

Elevating Program Evaluations: Using Psychometrics to Enhance a Program Evaluation

Tony Gage, MS

Director of Business Intelligence and Analytics

Email: agage@cealliance.com

Sarah Nisly, PharmD, MEd

Vice President, Outcomes and Clinical Impact

Email: snisly@cealliance.com

Clinical Education Alliance

Study Goals

This is the second in a series of studies that is guided by a single idea: Continuing education providers often leave valuable information and insights on the table when collecting and analyzing evaluation data. The first study in this series applied Item Response Theory (IRT) responses to a standard set of evaluation questions to establish and validate a single satisfaction scaled score metric. The key findings from that effort are summarized in Table 1. Based on the findings of that research, this study has three primary aims.

1. Increase the amount of test information (i.e., reliability) at higher levels of learner satisfaction.
2. Shift from a five-point to four-point agreement scale.
3. Maintain or improve scale reliability.

Methods

Satisfaction Scale Item Selection and Construction

Key stakeholders collaborated to update that standard evaluation form with the three above goals in mind. Although 23 items exist on the updated 2024 CEA standard evaluation, 13 items were included in this validation study (Table 2). Chosen items were a mix of questions that existed on the previous versions of the satisfaction scale and items that were written in the interest stated goals. Table 2 also indicates what rating scale is used for collecting responses: newness, change, agreement, or confidence.

Table 1: 2023 Research Leads to 2024 Evaluation Updates

Key Research Findings	Practical Implications	Actions taken in 2024 forms
1. The scale struggles to differentiate amongst individuals with high levels of satisfaction, which results in lower reliability of high satisfaction scores.	• Lack of confidence in precise location of positive ratings. • Difficult to truly differentiate amongst satisfied learners and • Harder to notice when learners think we have gone above and beyond expectations.	• Develop and pilot new questions that are more difficult for learners to endorse (i.e., harder to provide high ratings)
2. Some questions have highly correlated residuals (error terms), suggesting existing items are providing redundant contributions to scale scores.	• Redundancy in information being collected from learners, which requires extra time for learners without providing incremental information gains.	• Remove redundant questions from the form to make space for new questions.
3. The bottom three response options on the 5-point agreement scale are almost never utilized.	• This issue is causing much of what we observe in #1. • Learners largely agree or strongly agree in their ratings. • Results in what amounts to a 3-point scale rather than what was intended to be a five-point scale.	• Change from 5-point to 4-point agreement scale. • Update agreement scale anchors to provide more granularity in the measurement of learner satisfaction.

Table 2: Selected Satisfaction Scale Items

Item Name	Question Text	Rating Scale
NEW*	Approximately what percentage of content was new to you?	Newness
CHANGE	Based on your participation in this activity, do you intend to change your practice?	Change
ASSESS	Included appropriate and/or effective methods to assess my learning	Agreement
NONCLINICAL	Is helpful in growing non-clinical skills (e.g. self-advocacy, well-being, etc.)	Agreement
COMM	Promoting effective communication among members of the interprofessional team	Agreement
RIGOR	Was scientifically rigorous and evidence based	Agreement
SUPPORT	Understanding how MY ROLE supports others within the interprofessional team	Agreement
SOFTSKILL*	Provided guidance for soft skill (e.g. communication) improvement	Agreement
PREVEERROR*	Will help prevent patient errors or enhance patient safety in practice	Agreement
RELEVANT	Is relevant to my clinical practice	Agreement
KNOWLCOMP*	Improved my knowledge or comprehension on this topic	Agreement
CONFIDENT	How confident are you that you can make the intended changes?	Confidence
RECOGNIZE	Recognizing abilities and contributions of OTHER interprofessional team members	Agreement

* Indicates new items. Other items were previously on the scale.

Final item selection

Because of a low point-measure correlation (degree of correlation between responses on this question and total scores), the item labeled 'NEW' was removed from the scale. In all likelihood, this item represents an additional dimension related to learner satisfaction, and this does not conform to the unidimensional Rasch model being applied. The removal of this item results in a 12-item scale for further analyses.

Validation Analysis

Responses from 529 learners were calibrated using the WINSTEPS statistical program, which applies the Rasch measurement model. This is a 1-parameter IRT model that is commonly used in the field of educational measurement to construct scales of measurement. The Rasch calibration results in placing people and items on a common scale of measurement using a logit metric. Higher values on this metric are associated with more satisfaction (or higher difficulty in endorsing for items), whereas lower values are associated with lower levels of satisfaction (or lower difficulty in endorsing for items).

Results

Goal 1: Increase Test Information at Higher Score Points

Test Information Function

The first focus of the new items on this scale was to gain more test information at higher points on the scale. When comparing Figure 2 (Data from updated forms) to Figure 1 (Data from original form), the goal of gaining more information at higher points on the latent ability measure (x-axis) was achieved. The practicality of this result is gaining more reliability for scores that fall at higher points on the ability metric, which results in greater ability to differentiate among satisfied individuals.

Figure 1: 2022 Test Information Function

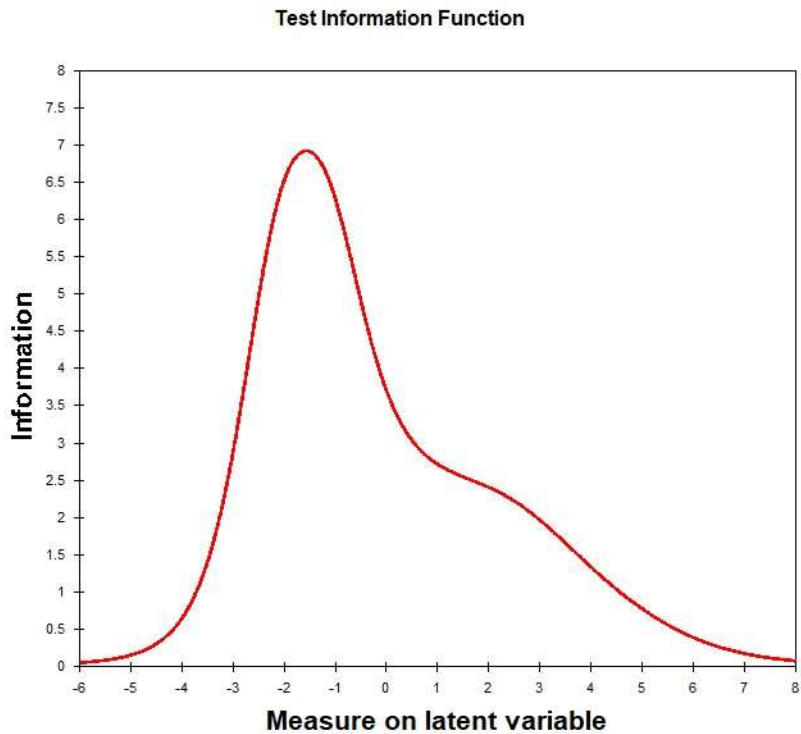
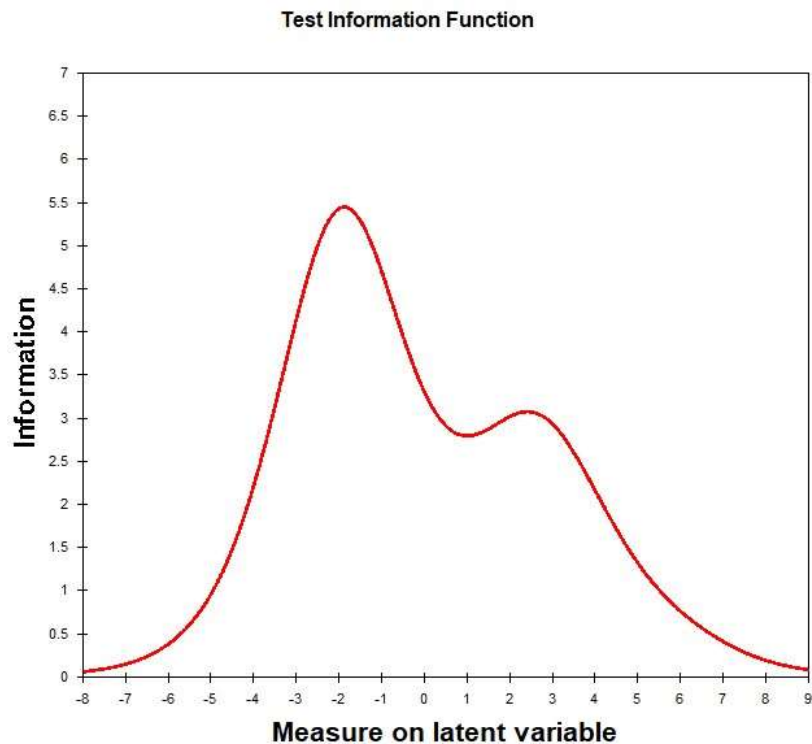


Figure 2: 2024 Test Information Function



Goal 2: Shift From a Five-Point to Four-Point Agreement Scale

Rating Scale Category Usage and Response Frequencies

One key goal was to successfully move from a five-point agreement to a four-point scale. The underlying measurement purpose for this is that prior research had suggested that not all options were used when learners were provided 5 options, and thus the scale was functioning effectively as a four-point scale anyway.

Table 3: Response Scales and Category Response Frequencies.

Rating Scale Name	Value	Rating Anchor	N	%
Newness	1	0%	11	2%
	2	25%	114	22%
	3	50%	186	35%
	4	75%	122	23%
	5	100%	96	18%
Change	1	I need more information before I will change my practice	52	10%
	2	My current practice has been reinforced	228	43%
	3	Yes, I plan to implement changes	249	47%

Agreement	1	Disagree	27	1%
	2	Somewhat Agree	182	3%
	3	Agree	1682	32%
	4	Strongly Agree	3399	64%
Confidence	1	Not At All Confident	15	3%
	2	Somewhat Confident	122	23%
	3	Confident	232	44%
	4	Very Confident	160	30%

Category Probability Curves

One big change that was made to this standard evaluation was to move from a five-point to four-point agreement scale. The main findings in previous research that led to this adjustment were twofold: First, there was very little usage of the bottom 3 rating options on the five-point scale, and second, a rating of '2' on a five-point scale was never the most likely response when looking at the category probability curves.

In Figure 3 and 4, each colored line plots the probability that a person with a given satisfaction score (x-axis) is to select each option. A key component of Rasch measurement is that each answer option on a rating scale should be most likely at some portion of the underlying latent trait ability. In Figure 3, you can see that the blue line (representing answer option 2 of 5) was never the most likely response. As Figure 4 shows, the issue of rating options that are never the most likely response has been resolved with the modification of the agreement scale. Furthermore, the order of when each answer option is most probable is correlated with overall scores. In other words, as you move from left to right each line each option that is most probable increases by one rating scale: red = Disagree, blue = Somewhat Agree, pink = Agree, and black = Strongly Agree.

Furthermore, a look at the response frequencies reveals that learners continue to predominantly use the top two agreement ratings, and thus the removal of one of the lower three ratings was justified (See Table 3).

Figure 3: 2022 Agreement Category Probability Curves

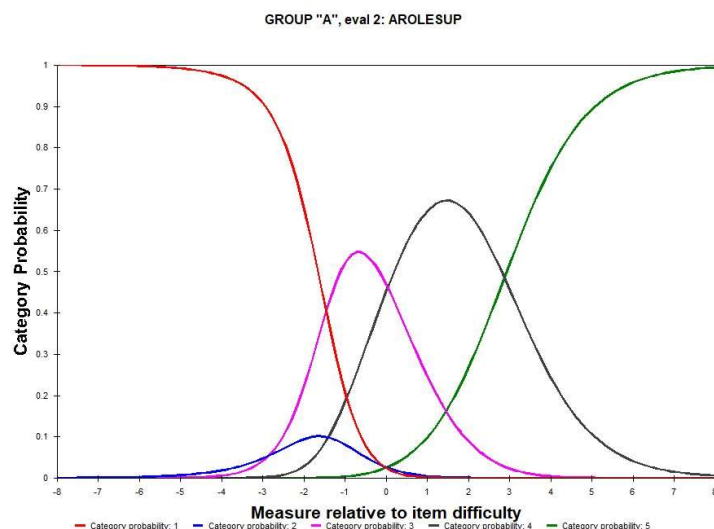
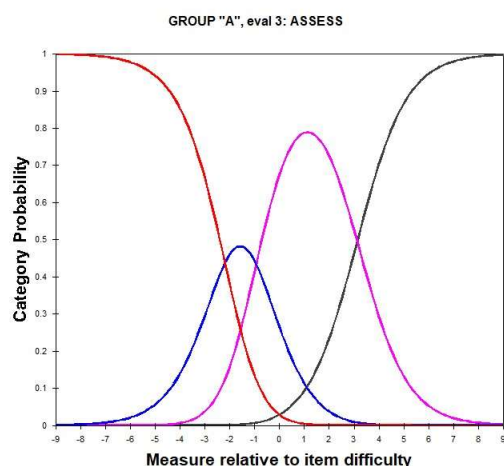


Figure 4: 2024 Agreement Category Probability Curves



Goal 2: Maintain or Improve Scale Reliability

Measure Order

In Table 4, the items are ordered based on their logit measure of item difficulty from most difficult to easiest (to endorse). One aspect of a quality metric is that the rank order of questions reasonably conforms to theoretical expectations. As an example, it is theoretically consistent that CHANGE and CONFIDENT are higher on the measure scale than the items of RIGOR, KNOWLCOMP, ASSESS, and. It is a theoretically lower bar to get a learner to agree that an activity is rigorous, improves knowledge, or adequately assesses their abilities, than to get the same learner to agree that they are confident in their planned changes or that they are going to make a change in practice. These data fit the theoretical model.

Point Measure Correlations

Point Measure correlations can be interpreted like any correlation, thus positive correlations mean that an increase in ratings on that question are also associated with increases in total scores.

All questions had notably positive point-measure correlations (See Table 4). This suggests a scale made up of items that are reasonably interrelated and measuring a single construct. This result suggests the scale should produce reliable results.

Table 4: Item Logit Measures and Point-Measure Correlations

Item Name	Measure	Point-Measure Correlation
CHANGE	2.14	0.36
CONFIDENT	2.10	0.47
RELEVANT	0.26	0.71
SOFTSKILL	-0.11	0.78
PREVEERROR	-0.21	0.81
COMM	-0.31	0.77
SUPPORT	-0.32	0.79
NONCLINICAL	-0.42	0.73
RECOGNIZE	-0.42	0.75
ASSESS	-0.68	0.78
KNOWLCOMP	-0.79	0.75
RIGOR	-1.26	0.70

Test Information and Scale Reliability

There are many ways of estimating the reliability (i.e., the degree of consistency expected from repeated measurements) of a psychometric measurement. One of the many traditional methods is to calculate Chronbach's Alpha for internal consistency. This value ranges from 0 to 1, with values closer to 1 indicating very high reliability overall. One reliability metric in the Rasch and other IRT models is test information. This measure literally describes the amount of information the metric has at a given latent trait score point. While traditional reliability estimates assume that reliability is the same across an entire measure, test information changes across the spectrum of possible scale scores (see figures 1 and 2). One way to compare one scale to another is the amount of test information at the mean ability score. Table 5 provides these two reliability metrics and shows that incremental gains have been attained.

Table 5: Reliability Estimates.

Reliability Estimate	2022 Form	2024 Form
Chronbach's Alpha (Internal Consistency Reliability)	0.72	0.80
Test Information at Average Person Score	1.38	2.25

Discussion

This study was successful in its stated goals of increasing test information at higher levels of satisfaction, moving from a five- to a four-point agreement scale, and improving reliability of the satisfaction metric. The steps to ensure the valid and reliable measurement of educational satisfaction have set the groundwork for gaining greater insights from this valuable evaluation data. The end goal of all of this work is to automate the scoring of evaluations as they come in from learners and calculate activity scores by averaging the person scores for those that complete the evaluation on each activity. An example of the type of data that can come out of this work is in Figure 5.

Figure 5: Histogram of Activity Scores from 2022 Form ratings



The next steps for this data are as follows:

- Equate New Form
 - Psychometric work necessary to make scores on the 2022 forms comparable to scores from the 2024 forms.
 - Committing to this work allows CEA to have multiple years of data available to begin development of next steps.
- Summative Reporting Additions to Educational Insights and/or Outcomes Reports
 - Program/Activity Scorecards
 - Faculty Scorecards
- Business Intelligence
 - How does this metric relate to other important business and educational metrics and contribute to the overall outcomes story?

References

1. Smith EV, Smith RM, eds. *Introduction to Rasch Measurement: Theory, Models and Applications*. JAM Press; 2004.
2. John Michael Linacre. Winsteps® Rasch measurement computer program (Version 5.6.0). Published online 2023. Winsteps.com