An alternative to Rasch analysis using triadic comparisons and multi-dimensional scaling

C Bradley1, R W Massof2
1Research Associate, Wilmer Eye Institute, Johns Hopkins University, Baltimore, MD, USA
2Professor of Ophthalmology, Johns Hopkins University, Baltimore, MD, USA

E-mail: cbradley05@gmail.com

Abstract. Rasch analysis is a principled approach for estimating the magnitude of some shared property of a set of items when a group of people assign ordinal ratings to them. In the general case, Rasch analysis not only estimates person and item measures on the same invariant scale, but also estimates the average thresholds used by the population to define rating categories. However, Rasch analysis fails when there is insufficient variance in the observed responses because it assumes a probabilistic relationship between person measures, item measures and the rating assigned by a person to an item. When only a single person is rating all items, there may be cases where the person assigns the same rating to many items no matter how many times he rates them. We introduce an alternative to Rasch analysis for precisely these situations. Our approach leverages multi-dimensional scaling (MDS) and requires only rank orderings of items and rank orderings of pairs of distances between items to work. Simulations show one variant of this approach - triadic comparisons with non-metric MDS - provides highly accurate estimates of item measures in realistic situations.

1. Introduction

Rasch analysis is often used to estimate the magnitude of some common, underlying property shared by a set of items1,2,3. The inputs to Rasch analysis are ordinal ratings assigned to the items by a group of people. The outputs are estimated measures of both the items and the people. These estimated item and person measures are placed on the same invariant scale, allowing differences between item and person measures at different points on the axis to be directly comparable because they share the same units. A common application is to take ordinal ratings assigned to items on a questionnaire and estimate the relative difficulties of the items (e.g. if the items are tasks) and the relative abilities of the people that rated the items.

In order for Rasch analysis to provide good estimates of person and item measures, it is necessary for there to be sufficient variation in the observed responses. The reason for this is because Rasch analysis assumes that there is always a probability associated with any person assigning a given rating to any item, even if the true person and item measures are known. This variability in responses is quantified by probability distributions centered on both the person measures and item measures, as well as on any thresholds the person uses to define the boundaries of the rating categories. These assumptions are realistic because neither a person's ability, nor his perception of...
item difficulty, nor the precise thresholds used to categorize items remain constant from trial to trial. These probability distributions - also interpretable as error distributions around some mean value - allow Rasch analysis to map observed frequencies in assigned ratings to distances between item measures on an axis.

For most applications the required variance for Rasch analysis to be applicable is supplied by the differing responses people have to the same item. That is, the variance comes from a population. We run into trouble if the goal is to estimate item measures based on responses from just a single person. When only one person is providing the responses, it is possible the person will assign the same rating to an item no matter how many times he rates that item. For example, if there are only two possible rating categories 0 and 1 and the person is asked to rate items as either "difficult" or "easy", we might obtain the same answer each time for items such as "create a website" or "swim 1 km". Even when there are multiple rating categories we may run into this lack of sufficient variance problem (e.g. rating a movie as "great", "fairly good", "not that good", or "terrible").

2. Multi-dimensional scaling
The purpose of this paper is to outline an alternative to Rasch analysis in cases where there is insufficient variance in the observed responses, specifically in the case where only a single person is rating all the items and for whatever reason the person's responses show little to no variance over time. Our approach leverages the well-known mathematical technique of multidimensional scaling (MDS). MDS estimates coordinates of points in a low dimensional space (in our case a 1D space) from a dissimilarity matrix whose entries represent some dissimilarity measure between pairs of points (items). Depending on the types of responses we choose to obtain from the person in question, we will either use classical MDS or non-metric MDS. Classical MDS treats the entries in the dissimilarity matrix as actual distances and attempts to minimize a loss function of those distances. Non-metric MDS assumes much less about the nature of the input data, essentially interpreting the entries in the dissimilarity matrix as rank orderings of the distances between pairs of items. Which method we use will depend on the nature of the input data, which in turn depends on how difficult it may be for a person with limited cognitive abilities to provide internally consistent responses. Importantly, MDS is in general not a substitute for Rasch analysis because it assumes all inputs are in the same units whereas Rasch analysis allows each person to assign ratings in his own units.

Suppose our goal is to estimate the relative utilities of different events or outcomes for a single person. Since we are interested in estimating utilities for just one person, information about how a population assigns utilities to the items is in principle irrelevant. Thus, our only source of relevant information is the person himself. Our task is certainly simplified if the person solves the problem for us. That is, if a person can estimate his subjective utilities for different items in an internally consistent way, we are done. We just need a method for testing whether his responses are internally consistent. One way to do this is to ask the person to not only directly estimate item utilities, but to also estimate the distances in utility units between pairs of items. Since these distances are measures of dissimilarities between item pairs, they serve as inputs to classical MDS, which can be used to estimate item measures on a 1D axis. If the estimated item measures are highly similar to the person-estimated item utilities, we are done. The problem is that these two sets of estimated item measures may very well differ from each other, necessitating a different approach.

3. Paired and triadic comparisons
If person-estimated item measures and classical MDS estimated item measures differ too much from each other, it is almost certainly because the task of providing internally consistent estimates
of item measures or distances between them is too challenging from a cognitive point of view. Humans are limited in their cognitive abilities and it is important that we require the person to perform a task that is easy enough so that his responses have good enough internal consistency. In general, it is easier for humans to simply rank order a set of items than to assign real numbers to them that accurately represent their relative magnitudes. This is true even if we are talking about general, it is easier for humans to simply rank order a set of items than to assign real numbers to perform a task that is easy enough so that his responses have good enough internal consistency. In

Humans are limited in their cognitive abilities and it is important that we require the person to of item measures or distances between them is too challenging from a cognitive point of view. Humans are limited in their cognitive abilities and it is important that we require the person to perform a task that is easy enough so that his responses have good enough internal consistency. In general, it is easier for humans to simply rank order a set of items than to assign real numbers to them that accurately represent their relative magnitudes. This is true even if we are talking about rank-ordering distances between items. If \( d(A, B) \) and \( d(C, D) \) represent the distances between items \( A \) and \( B \), and items \( C \) and \( D \), respectively, then it is much easier for a person to say whether \( d(A, B) > d(C, D) \) or \( d(A, B) < d(C, D) \) compared to numerically estimating \( d(A, B) \) and \( d(C, D) \). Thus, our goal should be to estimate item measures using only rank orderings, either of items or of distances between items.

In the simplest case, we could ask the person to rank order pairs of distances between items. On each trial, the larger distance could be assigned a rank of 2 while the smaller distance could be assigned a 1. Such a "paired comparison" approach, if done on sufficiently many pairs, will over time provide us with accurate estimates of the average rank value for the distance between any two item measures. The average ranks for all item pairs is precisely the type of input non-metric MDS can be applied to, giving us a way of estimating item measures using only rank information.

If it turns out that rank ordering \( d(A, B) \) and \( d(B, C) \), where there is one common item between the pairs, is easier than rank ordering \( d(A, B) \) and \( d(C, D) \), a third approach can be taken. First, we ask the person to rank order a triplet of items, say items \( A \), \( B \), and \( C \). If the person responds with \( B > A > C \), then we automatically know that \( d(B, C) \) is the largest distance between all possible item pairs. The second question is which of \( d(B, A) \) and \( d(A, C) \) is larger. Assign rank values of 3, 2 and 1 for the largest to smallest distances in this "triadic comparisons" approach and we will over many trials obtain estimates of the average rank value for the distance between any two item measures. Non-metric MDS can once again be applied to these average ranks to estimate item measures.

One key difference between the paired and triadic comparisons approaches is that paired comparisons can in principle compare every item pair to every other, ensuring that accurate estimates of average rank values are obtained, while triadic comparisons cannot because every comparison involves a common item. Thus, with triadic comparisons, we need evidence through simulations that estimated item measures are linearly related to true item measures. Because the number of unique triplets for \( N \) items is \( N(N - 1)(N - 2)/6 \), it makes little sense to apply this approach if \( N \) is large. One potential real world application where \( N \) is small involves estimating utilities of different possible treatment outcomes for individual patients. For example, 3200 patients were administered the Activity Inventory (AI), a questionnaire that asks patients to rate the importance and difficulties of 50 goals in daily life, from driving to recognizing people's faces to preparing meals without assistance. They rated the importance of each goal on an integer scale from 0 to 3 and the difficulty of each goal from 0 to 4. While it is not clear precisely how importance and difficulty ratings map to utilities, one hypothesis is that utility is the product of the two. Because there is no need for an intervention if either the importance or difficulty of a goal or task is zero, we can try to estimate the utilities of \( N = 12 \) items, the possible non-zero products of importance and difficulty ratings. With 12 items, there are 220 unique item triplets, leading to 220 triadic comparisons, which is within manageable size for a patient questionnaire.

We simulated the triadic comparisons approach by first bootstrapping utility values from the distribution of importance times difficulty ratings obtained from the 3200 patients. Noise was added on each of 1000 simulations. For each simulation, all 220 possible triadic comparisons were performed, with random noise added to item distances to prevent rank orderings from being
too deterministic. The average ranks for each item pair were computed and non-metric MDS applied to the resulting dissimilarity matrix to estimate item measures. As long as the standard deviation of the noise added to the distances was less than 1 (utility values ranged from 1 to 12 before noise was added), the estimated item measures had a highly linear relationship with the true item measures, with $r^2$ averaging around 0.97 over 1000 simulations. Separate simulations with 12 randomly chosen points from a uniform or normal distribution also yielded $r^2$ values averaging around 0.97 or higher. Skewing a uniform distribution by squaring the values led to an $r^2$ of around 0.95, while greater amounts of skewing or larger amounts of added noise progressively reduced $r^2$ (e.g. increasing the standard deviation on the distances to half the range at 6 or skewing the values by a power of 4 both lowered $r^2$ to the 0.85 range). These simulations show that the method of triadic comparisons with non-metric MDS provides good estimates of item measures for a single person when insufficient variance exists for Rasch analysis to be applicable.

References

[1] Rasch G 1960 Probabilistic models for some intelligence and attainment tests (Copenhagen: The Danish Institute of Educational Research)
[2] Andrich D 1978 A rating formulation for ordered response categories Psychometrika 43 561-573
[3] Masters G N 1982 A Rasch model for partial credit scoring Psychometrika 47 149-174
[4] Torgerson W S 1952 Multidimensional scaling, 1: theory and method Psychometrika 17 401-419
[5] Kruskal J B 1964 Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis Psychometrics 29 1-27
[6] Kruskal J B 1964 Nonmetric multidimensional scaling: a numerical method Psychometrics 29 115-129
[7] Borg I and Groenen P Kruskal J B 2005 Modern multidimensional scaling: Theory and applications 2nd ed (New York: Springer Verlag)
[8] Massof R W, Ahmadian L, Grover L L, Deremik J T, Goldstein J E, Rainey C, Epstein C, Barnett G D 2007 The Activity Inventory (AI): An adaptive visual function questionnaire Optom Vision Sci 84 763-774