MAIMON WORKING PAPERS No. 11 JULY 2023

THERAPY RESPONSE, RASCH MODELLING AND LATENT CONSTRUCTS [REVISED]

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

ABSTRACT

If claims for patient reported outcomes (PRO) or subjective claims for therapy impact are to be taken seriously by formulary committees and other health system decision makers they have to meet the standards for fundamental measurement. These standards, which are the basis for claims in the physical sciences, must be for single attributes: unidimensional, linear, interval and invariant in respect of the application of the measurement instrument. Unfortunately, apart from the few practitioners who subscribe to Rasch or modern measurement theory, claims for patient or subjective responses in health technology assessment (HTA), the overwhelming majority of PRO instruments, including composite generic multiattribute measures and those that add integers from ordinal constructs, fail to meet these standards. As a result, claims for therapy response rest on uncertain foundations. The purpose of this brief commentary is to propose, given the standards of Rasch measurement, how we can report on therapy response in terms of latent constructs and the extent to which respondents can report on response to therapy in terms of the possession of the latent construct that is manifested in instrument choice. This fulfills the claim that Rasch measurement is the necessary and sufficient means to transform ordinal counts, which are the hallmark of HTA, to unidimensional, linear, interval and invariant measures that capture the interaction between respondent ability and item difficulty.

INTRODUCTION

Value claims for therapy response, whether they refer to direct clinical measures, patient reported outcome claims, drug utilization and other resource utilization claims must be based on measures consistent with the standards of Rasch or fundamental measurement; they must refer to single attributes with unidimensional linear, interval and invariant properties [1]. Consistent also with the standards of normal science all value claims must be empirically evaluable and replicable; assumption driven simulation modeled imaginary cost-effectiveness claims are unacceptable. These measurement standards are not new; they have been the central to the growth of provisional objective knowledge. Indeed, the history of science can be equally well be described not just as the systematic search for new provisional facts but the history of measurement. It is against this background that we must make the case for the imperative of Rasch measurement in health technology assessment in articulating PRO or subjective claims for therapy impact defined in terms of latent constructs where the Rasch framework supports a measure of the manifestation of an attribute of the latent construct [2].

The purpose of this brief commentary is to illustrate the process by which a Rasch logit continuum can be transformed by a simple logistic function to a series of probabilities that serve as item weights. This creates what is described as a frontier for the highest weighted item responded to before and after a therapy intervention. This yields value claims for therapy response which are defined in terms of improved construct possession, where the Rasch framework accommodates the interface between respondent ability and item difficulty.

THE RASCH MODEL

Initially developed in the 1950s, the Rasch model for developing interval level single attribute measurement, is widely accepted in education, less so in psychology and only to a limited extent in health technology assessment. Indeed, Rasch measurement hardly, if ever, features in leading HTA textbooks and practice guidelines by organizations such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR); although in respect of ISPOR there are a number of papers in *Value in Health* that have applied Rasch standards assessment to existing PRO instruments to assess their relevance, even with the culling of items, for integer response assessment. The most recent guidance for HTA, CHEERS 2022 details standards for modelling submissions to journals makes no mention of Rasch and the importance of fundamental measurement in formulary submissions; instead, it continues to promote the composite and mathematically impossible, quality adjusted life year (QALY) [3].

This lack of recognition is unfortunate because Rasch it is accepted by measurement theorists as the unique necessary and sufficient means to transform ordinal counts to interval measures [4]. This follows from a long-standing confusion between observations, as ordinal scores, and measurement which must be interval. All observations are ordinal, as counts or stated levels of performance. But these are not a measure; they are not a *calibrated measuring system with a well-defined origin and unit* [4] . While observations or counts are always our starting point, the Rasch contribution is to establish the mathematical rules for transformation to interval scores. Rasch's contribution to the debate over measurement and the work, in particular, of Thurstone in the 1920s and his emphasis on establishing relative differences, was to make clear (i) that regardless of context a measure must retain its quantitative status (e.g., each test item must retain its level of difficulty) and (ii) that any application of an instrument must recognize, in probabilistic terms, the interaction between the ability of a respondent and the difficulty of the item [5 6]. Hence, for a given level of ability, the probability of a respondent responding successfully to an item must be less the more difficult the item. This is the basis for the Rasch model transforming ordinal counts to interval measures

THE RASCH TRANSFORMATION

It is important to emphasize that the Rasch model, and any measures that derive from the model, must focus on respect the probability of successfully responding to an item as a function of the difference between respondent ability and item difficulty. In the simplest case of a dichotomous questionnaire with items yielding a likelihood of such captured by Yes=1 and No=0, the items represent, following the fit to the Rasch model, the measured manifestation of a latent construct. Consider the latent construct (entity or trait) we may call quality of life, but defined not in narrowly clinical terms as evidenced by the notion of health- related quality of life, but a more holistic concept of quality of life as the needs of patients or caregivers [7] . The more needs that are met, the greater the quality of life where these needs may accommodate the impact of health status. From the analyst's perspective, the issue is one of manifesting these needs in a form that allows a count of observations for item responses; there is no immediate jump to measurement [8] This is typically achieved by extensive interviews which generate statements that capture the needs considered by the respondent sample in a target, disease specific population, with a first cut of a questionnaire by the selection of items. At this stage there is no attempt to assign difficulty and ability to the sample of respondents just from interviews. Given provisional item selection, a simple two-way matrix will suffice to capture the number of items successfully responded to (item difficulty) and the ability of the respondents (number of successful item responses).will yield the required probabilities.

This is the basis for the Rasch transformation where, to capture relative differences, the probability of a successful response defined by ability and difficulty is transformed to an odds ratio and then to logits for ability and respondent rows and columns where the logit is defined as the natural logarithm of the odds ratio ((ln(p/1-p)). The difficulty logits and ability logits are iterated against each other by joint maximum likelihood estimation to a preset convergence value for the maximum difference in person and item values. The logits are applied to a continuum or scale calibrated in logits where the average logit is arbitrarily set to zero with positive logits indicating greater than average probabilities and negative logits lower than average probabilities. This scale or Rasch continuum not only represents relative differences between logit values but the scale is unidimensional, linear, interval and invariant (typically represented as a Wright map). Rasch is unique in satisfying this condition.

It is not the intention here to give a detailed explanation of the various Rasch models for both dichotomous and polytomous instrumentation. This is more than adequately explained in text books, on-line videos and in journal articles. Indeed, there an extensive suite of software programs that apply Rasch rules to determine whether an interval scale can be created. The two main software packages are RUMM2030 (https://www.rummlab.com.au/) and WINSTEPS (https://www.winsteps.com/winsteps.htm). The former is applied in the Andrich and Marais textbook [8] the latter in Bond et al [2]. For those newly arrived to Rasch measurement, the latter book provides a non-mathematics focused introduction (with mathematics in Appendices. Reference should also be made to a review, including the Wright map, by Boone [9]. Rasch measurement is not easy to those with only 'classical' HTA training. To illustrate how the iteration process in Rasch meets a preset convergence value, using the WINSTEPS model, reference should be made to a useful online spreadsheet by Moulton (rasch.org/moulton.htm) [10]. This provides a detailed step by step explanation of convergence to a joint logistic score in the iteration process.

THE LATENT TRAIT CONTINUUM: RESPONSE TO THERAPY

Given our objective to create a measurable manifestation of the latent trait fulfillment, the real number line defined in transformed logits is the required measure, within the constraints of the number of items and the logits indicating the measure attributable to successfully responding to or affirming a statemen (true = 1; not true = 0) an item. The important point is that this transformation, which fits item to the Rasch model, maps those responses to a real number line applied in common to items and persons defined in logits. This means that, from the perspective of the items, the Rasch iterated transformed logit values for the items map to a unidimensional, linear, interval and invariant scale, invariance is defined as a logit value for each item in the instrument. The logit values take both negative and positive values, typically in the range +/- 3.5 logits, with the logits approaching +/- infinity asymptotically and the probabilities approximately 1 or zero for logits greater than this range. It is the only transformation from subjective ordinal responses to a scale with unidimensional, linear, interval and invariant properties defined in terms of logits that supports meaningful claims for therapy responses; the extent to which statements are affirmed.

To achieve this transformation to proportions or probabilities, which we might label the latent construct possession proportion, is a relatively trivial step. All that is required is to apply the logistic function where p is probability as $p = 1/1 + e^{-logit})$ to estimate the probability that supports the odds ratio from the Rasch logit. These are probabilities (or item weights) for the respective logits. These standards for the original Rasch logistic continuum are retained in the logistic transformation and the percentages that act as weights for each item in a questionnaire; if two logits are equally spaced apart then their corresponding probabilities will be equally spaced apart; the spacing reflecting the interaction between respondent ability and item difficulty to yield a likelihood estimate of successful responses to items. This is only achieved by Rasch measurement. As noted in the introduction to this module, these probabilities are the gold standard for assessing therapy response for disease specific manifestations of a latent construct,

.

Adopting the Rasch model for the latent trait quality of life, enables value claims to be made in terms of latent constructs; in needs fulfillment as a manifestation of quality of life, the extent to which needs are fulfilled are, given the fit of the items to the Rasch model, precisely articulated and presented as a value claim with application of an instrument with invariant properties where the probabilities capture the contribution of a successful outcome or the weighted addition of the affirmed item to the standardized latent trait score. This provides a basis for therapy response value claims which meet the required Rasch measurement standards.

Tables 3 and 4, for a questionnaire with 10 items ranked by difficulty and 10 respondents ranked by ability, provide an example of the distribution of responses to items prior to and post therapy. As the items are ranked by their difficulty, affirmation of a more difficult item assumes that the previous less difficult items are likely to have been affirmed. This may not always be the case but these missing values can be accommodated in the latent trait estimation; note that the respondents have no idea of the relative difficulty weights attached to each item. This has been determined at early stages in the Rasch modelling.

The maximum latent trait will equal the total weights for the items multiplied by the number of respondents. In this case the total item weights are 5,343 which gives maximum for the latent trait on a scale 0 to 0.534 (i.e., 5,343/100). Any combination of item responses will fall into this range. To make a claim for the proportion of the latent trait possessed or affirmed by respondents the response has to be transformed to this range. This means in Table 3 where the prior distribution of average item response affirmation is 0.321 the actual proportion is 0.321/0.534 = 0.601 or 60.01% for the degree of possession or affirmation of the latent trait.

The corresponding estimate for Table 4 with the post-therapy distribution of responses scores (added to the item distribution in Table 3) is 0.418 for possession or affirmation. Standardizing is 0.418/0.534 = 0.783 or 78.28%. The hypothesis is that the impact of the therapy interventions allows more respondents to realize the item statements and increase latent trait possession. The standardized proportions are still on a unidimensional, linear, interval and invariant scale; they retain the required measurement properties. Standardizing to the maximum possible latent trait value allows comparisons to be made, not just for prior and post therapy interventions but between questionnaires which differ in their item number, logit weights and number of respondents. In this case the impact of the new therapy, given an invariant questionnaire, is to increase latent trait affirmation or possession from 60.01% to 78.28%. The hypothesis in respect of the positive impact of a new therapy intervention is realized ($p = 0.003$) and effect size (Cohen's d) is 0.894. Provisionally, we can assume that the impact of a new therapy intervention on the latent trait possession is highly significant.

Presenting the increased affirmation or possession the latent trait manifested in the items that meet Rasch standards, the fit of the items, gives a coherent and robust bases for value claims presented as single attributes which are unidimensional, linear interval and invariant. At the same time, it is important to consider whether or not simple integer summation, assumed to be from the same questionnaire presented here which meets Rasch fit standards, have any intrinsic meaning or whether integer summation should be rejected.

Integer summation for a 10 item by 10 respondents will yield a proportion or percentage out of 100 possible integers assigned (or affirmed) to item responses where the items are ranked by difficulty and respondents by ability. What is missing is any attempt to build on this Rasch probabilistic framework where the probability of a successful response is a function of the difference between item difficulty and respondent ability; that is, to apply a logit scale transformed to probability weights for item difficulty. In the prior distribution of responses (Table 3) the total integer affirmatory response is 38/100 with the post therapy integer count 60/100 (Table 4); these percentages are far removed from the corresponding 60.01% and 78.28% of the weighted and standardized affirmation model. The failure of the integer count is that while the item fit follows Rasch standards no account is taken of the Rasch standard for the probability of a successful affirmation of an item as reflecting the manifestation of a latent trait. Instead, the integer count

TABLE 3

EVALUATING RASCH LATENT TRAIT POSSESSION: 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	-2.484	0.078	1		1	1		1	1	1	1	1
2	-1.437	0.192	2	2	2	2	2		2	2		2
3	-0.636	0.346	3	3	3		3	3	3	3		3
4	-0.156	0.461			4	4	4	4	4	4	4	4
5	0.0	0.500					5	5	5	5	5	5
6	0.310	0.577									6	6
7	0.805	0.690										
8	1.203	0.769										
9	1.704	0.846										
10	2.041	0.884										
Sum Item Weights			0.616	0.538	1.077	0.731	1.499	1.385	1.577	1.577	1.616	2.154
Average Item Weight			0.205	0.269	0.269	0.237	0.375	0.346	0.315	0.315	0.404	0.359
Sum average item weights 3.206												
Overall average item weights 0.321												
Latent trait possession 0.321/0.534 = 0.601 or 60.11%												

TABLE 4

EVALUATING RASCH LATENT TRAIT POSSESSION: 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	-2.484	0.078	1		1	1		1	1	1	1	1
2	-1.437	0.192	2	2	2	2	2		2	2		2
3	-0.636	0.346	3	3	3		3	3	3	3		3
4	-0.156	0.461	4	4	4	4	4	4	4	4	4	4
5	0.0	0.500		5	5	5	5	5	5	5	5	5
6	0.310	0.577				6	6	6	6	6	6	6
7	0.805	0.690							7	7	7	7
8	1.203	0.769								8	8	8
9	1.704	0.846									9	9
10	2.041	0.884										10
Sum Item Weights			1.077	1.499	1.577	1.808	2.076	1.962	2.844	3.613	3.921	5.343
Average of Item Weights			0.269	0.375	0.315	0.362	0.415	0.392	0.406	0.452	0.560	0.343
Sum average item weights 4.08												
Overall average item weights 0.418												
Latent trait possession 0.418/0.534 = 0.783 or 78.28%												

assumes that item affirmation with the application of probability weights is irrelevant. There is no concept of the interaction between item difficulty and respondent ability.

To this dismissal of integer responses where the instrument item fit meets Rasch standards, should be added the misuse of integers where the instrument is just a selection of responses that have been cobbled together with no concept of the Rasch probabilistic approach to transforming integer counts to a unidimensional, linear, interval and invariant measures. These must be rejected out of hand. All that is achieved (or failed to achieve) is to assume each item or statement is equally difficult (the same weight) and that, presumably, each respondent has the same ability to realize an item statement. In which case a single item would suffice. All we have is a subjective or ordinal count which is only the first step in the Rasch necessary and sufficient transformation to an interval measure.

CONCLUSIONS

The Rasch analysis presented here and the transformation of the Rasch logits on a unidimensional, linear, interval and invariant scale, which is maintained with the transformation to probabilities, presents a new focus on value claims for products and devices. Not unreasonably, this can be described as the Rasch gold standard for therapy response. In this formulation we have estimates with the required measurement properties which capture the extent to which item successful responses drive estimates of the extent to which we can claim the extent which a latent construct, manifested in a questionnaire that meets Rasch model standards, is possessed by a target patient population. This, it should be noted, raises an issue for Rasch standards: the criteria for a latent trait that manifests a range from 0 to 1 to assess therapy response.

The analysis is consistent with standards that should be met in HTA value claims: the claims should disease or target patient population specific with a single attribute measure that is invariant across applications. In this case the item probabilities are fixed in all applications where all that is required are item responses to calibrate measure of latent construct possession and response to therapy in these terms. The value claims are empirically evaluable and replicable as well as meeting Rasch measurement standards. This framework for value claims puts to one side the common approach of aggregating scores, whether for dichotomous or polytomous instruments. The Rasch framework focuses on the interactions between items and ability, with linear and interval measure of the logit hierarchy of item difficulty. This is overlooked entirely in the majority of PRO instruments in HTA. The framework presented here is for a new start in value claims and one that is consistent with the standards of normal science and fundamental measurement.

REFERENCES

[1] Langley P. Nothing to Cheer About: Endorsing Imaginary Economic Evaluations and Value Claims with CHEERS 22 [version 1; peer review: 2 approved] F1000Research 2022, 11:248 https://doi.org/10.12688/f1000research.109389.1

[2] Bond T, Yan Z, Heene M. Applying the Rasch Model: Fundamental measurement in the human sciences. 4th Ed. New York: Routledge, 2021

[3] Husereau D, Drummond M, Augustovski F et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 22) Statement: Updated reporting guidance for health economic evaluations. *ValueHealth*. 2022;25(1):3-9 https://www.valueinhealthjournal.com/article/S1098-3015(21)03146-6/pdf

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

[5] Thurstone L. A method of scaling psychological and educational data. *J Educ Psychol.* 1925;15:433-51

[6] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske, Institut, 1960

[7] McKenna S, Doward L. The needs-based approach to quality of life assessment. *ValueHealth*. 2004; 7(5): Supp 1. S1-S40 https://www.valueinhealthjournal.com/article/S1098-3015(10)60227-6/fulltext

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

[9] Boone W, Rasch Analysis for Instrument Development: When, Why and How? CBE – Life Sciences Education. 2016; 15:rm4 1-7 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5132390/pdf/rm4.pdf

[10] Moulton M. Rasch Estimation Demonstration Spreadsheet. 2003 rasch.org/moulton.htm