Research and Development Division
Professional DynaMetric Programs, Inc.

ProScan Psychometric Assessment and Evolution: A Modern Approach Using Item Response Theory (IRT)

Prepared By:
Lauren Matheny, PhD, MPH.
September 2024

Contents

Background Information 5

The Rasch Rating Scale Model: A Special Case of Item Response Theory 5

Study Purpose 5
Methods 5
Table 1. Sample Size (n) for each Round of Data Collection 6

Results 6

Table 2. Reliability Estimates for Each Trait 7
Table 3. Validity Estimates for Each Trait 7

Conclusions 7

References 8

Figures 9

Figure 1A – 1E: Category response curves for the five latent constructs (ProScan traits) . . . 9
Figure 1A. Dominance 9
Figure 1B. Extroversion 10
Figure 1C. Pace 10
Figure 1D. Conformity. 11
Figure 1E. Logic. 11

Background Information

Assessing evidence of reliability and validity is a crucial step in the development and evolution of instruments used to measure affective traits. Reliability is the extent to which random sources of measurement error are minimized and refers to how dependably or consistently a test measures a characteristic (Henson, 2001). Validity is defined as the extent to which an instrument measures what it is intended to measure (Messick, 1989). Assessing reliability and validity over time, rather than at a single time point, is essential to evolve the instrument as society and language also continues to progress. This approach of continual and rigorous assessment facilitates the advancement of an instrument that is crucial to remain relevant in the modern world.

The Rasch Rating Scale Model: A Special Case of Item Response Theory

Classical test theory (CTT) is often used to psychometrically assess instrument scores and is based on the principle that the observed score is partitioned into the true score plus error. A more modern alternative to CTT that has been found useful in the development of instruments that measure traits is the Rasch rating scale model, which is a specialized form of item response theory (IRT; Andrich, 1978b; Rasch, 1980; Tennant et al., 2004; Wright & Masters, 1982). Rasch analysis is a probabilistic mathematical technique that is used to assess psychometric properties of latent constructs. Item response theory is a more modern approach to psychometric assessment than classical test theory and is a set of mathematical models that describe the relationship between an individual's ability (or trait) and how that individual responds to items on the scale. The Rasch model has two main principles that are 1) the easier the item on the scale, the more likely that the item will be chosen or "passed," and 2) the higher the ability or trait level a person possesses, the more probable the person will "pass" or endorse an item compared to a person with a lower ability or trait level (Tennant et al., 2004). The scale of interest is measured in terms of item difficulty and generates estimates of locations of individual items (item difficulty) and ability level along a common interval-level scale (log-odds). While similar to a Classical Test Theory (CTT) approach in that both CTT and IRT methods assess psychometric properties including reliability and validity, IRT offers several advantages over classical assessment methods. Item Response Theory (IRT) offers a more nuanced analysis of survey items by assessing how each item performs across different levels of ability, allowing for more accurate measurement of individual abilities. Unlike Classical Test Theory (CTT), IRT accounts for varying item difficulties and respondents' differing levels of the underlying trait being measured, providing a more detailed understanding of how items differentiate between survey-takers. Additionally, IRT's models are more generalizable and less dependent on specific sample characteristics, making it easier to apply results across diverse populations.

Study Purpose

The purpose of this study was to confirm that the newly proposed model (ProScan traits with corresponding new words) had superior psychometric properties to that of the previous ProScan model for each trait, including Dominance, Extroversion, Pace, Conformity and Logic.

Methods

There were 10,349 individuals included in this analysis. The assumption of unidimensionality, which is required to conduct a Rasch analysis, was tested using principal component analysis (PCA) and scree plots to assess factor structure for each trait. Initial psychometric assessment, including reliability

and validity, of each ProScan Survey trait, was conducted using the Rasch model. Four rounds of data collection were conducted to test and assess new items for each trait and confirm using a new dataset each time (Table 1).

Table 1. Sample Size (n) for each Round of Data Collection	
Round 1	2,091
Round 2	2,255
Round 3	3,240
Round 4	2,763

Sample Size

The sample size was N = 10,349, which exceeded the recommended minimum sample size of 300 participants (Clark & Watson, 1995). Although 100 participants is the minimum sample required for Rasch analysis, it was important to capture more individuals for generalizability purposes (Green & Frantom, 2002).

Reliability

Evidence of reliability of responses to the scale was assessed using person reliability, item reliability and item separation. Person reliability is analogous to Cronbach's alpha (internal consistency), with values >.70 deemed acceptable. Item reliability is a measure of item difficulty consistency, meaning that items are consistently measuring lower or higher activity levels, with values >.90 deemed acceptable. Item separation indicates how efficiently the scale items can categorize patient activity levels, with values >3.0 deemed acceptable (Linacre, 2020).

Validity

Evidence of validity was assessed using mean square infit and outfit values that are indicative of acceptable or poorly fitting items. A value for mean-square infit and outfit is assigned for each item within each trait. Items were then evaluated for validity by assessing outfit mean-square (MNSQ) and infit MNSQ statistics, with acceptable values between 1.5 and .5. Values closest to 1.0 indicate the greatest evidence of validity, with the least amount of "noise" or error in the data (Linacre, 2002). Wright item person maps and category response curves were assessed to identify poorly fitting items and assess person ability and the ability of the instruments' items to capture differing levels of the trait being measured. Category response curves are plots used in IRT that show the probability of selecting each response category as a function of the underlying latent trait (θ). Each curve in the plot represents a different response category, with the x-axis showing the latent trait and the y-axis showing the probability, illustrating how the likelihood of each category varies across different levels of the trait. To perform this portion of the analysis, WINSTEPS version 4.0.1 (Beaverton, Oregon) was used.

Results

Reliability

The ProScan Survey demonstrated excellent evidence of reliability, Reliability, including person reliability, item reliability and item separation, for each trait for the final ProScan Survey were documented (Table 2).

Table 2. Reliability Estimates for Each Trait			
Trait	**Item Reliability**	**Person Reliability**	**Item Separation**
Dominance	1.00	.77	18.06
Extroversion	1.00	.74	14.77
Pace	1.00	.72	22.83
Conformity	.99	.79	10.31
Logic	1.00	.77	17.97

Validity

The ProScan Survey demonstrated excellent evidence of validity. All mean-square infit and outfit values were within the acceptable threshold of .5 to 1.5 (Table 3).

Table 3. Validity Estimates for Each Trait		
Trait	**Mean-square Infit range**	**Mean-square Outfit Range**
Dominance	.83 – 1.23	.84 – 1.29
Extroversion	.74 – 1.40	.70 – 1.38
Pace	.83 – 1.08	.83 – 1.08
Conformity	.80 – 1.18	.81 – 1.20
Logic	.74 – 1.11	.85 – 1.12

The results indicate that the category response curves for the items across the five latent constructs (ProScan traits) exhibit well-defined thresholds, with each response category having a distinct peak at different levels of the latent trait. This suggests that the items are effectively differentiating between various levels of the underlying constructs, demonstrating good measurement properties (**Figures 1A – 1E**).

Conclusions

The finding of this study, utilizing the Rasch rating scale model, which is a specialized form of item response theory, along with the previous research employing big data and Classical Test Theory (CTT) methods for confirmatory factor analysis, highlight the ProScan's exceptional reliability and validity, especially with the inclusion of the newly updated items. The robust nature of these results is evidenced by their successful replication across large sample sizes which is necessary for big data analytics, further reinforcing the psychometric strength of the instrument. The revisions made to the ProScan have yielded an enhanced version of the tool, which consistently delivers dependable and accurate scores. Crucially, the reproducibility of these scores has been demonstrated, adding a layer of confidence to the updated ProScan. This study has validated the proposed model structure, thereby affirming the credibility of the ProScan for both its application and score interpretation. Instrument development is inherently a cyclical process, demanding multiple iterations of data collection, statistical analysis, and refinement. However, this critical process is often disregarded due to the significant investment of time and resources it requires. By investing valuable resources and technology into the continual assessment and refinement of the ProScan survey, the tool remains a reliable and up-to-date modern instrument, allowing users to confidently rely on its accuracy and effectiveness.

References

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2(4), 581-594.

Andrich, D. (1978b). A rating formulation for ordered response categories. Psychometrika, 43(4), 561-573. https://doi.org/10.1007/BF02293814

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? Medical Care, I7-I16.

Clark, L. A., & Watson, D. (1995). Constructing Validity: Basic Issues in Objective Scale Development. Psychological Assessment, 7(3), 309-319. https://doi.org/10.1037/1040-3590.7.3.309

Green, K. E., & Frantom, C. G. (2002). Survey development and validation with the Rasch model. International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC, United States.

Linacre, J. M. (2000). Comparing "partial credit" and "rating scale" models. Rasch Measurement Transactions, 14(3), 768.

Linacre, J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. Rasch Measurement Transactions, 19(3), 1032.

Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? Journal of Outcome Measurement, 2, 266-283.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions, 16, 878.

Linacre, J. M. (2020). Reliability and separation of measures. Winsteps. https://www.winsteps.com/winman/reliability.htm

Linacre, J. M., & Wright, B. D. (2000). Winsteps. http://www. winsteps.com/index.htm

Figures

Figure 1A – 1E: Category response curves for the five latent constructs (ProScan traits). These figures present category response curves for one item from each of the five latent constructs. The x-axis represents the latent trait (θ) specific to each construct, and the y-axis shows the probability of selecting each response category. Each curve corresponds to a different response category on a Likert scale (1 to 5), illustrating how the probability of endorsing each category varies across levels of the latent trait. These CRCs demonstrate the item's effectiveness in distinguishing between different levels of the underlying constructs.

Figure 1A. Dominance

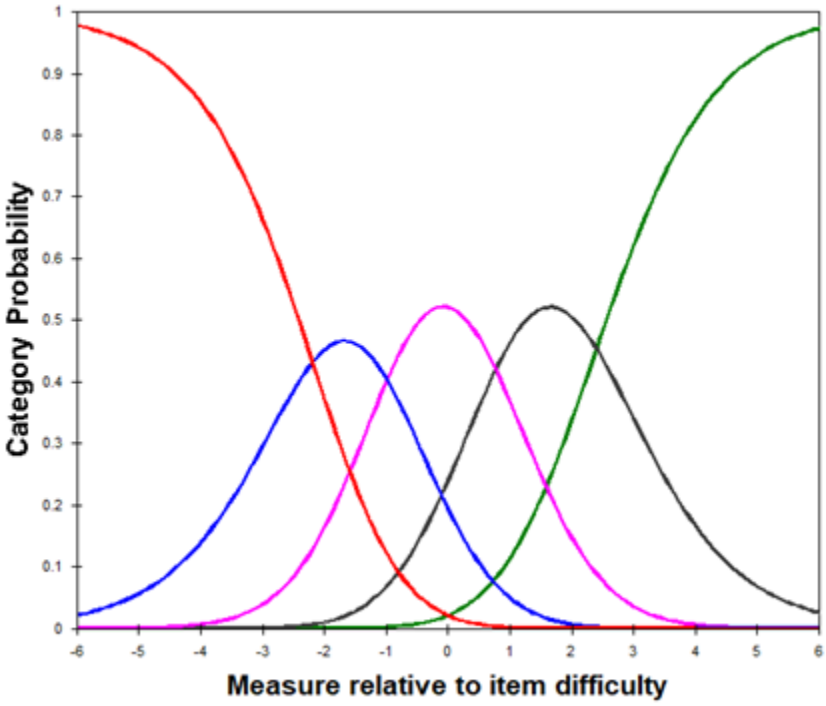

Figure 1B. Extroversion

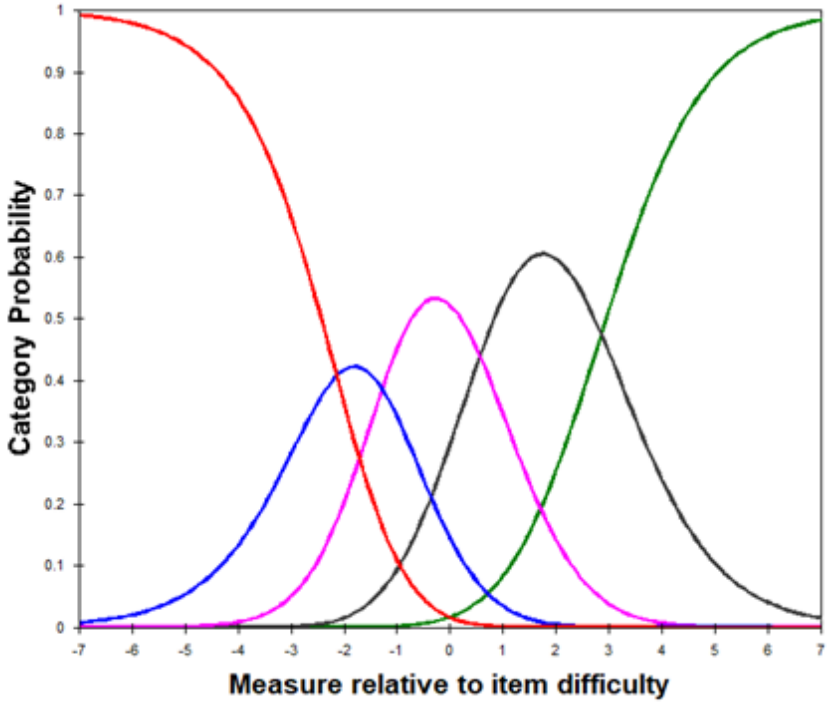

Figure 1C. Pace

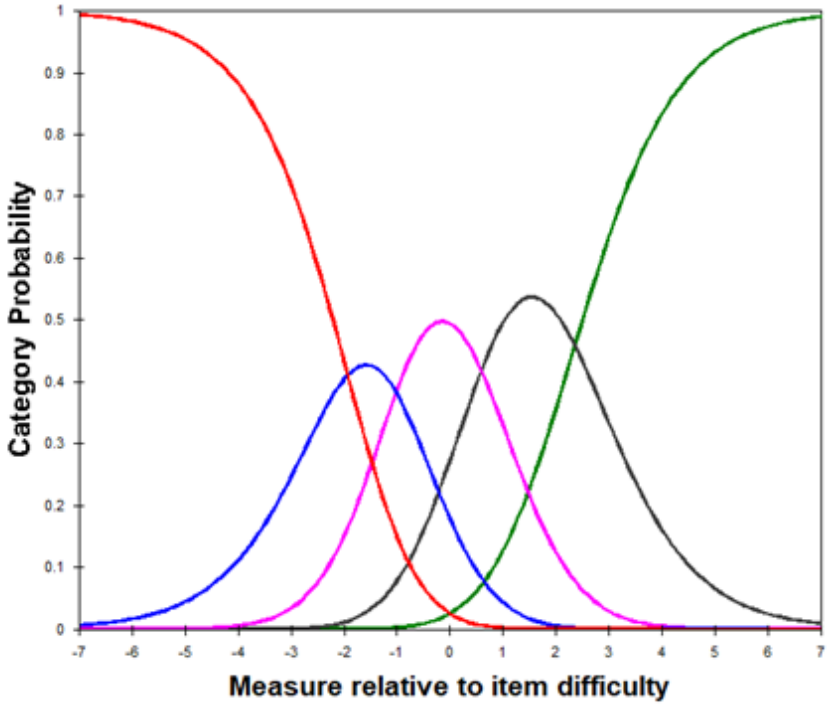

Figure 1D. Conformity

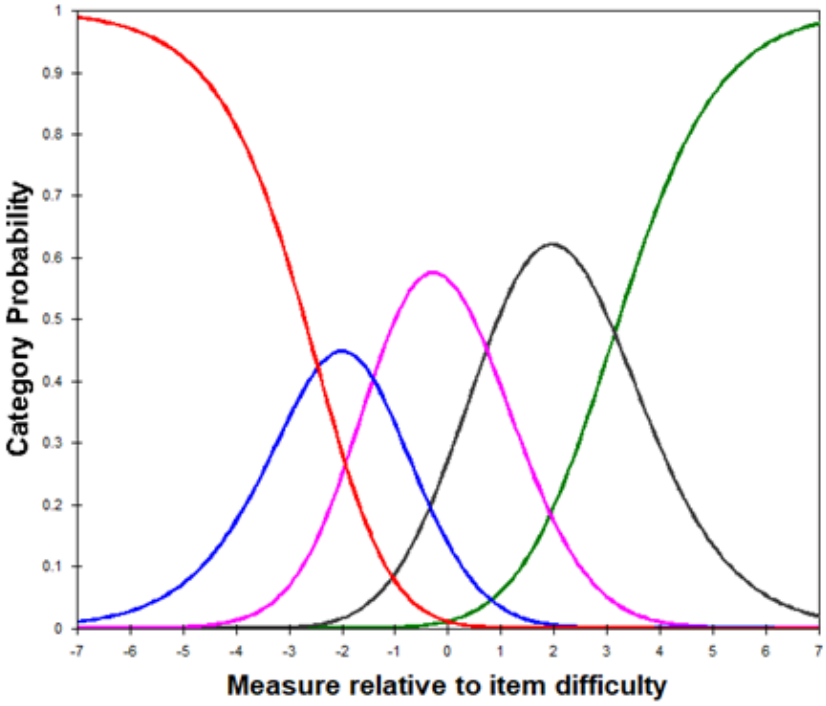

Figure 1E. Logic