Transitional Categories and Usefully Disordered Thresholds

by

John Michael Linacre
Research Director, Winsteps.com

Abstract:

The definition and analysis of rating scales or partial-credit scales is central to research in the social sciences. Rating scale categories are intended to manifest latent variables in an ordered, exclusive and exhaustive manner. Some rating-scale categories may be infrequent, transitional categories between dominant categories. Prior advice has been to eliminate low-probability categories whose Rasch-Andrich thresholds are disordered. Examples in the physical science, and a hypothetical example in the social sciences, suggest that the prior advice may be counter-productive. The advance of social science may well require the analysis of more and narrower transitional categories, and the unambiguous communication of the implications of those categories to the target audience.

Introduction:

The definition and functioning of the categories of rating scales (or partial-credit items) are crucial for making inferences from polytomous datasets. In the minds of raters and respondents, the categories function best when they are perceived to be discrete, specific and of equal size, as depicted in Figure 1.

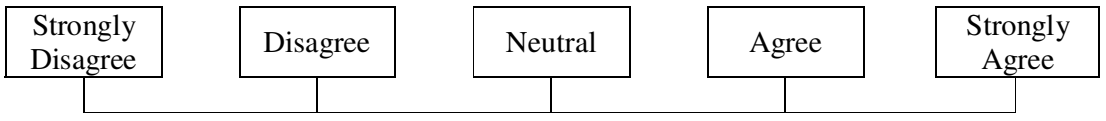


Figure 1. A Likert Scale as perceived by raters.

Under Rasch model conditions, the categories of a rating scale or partial-credit item are modeled to be qualitatively ordered along an infinite unidimensional latent variable. The categories are mutually exclusive and exhaust the latent variable. The categories are of unequal size. The top and bottom categories are always infinitely wide. Figure 2 illustrates this.

Figure 2 already suggests a focus of investigation. What is a relevant category and what is not? For instance, is "Neutral" a relevant category on an Agreement scale or is it an opportunity for the respondent to avoid giving a substantive rating? In socially-sensitive situations, "Neutral" may even be a subtle way of indicating disagreement with the socially-accepted norm which pressures respondents to respond "Agree", or preferably "Strongly Agree". The ambiguities in a rating scale worsen when the middle category is labeled "Don't Know". This compounds two dimensions into one rating scale: Agreement-Disagreement and Knowledge-Ignorance. A similar confusion of dimensions arises when the middle category is labeled "Not applicable". What statistical indicators

← Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree →

Figure 2. A Likert Scale as modeled for Rasch rating-scale analysis

← Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree →
← 0	1	2	3	4 →

T_1 T_2 T_3 T_4

Figure 3. Category scoring and category thresholds, T_2, T_3, ...

confirm that the rating-scale categories are ordered, exclusive and exhaustive? Only one aspect of this line of inquiry will be followed here, but already the challenges encountered in extracting and communicating meaning from rating-scale data are becoming evident.

An increasingly perplexing aspect of rating scale construction and analysis is the method of depicting and communicating categories which occupy only a small interval on the latent variable. Perhaps Figure 2 depicts the rating scale for a polarizing issue in society in which the frequency of observing "Neutral" is very low, so that the interval corresponding to "Neutral" on the latent variable is very narrow. Perhaps the "Neutral" category is a transition category, occupied by those few people who are transitioning from one extreme viewpoint to the other. These people may be crucial for investigation by opinion-leaders at both ends of the spectrum, because these people embody the reasons for rejecting one attitude and accepting the other attitude. Previous recommendations, such as Linacre (2002), have encouraged analysts to eliminate transitional categories from their datasets, but this paper suggests that those recommendations may be counter-productive to the advance of social science.

Figure 3 illustrates that, for the purposes of Rasch analysis, the categories are scored with sequential cardinal numbers, usually 0, 1, 2, .. or 1, 2, 3, ... The numbers ascend in the direction of more of the conceptual latent variable. The category boundaries between categories are termed "thresholds", numbered by the higher of the adjacent categories, for instance, T_1, T_2, ... Figure 3 also raises important questions. T_1, T_2, .. are the thresholds at the boundaries of the categories, but what is the exact definition of a "threshold"? And what becomes of the definition when a category is rarely or never observed in a dataset? This paper suggests several definitions, each supporting different inferences based on the functioning of the rating scale.

Transitional Categories: Explosive Examples

Transitional categories are usually narrow intervals on the latent variable representing the transition between dominant categories, but they can be central for inference. In many situations, it is vital that the specific details of transitional categories be preserved during data analysis. Otherwise incorrect, even dangerous, inferences may be drawn from the data.

"Liquid" is a transitional category between "solid" and "gas". For instance, "water" is solid "ice", below 0°C and gaseous "steam" above 100°C. Water is liquid for only a narrow temperature range,

← Too lean →	← Explosive range → (transitional category)	← Too rich →

0% saturation T_1 = LEL Lower Explosive Limit T_2=UEL Upper Explosive Limit 100% saturation

Figure 4. Explosive state of fuel vapor in a fuel tank

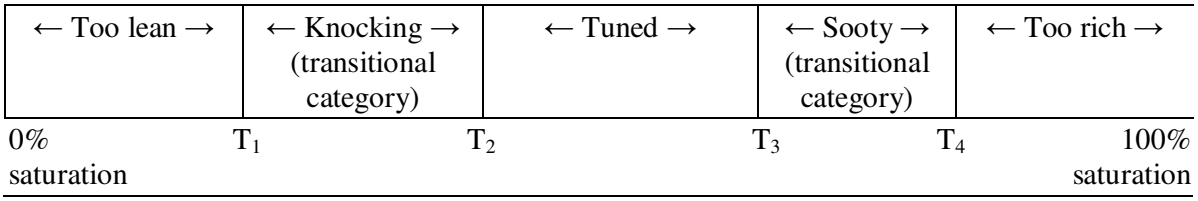


Figure 5. Explosive state of fuel vapor in an internal combustion engine

but many decisions of daily life are based on the probability that water will be solid, liquid or gas in a particular situation. The inferential process for water appears to be exact because it has been so thoroughly investigated and is so familiar to us, but in other life-threatening situations the probabilities and their inferences are much less clear-cut.

When fuel tanks are being filled or emptied, cleaned or repaired, the air space above the fuel is filled with fuel vapor. This can be explosive. It is vital for workers to know when the vapor is potentially explosive and when it is not. Figure 4 (based on Alberta, 2010) shows the explosive state as a transitional category between 0% saturation of the air with fuel to 100% saturation.

In Figure 4, T_1, the "Lower Explosive Limit" (LEL), is the lower threshold of the transitional "explosive range" between "too lean" and "too rich", and T_2, the "Upper Explosive Limit" (UEL), is the upper threshold. Figure 4 is presented deterministically, but the process is probabilistic. There are factors which facilitate explosions, such as open flames, and also which inhibit explosions, such as lower ambient temperature. Safety procedures are intended to lower the probability of observing an actual explosion down to zero. However, it would be dangerously misleading to eliminate the "explosive" category from an analysis of fuel-vapor saturation, no matter how rarely explosions are observed.

Figure 4, however, can be interpreted in an opposite manner. Internal combustion engines depend on explosive fuel-air mixtures. The mixing device, such as a carburetor, must be adjusted so that it emits fuel vapor that will be in the explosive range during engine operation. For carburetor adjustment, Figure 4 is too vague. More categories are needed. Figure 5 illustrates this. The original transitional category, "Explosive range", category has become the "Tuned" category bracketed by two more transitional categories "Knocking" and "Sooty". Engine mechanics are trained to recognize all three transitional categories, and then to infer what adjustments are required to improve engine functioning. Again, elimination of any of the categories from an analysis of carburetor behavior would be misleading, no matter how rarely any of the categories are observed.

These everyday examples illustrate the importance of recognizing transitional categories and making inferences based on them. This is true regardless of the width of the transitional category, such as the "explosive range" or the probability of observing their consequences, such as an explosion.

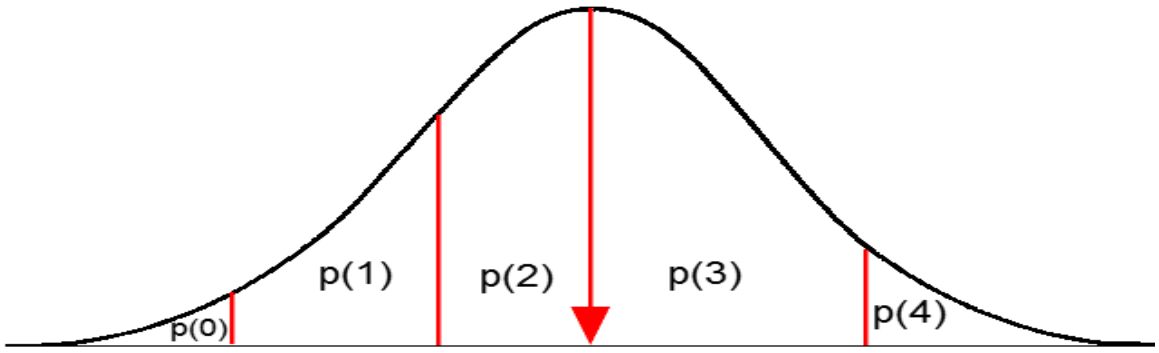
Conceptualizing Thresholds

The discussion of the "explosive range" assumed that the threshold values, T_1 etc., can be substantively defined and then determined through relevant physical experiments. Psychological processes are initially hazy, as were physical processes originally. Practitioners observe a transitional category based on its consequences, such as an explosion, and then they, along with scientists, investigate and isolate the key elements of the process which caused the consequences. This may not be easy, even in physics. For instance, investigation into familiar physical transitional states such as car accidents and lightning strikes, and even improbable states such as earthquakes and asteroid

strikes, continues worldwide. Ultimately useful definitions for the categories are agreed and their thresholds quantified. In psychometrics, this ultimate stage has not yet been reached. Consequences have been, and continue to be, observed, such as success and failure in every day situations requiring simple arithmetic. Based on these consequences, provisional latent variables have proposed, such as "math ability". Simultaneously scientific probes of math ability have been devised, such as items on a math test.

On first view, the response mechanism to a math item is trivial. There is a two-category rating scale. The lower category is "wrong" and the upper category is "right". The threshold between them is at point where the respondent has a 50% probability of obtaining a "right" answer and a 50% probability of obtaining a "wrong" answer. But already arbitrary and probabilistic aspects of this two-category rating scale are emerging. What is the correct answer and who decides whether it is correct? Does 1+1 = 2 (decimal) or 1+1 = 10 (binary)? What if the respondent answers correctly or incorrectly due to irrelevant factors such as a fire-alarm sounding? And is 50% probability a useful threshold value? Consider the "explosion" rating scale in Figure 3, safety officers would not want LEL reported as the value where there is a 50% chance of an explosive saturation. Substantively this is close to the "flash point" for the fuel vapor. Instead they want the LEL to be the point where there is effectively no chance of an explosion. This may be 1 percentage point or more of saturation lower than the flash point. In societal terms, 50% probability of success on a math item is not good enough. "Mastery" is required for practical purposes. 100% mastery is impossible to attain, but 80% mastery (4 successes in every 5 attempts) is good enough, particularly if the learner is taught to verify their own work in crucial situation. Then the expected failure rate of 20% might approach 20% * 20% = 4%, so that the combined mastery level could be 96%. Accordingly, a more useful threshold level between "right" and "wrong" for a math item could be 80% chance of success, rather than 50% chance of success.

For polytomous rating scales, the scoring of the categories and the choice of threshold definition are more complex, but equally essential, if the findings of an analysis are to understood by the target audience, and then to become productive bases for decision-making.



← Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree →
← 0	1	2	3	4 →

Rasch-Thurstone: T_1 T_2 T_3 T_4

Figure 6. The curve represents the probability density of a continuous latent random variable induced by the reaction of a person to an item. The observed rating for that person is determined by the interval in which the random variable falls. (After Andrich, 1978, Figure 1). Thresholds T_1, T_2, ... are the Rasch-Thurstone thresholds. p(0), p(1), ... sum to 1.0.

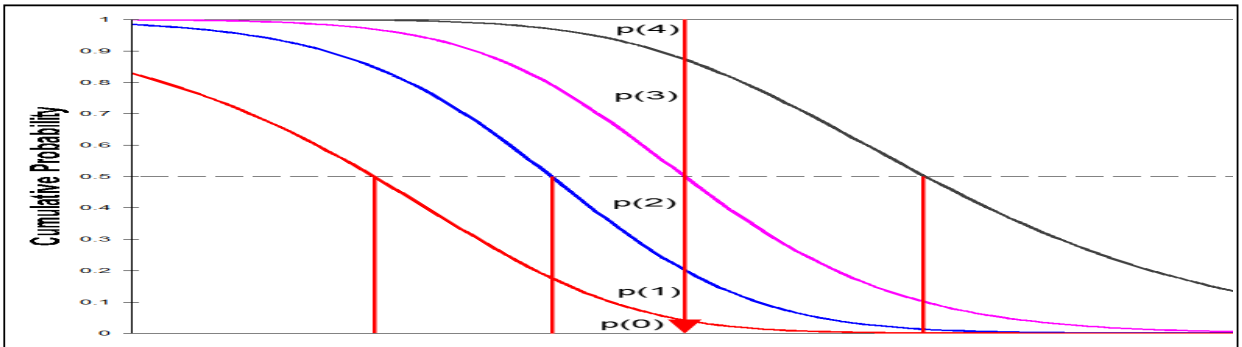
Definitions of Rasch-based Thresholds:

Andrich (1978) implicitly proposes three definitions for the thresholds of a rating-scale category. The first definition is illustrated in Figure 6. It shows the probability density for the possible observed ratings for one respondent to one item. The x-axis is the latent variable. The respondent is model as a random variable with a finite probability of being observed anywhere along the latent variable. The bounded areas under the probability-density curve are the category probabilities, p(0), p(1), ... which all sum to 1.0. The boundaries between the areas are the thresholds. This definition is also implied in Thurstone and Chave (1928, Fig. 1 and p. 15): "A ... characteristic might also be indicated graphically in terms of the scale, namely, the range of opinions that any particular individual is willing to indorse. ... [and] the mean position that he occupies on the scale."

In Figure 6, inspect T_3, the threshold between category 2, "Neutral" and category 3, "Agree". We can see that, in this example, T_3 corresponds to the mean position of this respondent on the rating scale. The respondent has 50% probability of being observed in categories 0, 1, 2 and 50% probability of being observed in categories 3 and 4. In a practical situation, we might consider categories 0, 1 and 2 as "wrong" and categories 3 and 4 as "right", then Threshold T_3 approximates the threshold of an equivalent dichotomous "Agree or not" item.

In this paper, thresholds defined in this manner are termed "Rasch-Thurstone" thresholds. A respondent whose position on the latent variable aligns with a Rasch-Thurstone threshold for an item has a 50% probability of being observed in the categories above the threshold and a 50% probability of being observed in the categories below the threshold. Though the underlying process is probabilistic, the shape of the bell-shaped curve in Figure 6 is not normal but, for our purposes, is determined by the relevant polytomous Rasch model.

Figure 6 shows the probability intervals along the x-axis corresponding to observations made at one point on the latent variable. This is generalized in Figure 7 to all points on the latent variable. The probability intervals are now shown vertically for each point on the latent variable, depicted as the x-axis. At each point on the latent variable, every category has a finite probability of being observed, and, since the categories are exclusive and exhaustive, the sum of the category probabilities is 1.0. In polytomous Rasch models, adjacent categories are modeled to be local Rasch dichotomies, a consequence is that the cumulative probability curves are monotonically descending when summed



← Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree →
← 0	1	2	3	4 →

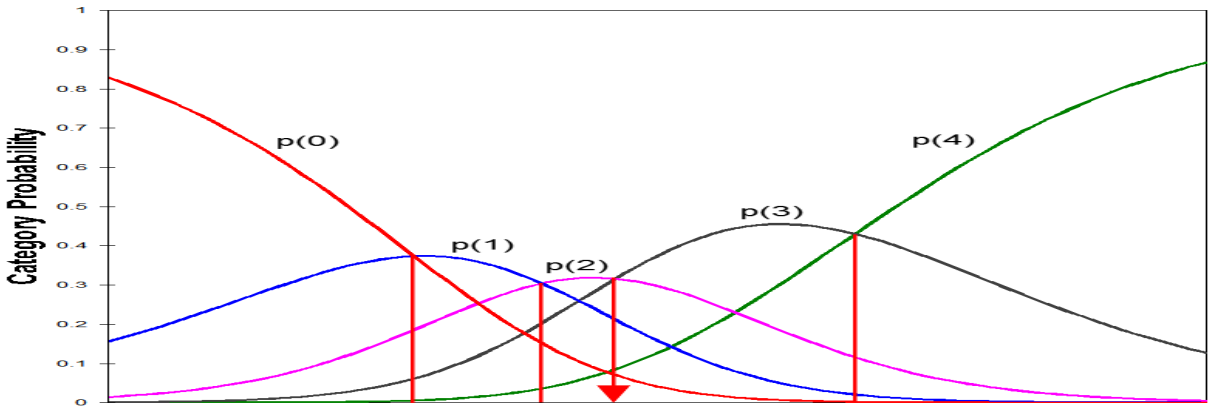
Rasch-Thurstone: T_1 T_2 T_3 T_4

Figure 7. The cumulative probabilities for the responses to one item by respondents located at each point on the latent variable. Each Rasch-Thurstone threshold is the point on the latent variable where the cumulative category probabilities for categories below the threshold are 0.5.

from the lowest category upward.

Andrich (1978) derives the Rasch rating-scale model for polytomous data, and then proposes another definition for category thresholds based on it. Figure 8 illustrates this definition. The thresholds, A_1, A_2, ... are here termed the "Rasch-Andrich" thresholds. They are located at the points on the latent variable where adjacent categories are equally probable. These thresholds manifest the same category probabilities as the Rasch-Thurstone thresholds, but expressed by individual category rather than cumulatively. Consequently the values of the Rasch-Andrich thresholds differ from the values of the Rasch-Thurstone thresholds. In Figure 8, the Rasch-Andrich thresholds are more central than the Rasch-Thurstone thresholds, so that, according to this definition, the category intervals are narrower. Notice that the Rasch-Andrich threshold A_3, between "Neutral" and "Agree" is offset relative to the cumulative Rasch-Thurstone threshold T_3. The decision points for "Agree" versus "Not Agree" differ.

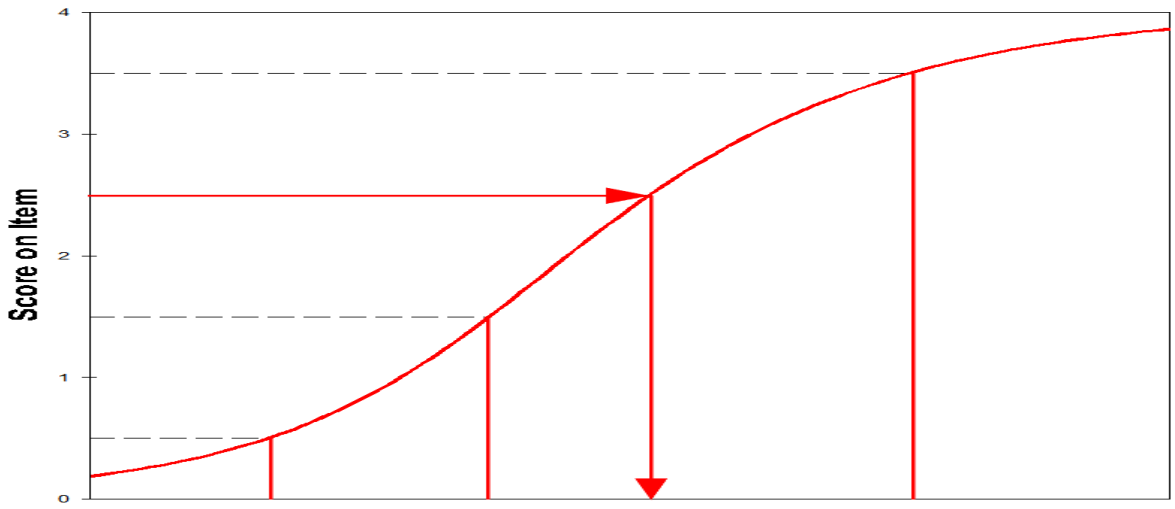
Andrich (1978) also discusses the "integral scoring functioning". For Rasch models, this is based on the observed score on the item by the respondent, in other words, the sum of the observed category values. The Rasch-Andrich model also allows us to compute the expected score on each item at each point on the latent variable. These are shown in Figure 9. Thresholds H_1, H_2, ... are the "Rasch-half-point" thresholds. The thresholds bound the intervals on the latent variable in which the half-rounded expected score is the category value. For instance, for a respondent located at Rasch half-point threshold H_1 relative to the item difficulty, the expected score on the item is 0.5. At H_2, the expected score is 1.5. So, between H_1 and H_2, the half-rounded expected score is 1.0, the value corresponding to category 1. The expected score below H_3 is less than 2.5 (which half-rounds to 2 or less), and the expected score above H_3 is greater than 2.5 (which half-rounds to 3 or more). Notice that in Figure 9, the Rasch-half-point thresholds are more dispersed than the Rasch-Thurstone thresholds, so that the category intervals are wider.
Three definitions for rating-scale category intervals have been presented here. They are related



← Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree →
← 0	1	2	3	4 →

Rasch-Thurstone:	T_1	T_2	T_3	T_4
Rasch-Andrich:	A_1	A_2	A_3	A_4
← 0	1	2	3	4 →

Figure 8. The curves represent the probability of observing each category of an item at each point of the latent variable. (After Andrich, 1978, Figure 3). The probability values at each point on the latent variable always sum to 1.0 vertically. Thresholds A_1, A_2, ... are the Rasch-Andrich thresholds.



← Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree →
← 0	1	2	3	4 →

Rasch-Thurstone:	T_1	T_2	T_3	T_4

Rasch-Andrich:	A_1	A_2	A_3	A_4
← 0	1	2	3	4 →

Rasch-half-point:	H_1	H_2	H_3	H_4
← 0	1	2	3	4 →

Figure 9. The curve is the expected-score ogive for an item. It is the Rasch-model-predicted item characteristic curve (ICC) for the item. Thresholds H_1, H_2, ... are the Rasch-half-point thresholds, corresponding to expected scores on the item with half-score-points..

mathematically. For a rating scale with categories 0,m, then
at Rasch-Thurstone threshold T_k, $p(0)+...+p(k-1) = p(k)+...+p(m)$
at Rasch-Andrich threshold A_k, $p(k-1) = p(k)$
at Rasch-half-point threshold H_k, $\text{sum}(j*p(j)) = k - 0.5$ for $j=0,m$

There are other definitions for category intervals, such as, for instance, the intervals on the latent variable in which each category is the most statistically informative. For the rating scale in Figure 9, the category-information thresholds are more central than the Rasch-Andrich thresholds.

For estimation purposes, the Rasch-Andrich thresholds provide the simplest parameterization of a polytomous Rasch model and so are the most convenient. But deciding which threshold formulation is best for inference and communication is more challenging. Much depends on the thinking-processes of the intended audience and the decisions that are to be made.

For transition categories, the Rasch-Thurstone thresholds are usually the least confusing. In particular, when a crucial transition category is rarely observed in a dataset, only the Rasch-Thurstone thresholds indicate to a non-technical audience that the transition category is squeezed between two dominant categories. The Rasch-half-point categories make the category appear too wide. The Rasch-Andrich thresholds make the transition category appear in some way negative or ignorable.

A Hypothetical Example of Thresholds and Transitional Categories: Judge Agreement

Transitional categories correspond to narrow intervals on the latent variable. They may indicate growth states of short duration. They are usually less frequently observed than the neighboring dominant categories. A result is that the probability of observing transitional categories tends to be low, but, as in the "explosion" example, users of the measures certainly need to be aware of where on the latent variable there is some probability of them occurring.

Imagine this hypothetical, but realistic, situation. A sample of candidates of differing abilities perform a series of items (tasks) of differing difficulties. Two judges independently rate the candidate performances with pass = 1 and fail = 0. The recorded score for each candidate on each item is the sum of the two judges' ratings.

For this hypothetical example, there are 50 candidates with a uniform distribution of ability from -4.5 to +3.5 logits, 30 items with a uniform distribution of difficulty from -3 to +3 logits, and two judges, one more lenient at -0.5 logits, and the other more severe at +0.5 logits. Each judge's 0-1 ratings are simulated with the Rasch dichotomous model:

$$\log(\text{Probability of observing 1} / \text{Probability of observing 0}) = \text{Candidate ability} - \text{Item difficulty} - \text{Judge severity} \quad (1)$$

The pair of ratings for each candidate-item performance are summed into a 0-1-2 observation, and these data are analyzed with the Rasch-Andrich Rating Scale Model. Figure 10 shows the resulting probability density curves. 0 indicates that both judges' ratings are 0, a clear fail. 2 indicates that both ratings are 1, a clear pass. 1 indicates that one judge's rating is 0 and the other judge's rating is 1. The judges disagree. This is an ambivalent transition state between passing and failing on the item. The points of equal probability between adjacent categories are the Rasch-Andrich thresholds, arrowed in Figure 10. Figure 11 shows the same information but depicted as cumulative probability density curves, rather than as individual category probability density curves. The points of 50% cumulative probability between higher and lower categories are the Rasch-Thurstone thresholds, arrowed in Figure 11.

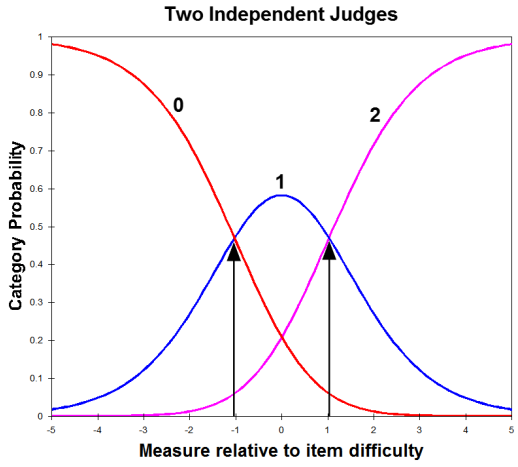


Figure 10. Probability of summed scores given by two independent judges. The Rasch-Andrich thresholds (black arrows) are ordered, 0-1 is to the left of 1-2.

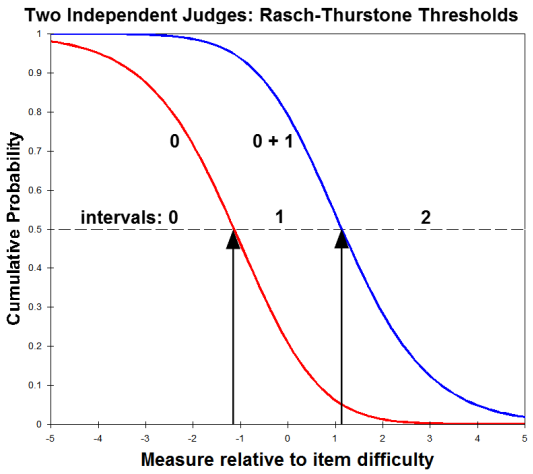


Figure 11. Accumulated probability of summed scores given by two independent judges. The Rasch-Thurstone thresholds (black arrows) are always ordered, 0-1 is to the left of 1-2.

The Examination Board sees Figure 11 and are alarmed that there is a 2.6 logit transition-interval between the judges agreeing (with probability greater than 50%) on failure (0) for a low performer and agreeing on success (1) for a high performer. This is a considerable distance on the latent variable, perhaps equivalent to more than two years of academic growth. Accordingly they devise a scheme. Whenever the judges disagree, they will be asked to re-rate the item, again independently, and their revised ratings, changed or not, disagreeing or not, will be accepted. The judges are not told whether their revised ratings agree or not. In this hypothetical example, the relevant ratings are re-simulated using Equation (1).

Figure 12 shows the resulting probability curves. Category 1, indicating judge disagreement, is never modal. It is never the most probable category at any point on the latent variable. Its probability density curve is never higher than those of the other categories. Consequently its points of equal probability with adjacent categories, the Rasch-Andrich thresholds, are disordered relative to the latent variable. How are these curves related to intervals on the latent variable? Does threshold disordering imply that the interval for category 1 on the latent variable has become negative?

Figure 13 shows the cumulative probability curves and the Rasch-Thurstone thresholds. The judges continue to rate independently, but the transition-interval representing category 1, which is between the Rasch-Thurstone thresholds, has reduced from 2.6 logits to 1.6 logits. It is positive and continues to be somewhat large. The Examination Board can see this clearly. A one logit reduction in the threshold-interval is not as big as the Examination Board would have liked, but is definitely an improvement.

If a continued process of rerating reduced the frequency of a combined score of 1 yet further, what would have happen? In Figure 12, the peak of the probability density curve for "1" would become lower, so that the interval between the disordered Rasch-Andrich thresholds would be wider and more negative. In contrast, as the probability of observing "1" reduces, the cumulative curves in

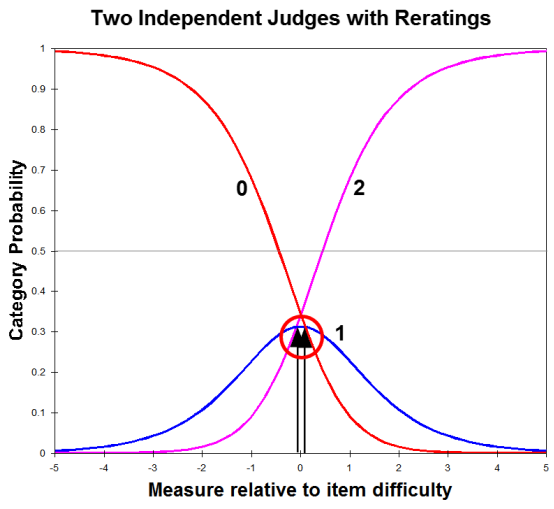


Figure 12. Probability of summed scores given by two independent judges. When the judges originally disagreed (category 1), the judges were instructed to re-rate independently once. The Rasch-Andrich thresholds (black arrows in red circle) are disordered, 1-2 is to the left of 0-1.

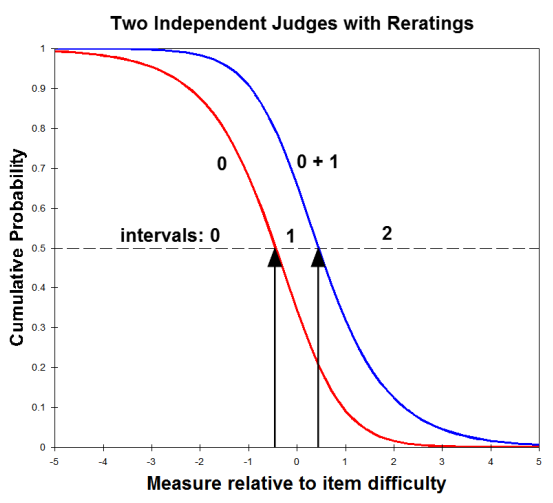


Figure 13. Accumulated probability of summed scores given by two independent judges with re-rating. The Rasch-Thurstone thresholds (black arrows) are always ordered, 0-1 is to the left of 1-2.

Figure 13 become closer together. For a non-technical audience, the "smaller" interval in Figure 13 is much each easier to understand than the "larger" negative interval in Figure 12, despite the fact that Figures 12 and 13 depict exactly the same probabilities.

The Examination Board considers reducing the transition-interval further by rescoring all combined scores of 1. Automatically scoring such a candidate as a clear pass (2) or a clear failure (0) would nullify one of the judges and provide a misleadingly good or bad report about the candidate's performance . Scoring the judge's decision for the candidate as "undecided" (missing data) would lose the information that the candidate's performance is close to the pass-fail point for the item. The Board also considers instructing the judges to rerate disagreements again, but decides that this would provoke judge misbehavior. The Board's decision is that a combined score of 1, after rerating, indicates genuine, not merely accidental, disagreement between the judges about the competence of a candidate who is close to the pass-fail decision on the item. The existence of a combined score of 1 would be useful information when the Board ultimately makes decisions about candidates whose overall performance is very close to a criterion cut-point.

This is a hypothetical situation, but it does illustrate that non-modal categories, with their disordered thresholds, are not necessarily deleterious to measurement, but can be advantageous because they indicate ambivalence in the transition between two states, in this case a clear fail and a clear pass. The example also illustrates that the cumulative probability curves and Rasch-Thurstone thresholds are more effective for communicating to non-technical audiences about transitional categories than the probability density curves and Rasch-Andrich thresholds.

Conclusion:

In the physical sciences, improvements in measurement enable the substantive identification of narrower and narrower categories on the relevant measurement scale. The equivalent process in the social sciences leads to the substantive identification of more and more transition categories. The "explosive" example shows how one transition category in the measurement of explosive-saturation developed into three transition categories with the invention of the internal combustion engine. Surely the social sciences will advance in a similar way. Better measurement will facilitate advances in substantive theory and practice, which will lead to rating scales and other data-collection devices with more and more qualitatively ordered categories, many of which will be narrow transition categories. Accordingly, this paper suggests that transition categories should no longer be viewed as threats to valid measurement, but rather as an integral and increasingly important part of the advance of social science. Consequently, social-science measurement must improve the analysis and communication of transitional categories rather than attempt to eliminate them.

References:

Andrich, D. A. (1978) A rating formulation for ordered response categories. *Psychometrika, 43, 4,* 561-573.

Linacre, J. M. (2002) Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement 3, 1,* 85-106.

Thurstone, L. L. & Chave, E. J. (1928) *The Measurement of Attitude.* Chicago: University of Chicago Press.

Government of Alberta, Employment and Immigration (2010) *Workplace Health and Safety Bulletin: Controlling Explosive Atmospheres in Vessels, Tanks and Piping Systems.* Alberta, Canada: author.