MAIMON WORKINGPAPERS No. 13 AUGUST 2023

VALID CLAIMS FOR THERAPY RESPONSE: THE RASCH RATING SCALE POLYTOMOUS MODEL

Paul C. Langley, Ph.D., Adjunct Professor, College of Pharmacy, University of Minnesota, Minneapolis, MN

ABSTRACT

It has long been recognized that if we are to create valid claims for therapy response based on subjective counts and observations, that the Rasch model is the only valid basis for establishing those claims. The Rasch model is the only analytical framework that is necessary and sufficient to transform ordinal responses to interval scores. Most importantly, unlike any other instrument that claims to produce a useful scale to assess response to therapy, the Rasch model creates the requirement for unidimensional or single attribute measures with linear, interval and invariant properties. The question of interest is how analysts should proceed from the Rasch output in logits to a scale that is more attractive to the user in making claims for therapy response. A proposed solution to this was demonstrated in a recent commentary where, for the dichotomous Rasch model, a logistic transformation to probability weights provided the basis for claims framed in terms of the possession of a manifested latent trait. In this commentary, the argument is taken one step further in demonstrating how the same analysis can be applied to the Rasch analysis of instruments with ordinal, polytomous responses.

INTRODUCTION

It is now accepted, for those with an understanding of the importance of measurement theory, that if we are to evaluate the response to therapy based on subjective or ordinal responses from patient reported outcomes (PRO) instruments that we need to transform those observations into measurement scales with single attribute or unidimensional, linear, interval and invariant properties. This transformation, which has been recognized by measurement theorists for over 60 years, is unique as the necessary and sufficient condition that focus on therapy response in probabilistic terms where the successful response to either a dichotomous or polytomous item structure must be a function of the difference between item difficulty and respondent ability [i]. This is widely accepted with the unfortunate implication that the overwhelming majority of PRO claims in health technology assessment (HTA) fail this standard. The Rasch framework for modeling subjective responses, transforming them from ordinal to interval scales, is unique; we have no option but to apply Rasch standards [ii iii].

A recent commentary in this series proposed a solution to make Rasch output more tractable for users who may be considering applying standard statistical tools to support claims for therapy response [iv]. In the case of a dichotomous modeling, the standard Rasch respondent abilities and

item difficulty on a common logit scale where, centered on zero, the logits typically occupy a range from +/- 3.5 logits. It was proposed that, by application of a logistic function, the various logit values for item difficulty could be transformed to probabilities without impacting the Rasch measure and its characteristics. It was further proposed that these probabilities could be considered weights to be applied to item responses (true/not true) to assess the extent to which the respondents to the questionnaire could be said to possess the manifestation of a latent trait captured by the instrument. As probability weights applied to items it was possible to estimate the proportion of the latent manifestation possessed by each respondent, by simple item response weight addition divided by total sum of weights as a possession scale. The final step was to apply standard statistical assessment for mean value and standard deviation to support, at baseline, the average trait possession and then, following a therapy intervention, the extent to which the possession had 'improved' together with estimates of effect size. It is to be noted that statistical operations are only allowed if a measure has unidimensional, linear and interval properties which is achieved only the result of applying Rasch standards to subjective ordinal counts or observations.

THE RASCH POLYTOMOUS MODEL

The Rasch analysis of polytomous instruments, where items differ by difficulty as required by Rasch, the item difficulty is estimated from initial counts of the response categories for the item. However, as a polytomous or Likert-type instrument each item asks for an ordinal response for what we may call the extent of possession of the latent trait given the difficulty thresholds of the item. The issue addressed by the Rasch model is how to accommodate the ordinal responses for each item to create a logit scale with the required properties. The solution is straightforward: the simplest approach, the Rasch Rating Scale Model, establishes the relative difficulty of each item in an instrument, as with the dichotomous model, but with the Likert-type category structure of possible responses given a single rating scale structure common to all the items on the scale. This enables us to infer that the response categories for the Likert items are the same in terms of the differences between the difficulty thresholds. It is important to note that the items are then ranked by increasing difficulty of responding to a particular item's threshold.

The term threshold needs to be made clear. The threshold, measured in logits, is on a scale identical to that for dichotomous responses; a common logit scale, centered on zero, that captures both item difficulty, the difficulty of the individual item response level, and the ability of the respondent. The threshold logit is the level at which the probability of being observed in a given response category below the threshold is exceeded by the probability of being observed in the next highest response category. That is, to express it somewhat differently, the level at which the probability of failure to endorse a given response category, below the logit, turns to the probability of successfully endorsing the category above the threshold. The response category selected by the respondent given the options of a Likert scale can then be interpreted in terms of failing to agree with the other response categories on offer (e.g., choosing response category B from the option of ranked response categories from A < B < C < D). .It is worth remembering that the logit scale is also a probability scale as each logit by applying a logistic function can be transformed to a probability scale with the same measurement properties as the logit scale. This s critical if we are to develop a meaningful claim for therapy response.

Where there are four response options there will be three thresholds. That is a threshold for A to B, a threshold from B to C and a threshold from C to D. To emphasize the threshold approach: the Rasch model says nothing about the size of the step necessary to move across each threshold. The Rasch modelling detects the threshold structure in the data set and estimates a set of common thresholds that apply to all items in the questionnaire. The Partial Credit Rasch model relaxes this requirement for common thresholds by allowing differing numbers of response levels for different items or different item threshold where each item has the same number of response categories. The modelling so0lution proposed here can be extended to both of these options.

It is important to make clear how the approach in the Rasch model contrasts with the usual *add em-up* approach to Likert based instruments. The traditional approach of integer summation assumes all items are of equal difficulty and that the thresholds between steps are on equal distance or value. Neither of these assumptions characterize the Rasch model. All is required is to recognize that items are ordered and response categories within items are ordered. The data, in other words, are treated as ordinal; the first step in the Rasch transformation. In the *add em'up* belief system, all that is produced is an ordinal integer scale which cannot support claims for therapy response as it lacks linear and interval properties; rather than apply standard parametric techniques we are forced to rely on non-parametric statistics (e.g., median and modal values rather than means).

In application of the Rasch model each item has its own difficulty estimate although sharing the same threshold structure. The thresholds divide the latent trait continuum into intervals and their positions along the continuum, reflecting how participants respond to the item The estimate of item difficulty and the ranking of items by this one statistic is represented by a set of parameters, one for each category of the item (typically denoted as delta logit values) which are estimated from observed individual responses to the item categories. Maximum likelihood is used to estimate the parameters of the Rasch model including the delta values that best explain the observed responses across individuals. The delta values indicate the difficulty of endorsing each category of the item with higher values indicating more difficult categories. The sequence of delta values provides an indication of the item's position on the latent trait continuum and its overall level of difficulty. The positions of these thresholds in the Rasch Rating Scale Model are determined relative to the item's difficulty The intervals between these thresholds correspond to regions where participants are likely to indicate a choice. The threshold position for each item difficulty is a key metric as it can indicate redundant items with the same difficulty level or the absence of items to provide a ranking of item difficulties that best map into the distribution of abilities on the Rasch logit real number line. In the last resort the item difficulty estimates are describe as the balance point at which the highest and lowest categories are equally probable so that the threshold locations are relative to each item's difficulty estimate. The result, without going into the fine details of the Rasch modeling, is an indication of the relative difficulty of an item and a threshold structure that is common to all items. This is the starting point for our hypothetical example to illustrate how we might assess response to therapy from the Rasch model results.

RASCH RESPONSE TO THERAPY

If a Rasch standard questionnaire is to be developed there is a critical point to be observed in determining the response categories. The first response option for each item should refer to the absence of the latent trait which is equivalent to the "not true" response in the dichotomous model. In a four-response category item the we might have "not true", "sometimes true", "often true" and "always true''. We cannot assume that irrespective of ability or item difficulty (e.g., "sometimes rarely true") there will always be sone degree of possession of the manifested latent trait measure. If we dropped this assumption then we face practical difficulties with four responses and three thresholds where, given the intention is to apply threshold logit values as weights, then we cannot proceed. The first response, therefore, is a null response with no elements present of the latent trait. This leaves us with the three logit thresholds for the four response categories with zero impact on possession of the first response.

The hypothetical example utilizes 10 respondents with increasing ability and 5 items with increasing difficulty (Table 1). There are three threshold values in logits and their probability values (-1.609 or p=0.166; -0.255 or p = 437; and 1.099 or p = 0.750). As this is a Rasch Rating Scale Model the three logits/probabilities repeat for each item response level to give a possible 15 responses. Null responses are indicated for items where there is no response to any of the 3 remaining items. Each respondent records a maximum of five responses. The probabilities for each threshold are the weights that attach to that response indication the degree of manifested latent trait possession for that response. The more favorable responses indicating a greater contribution to the individual possession of the latent trait.

The maximum value for the manifested latent trait is the sum of probability weighs where each respond scored is for the highest threshold (i.e., p = 0.750 x 5 = 3.75). It is then possible to estimate the proportion of the manifest latent trait possessed by each respondent. These range from 0.88 (respondent 1) to 0.916 (respondent 10) in the baseline distribution of responses. As these possession proportions, range 0 to 1, are on a linear and interval scale, we can apply standard techniques to estimate the mean possession (0.502) and standard deviation (0.287) [assuming an approximation to a normal distribution].

The next step is to consider a hypothesis that a therapy intervention will significantly increase the manifested latent trait possession (Table 2). The impact of the intervention is to shift the distribution of responses towards a greater ability to respond to the more difficult items and response levels within each item. Assuming that the threshold logits remain unchanged, the result is a range of possession proportions of 0.332 to 1.0 (complete possession) with a mean value of 0.796 and standard deviation of 0.257. Comparing the differences in mean values (0.502 vs. 0.796) it supports the hypothesis of a significant possession difference (0.294; p = 0.0267) and a significant effect size (Cohen's d = 1.079).

Note the caveat here is that as the distribution of responses changes the threshold logits and probabilities will change. The options are to stay with the baseline logit thresholds and probability weights for each item response, compare therapy response where the prior and post logit thresholds

TABLE 1

EVALUATING RASCH LATENT TRAIT POSSESSION: POLYTOMOUS PRIOR RESPONSE DISTRIBUTION

Items Increasing Difficulty	Item Logit	Item Probability Weight	Respondents (1 – 10) Respondent Ability increasing ….									
			1	2	3	4	5	6	7	8	9	10
1.1	-1.609	0.166			1	1		1				
1.2	-0.255	0.437		2			2		2	2	2	2
1.3	1.099	0.750										
2.1	-1.609	0.166	1	1		1						
2.2	-0.255	0.437			2		2	2				
2.3	1.099	0.750							3	3	3	3
3.1	-1.609	0.166	1	1		1	1					
3.2	-0.255	0.437						2	2	2		
3.3	1.099	0.750									3	3
4.1	-1.609	0.166		1	1	1	1					
4.2	-0.255	0.437						2	2	2	2	
4.3	1.099	0.750										3
5.1	-1.609	0.166										
5.2	-0.255	0.437				2	2	2				
5.3	1.099	0.750							3	3	3	3
Null items			3	1	2	0	0	0	0	0	0	0
Sum Item Weights		6.765	0.332	0.935	0.769	1.870	1.643	1.914	2.811	2.811	3.124	3.437
Latent Trait Possession	Total possession = 3.75 Mean possession = 0.502 SD = 0.287		0.088	0.249	0.205	0.526	0.438	0.510	0.750	0.750	0.833	0.916

TABLE 2

EVALUATING RASCH LATENT TRAIT POSSESSION: POLYTOMOUS POST RESPONSE DISTRIBUTION

Items Increasing Difficulty	Item Logit	Item Probability Weight	Respondents (1 – 10) Respondent Ability increasing ….									
			1	2	3	4	5	6	7	8	9	10
1.1	-1.609	0.166										
1.2	-0.255	0.437	2	2	2	2	2	2	2			
1.3	1.099	0.750								3	3	3
2.1	-1.609	0.166										
2.2	-0.255	0.437	2	2	2	2	2	2				
2.3	1.099	0.750							3	3	3	3
3.1	-1.609	0.166	1				1					
3.2	-0.255	0.437		2		2		2				
3.3	1.099	0.750			3				3	3	3	3
4.1	-1.609	0.166	1	1								
4.2	-0.255	0.437			2	2	2	2	2			
4.3	1.099	0.750								3	3	3
5.1	-1.609	0.166		1	1							
5.2	-0.255	0.437				2	2					
5.3	1.099	0.750						3	3	3	3	3
Null items			1	0	0	0	0	0	0	0	0	0
Sum Item Weights			1.206	1.643	2.261	2.185	1.914	2.500	3.124	3.750	3.750	3.750
Latent Trait Possession	Total possession = 3.75 Mean possession = 0.796 SD = 0.257		0.322	0.438	0.603	0.583	0.510	0.667	0.833	1.000	1.000	1.000

differ or take the average of the corresponding logit thresholds. In this case we have remained with the prior threshold values defined as probability weights.

CONCLUSIONS

This example may seem trivial, but it addresses a key question in health technology assessment: a meaningful assessment of therapy response where the observations are unidimensional with a credible single attribute having linear, interval and invariant properties. A standard rarely if ever achieved in HTA PRO claims. This is guaranteed in this case as we are dealing with transformations of a logit real number line or latent trait continuum to probabilities. The exercise underscores the central role of Rasch measurement in providing the unique necessary and sufficient means to transform ordinal to interval scales, illustrating its application to polytomous instruments. These are common in HTA, but few are evaluated through a Rasch transformation to a logit measurement scale. They have little if anything to contribute to claims for therapy response.

REFERENCES

[i] Wright B, Linacre J. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehab.* 1989; 70(12):857-60 https://www.researchgate.net/publication/20338407_Observations_are_always_ordinal_measurements_however_must_be_interval/link/5563b02408ae9963a11ef326/download

[ii] Andrich D, Marais I. A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences. Singapore: Springer, 2019

[iii] Bond T, Yan Z, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 4th Ed. New York: Routledge, 2021

[iv] Langley P. Integers, Linear Transformations, Logistic Transformations and Value Claims for Therapy Response. Maimon Working Papers No. 12 July 2023 file:///C:/Users/Paul/Downloads/Maimon-logistic-functions-V7%20(3).pdf