is implied that effectively there is a different instrument which manifests responses differently, and not just that there is the same instrument but with units in which the outcome read is halved. Roskam asks "what is the meaning of decreasing ω? Do we have another item? (p. 33) The answer is, in a very real sense, "yes!" When categories are grouped before the data are collected, we *effectively* have a different item in terms of its effect on the person responding. When persons are required to take every second step up the Mexican pyramid, we have a different instrument from when they take every step. Thus the location of the object is not taken to be manifested in the same place with the two instruments, with the location read more accurately when the units of the instrument are smaller, but that the outcome of the observations themselves changes. I reiterate what Roskam has also stressed throughout, that whether or not the response process satisfies the URM or MPM, is an empirical question. However, according to the MPM, if the new instrument has a unit (size of the categories) with smaller units which operate as such, it implies a greater precision which in turn implies a smaller variance.

One would imagine that every instrument has its own level of precision in manifesting responses. Roskam also asks why an instrument is inaccurate in the first place. No single answer can be given; the inaccuracy depends on a constellation of factors that generate different outcomes on different replications of measurement, and to the degree that the variance in these outcomes can be reduced by controlling these factors while measuring the same substantive variable, to that degree the precision of measurement is increased and the joining assumption must be destroyed.

Likewise, it may happen that the object of measurement itself has some actual variance in location from occasion to occasion of measurement, but then the object would have to be characterised by a location and variance. In the URM and the MPM, and most models for graded responses, the object is characterized only by a location, and therefore in theory its location can be measured infinitely precisely with an infinitely precise instrument. The model does not imply that one, in fact, has an infinitely precise instrument, or that a person really does not vary at some very fine level of precision. The fact that one may not be able to construct an instrument with infinite

precision means that we would not invoke infinitely small units in the MPM; one would invoke the actual size of the unit satisfied in the MPM (if it were satisfied at all). However, this does not prevent us from appreciating that the model implies that if an instrument were infinitely precise, it would have an infinitely small variance in outcomes, and that, therefore, the joining assumption would not hold—it would not hold in social or physical measurement.

Returning to how persons may actually respond when confronted with different categories to what is ostensibly the same item, it is again agreed that this must be determined empirically. Perhaps there is a limit to how many categories people can operate with in different circumstances. For example, if people are given 10 categories rather than 5 in some situation, it may be that there is no increase in precision: having more categories available does not ensure that the categories are used correctly and that they will give greater precision. It may be that the discrimination between some categories is 0, and that they should then be grouped into 5 categories. That is, it may be determined empirically that the instrument is not more precise with the 10 given categories than with 5 categories. However, if respondents do use the 10 categories in such a way that their responses are more precise and give more information about their location than when they use 5 categories, then the joining assumption will be destroyed in the data. This insight is gained from the analysis of the MPM, and generalized to the URM. To emphasize the distinctive perspective, I would prefer that the joining assumption did not hold: if it did hold, then it would mean that the extra categories were not providing extra information. Of course, nothing prevents persons from throwing away information, but if the data conform to the URM, then dichotomized data will not conform to the URM and the advantages in using the model cannot properly be exploited. In practice, and for some purposes, the effects may be small, and there may be other reasons why it is convenient to dichotomize and group categories, but that is a different matter from understanding what aspect of the URM prevents the joining assumption from holding. To reiterate, an item with 10 categories which operates as intended, is in a material sense a different item from one with 5 categories, *most importantly even when the rest of the wording in the item is the same*, in which case the variable measured is the same. The number and definition of categories is an integral part of the item: changing these changes the item from the point of view of the response process and its effect on the distribution of the responses.

The above analysis leads to a conclusion directly opposite to the comment "Neither ξ-invariance nor the joining assumption have anything to do with parameter estimation." (p. 30) in the rejoinder. They have everything to do with parameter estimation in the sense that in the URM the parameter estimate is the same (ξ-invariance holds), but the precision of this estimate changes and the joining assumption fails.

5. Integer Scores and Meaning of Units

Roskam notices that the integer scoring plays a role in the analysis of the MPM, but over-interprets this as designating successive response categories. "Andrich shows that the distribution of responses is affected by what is called the 'size of the unit' (ω_i). He is neither specific about what is meant by the 'size of the unit,' nor about the response categories involved. He designates the response categories by integers, and he defines the 'distance' between the 'thresholds' or category boundaries as equal and item specific (ω_i)." (p. 32). It is important to appreciate that the integers do not appear because the distances between thresholds are equal. The integers appear in the URM *as a consequence of the equal discriminations at the thresholds*, and not as a consequence of the equidistance between the thresholds—they are a count of the num-

ber of equally discriminating thresholds exceeded. Thus in the URM, the scoring function is the set of positive integers, even though the number of categories is finite, and even though the thresholds may not be equidistant and are to be estimated. The equidistant thresholds in the multiplicative metric means that instead of having a parameter for each threshold, it is possible to use a single parameter which is the distance between all successive thresholds, and this simplifies the expression. By analogy to physical measurement, this common distance is called the unit.

This count of the number of thresholds is exploited in the case of the MPM when the thresholds are equidistant because one can say that if one exceeds k thresholds in half units, then it is equivalent to exceeding $k/2$ thresholds in the original units. For example, if one exceeds 50 thresholds in the units of half a yard, then this is equivalent to exceeding 25 thresholds expressed in units of a yard. This reexpression is simple to appreciate when the number of thresholds exceeded in half units is an even integer. If it is an odd integer, then the reexpression is generalized to give the equivalent rational number in the original units but many of these could not be realised if the instrument actually operated in the original units. For example, if the number of thresholds in half units exceeded is 51, then this is equivalent to 25.5 original units, even though when responses are in original units one cannot actually record exceeding 25.5 units. This is a key point: it explains why when different sized categories work differently from the original categories, they provide a different distribution giving a more precise estimate of the location of ξ which in turn violates the joining assumption. To reiterate, the integers arise out of the specification of invariance of comparisons and the requirement of sufficient statistics—they are not simply assigned to the categories and they do not rest on the condition of equal distances between the thresholds specified in the MPM.

6. Normal Approximation to the Poisson

Finally, Roskam refers to Müller's extension of the rating model[1] to a continuous case, and says that it might salvage the joining assumption and the MPM. This is exactly what I consider it will not do, and reflects a distinct incompatibility between our perspectives. Although Müller derives the model by making the thresholds get infinitely close together, dichotomizing, or polychotomizing a continuous distribution is exactly what is not what is reflected in the Rasch models. Although the MPM tends to the normal as the unit size tends to 0, it always characterises a discrete distribution whose mean is related to the variance—in contrast, the normal has two parameters, a mean and a variance and, in dichotomizing it after the fact, the variance of this distribution is explicitly fixed and the joining assumption is satisfied. This, as noted by Roskam, is the basis of the so-called Samejima (1969) model. Rather than taking discrete categories as approximations to some continuous process, from the analysis in the paper, the measurements are always discrete no matter how small the unit, and imposing a continuous distribution is seen as the approximation, rather than the other way around. If a process is characterised by a Rasch model, then it cannot be characterised by the Samejima model—the processes are incompatible.

There are many further issues to be considered, some of which, including Roskam's observations of the effects on the variance when course groupings are made after the fact, are explored in Andrich (1993c) and are as yet unpublished formally. In fact, this observation in the rejoinder is particularly relevant, and may usefully illustrate the kind of relocation of observations when categories are grouped before the data are collected and the effects it can have on the variance and, thereby, on the joining

[1] I am very familiar with this work, Müller having used the reparameterisation (Andrich, 1982), which he then extended excellently to the continuous case.

assumption. That there is a relationship to this well-known effect is not trivial, and it may help understand further the otherwise initially counterintuitive result that in measurement, and in psychological response processes, the joining assumption should not hold if more categories are intended to provide more precise estimates.

7. Summary

I have tried to indicate not only the analytic differences in some of the points made in the paper and the rejoinder, but also the perspective from which these differences arise. The point of the paper, and these remarks, is to show how the special case of the MPM revealed why the joining assumption and dichotomization are not, in general, properties of the URM for graded responses, and, thereby to identify the circumstances where one would require that this property did not hold in empirical graded responses.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69–81.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 357–374.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47*, 105–113.

Andrich, D. (1988, April). *A scientific revolution in social measurement.* First meeting of the American Educational Research Association's Special Interest Group on Rasch Measurement, New Orleans.

Andrich, D. (1993a). *Where linguistics, physical measurement and social measurement converge.* Paper presented at the Fourth International Conference on Bibliometrics and Scientometrics, Berlin.

Andrich, D. (1993b). *Theoretical and empirical evidence on the dichotomization of graded responses.* Paper presented at the Annual Meeting of the AERA, Atlanta, Georgia.

Andrich, D. (1993c). A law of the distribution of measurement error as a function of the effective unit of measurement (Report No. 3). Murdoch, Australia: Murdoch University, Social Measurement Laboratory.

Andrich, D. (1995). Models for measurement, precision, and the non-dichotomization of graded responses. *Psychometrika, 60*, 7–26.

Harris, J. (1991). Consequences for social measurement of collapsing adjacent categories within items with three or more ordered categories. Unpublished Master of Education dissertation, Murdoch University, Australia.

Jansen, P. G. W., & Roskam, E. E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika, 51*, 69–91.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed. enlarged.) Chicago: The University of Chicago Press.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika, 52*, 165–182.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 89–108). Chicago: Science Research Associates.

Roskam, E. E. (1995). Graded responses and joining categories: A rejoinder to Andrich's "Models for measurement, precision, and the nondichotimization of graded responses. *Psychometrika, 60*, 27–35.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs, 34, 2*, (No. 17).

Saussure, F. (1959). *Course in general linguistics.* London: Peter Owen.

Manuscript received 6/30/94
Final version received 7/1/94