Using Rasch Analysis to Construct a Trust in Medical Technology Instrument

A measure of trust in medical technology is important to the assessment and design of technology in health care systems where users might not use a machine that is believed to be untrustworthy (Parasuraman, 1997). Parasuraman (1997) argued that issues of trust, mental workload, and risk can influence automation use, but system factors and individual differences make it difficult to predict automation usage. This may be particularly true in dynamic systems such as healthcare, in which patients and physicians have different issues regarding trust, workload and risk. Lee and Moray (1994) found that participants did not chose automation options if they had lower trust in the automation than their confidence in their own ability. These issues are particularly important in the culture of medicine, where doctors are expected to have a high confidence in their own abilities and patients may have a lower confidence in their abilities. The data described in this validation study were collected from students at a large rural Research I university. Eighty participants (52 female and 22 male) completed the measure for preliminary analyses.

Instrumentation

The *Trust in Medical Technology Instrument* (TMT) measures a persons trusting attitudes towards medical technology, the instrument has 72 items. Each item uses a 5-point likert type scale. The items for the instrument were derived from literature about trust in technology and placed in the healthcare domain. Item categories were created based on an internal model of the construct which specified the components of the construct; an external model of the construct which showed the relationship of the construct to other constructs and a process models that showed where trust in medical technology lies in the human information process. Each item was categorized according to their intended subscale, trust is medical technology, trust in physician and trust in medical system. Example items include, “I trust blood pressure machines more than I the manual method” and “Doctors sometimes pretend to know things when they really aren’t sure,” and “If a mistake was made in my health care, the health care system would try to hide it from me.”

The instrument development process involved several administrations of the instrument. During item prototyping, three subject matter experts from a human factors graduate program were asked to look at the initial pool of items. The experts were given the instrument ahead of time and were asked to provide feedback of the items’ accuracy, communicability, difficulty, bias, importance, conformity, specifications and relevance. Based on the results of the initial review of items changes were made and new items were developed. The new instrument includes items related to trust in physician (Kao *et al.*, 1998), trust in technology (Jian *et al.*, 1998b), and trust in the healthcare system (Rose *et al.*, 2004).

Next, we collected validation data for the instrument. The remainder of this article summarizes the results of that validation study.

Administration

The validation study data was collected through an online survey administration tool and administration took place in several steps.

1. Participants were provided with a brief overview of the study.

   You are being asked to participate in study about your healthcare experiences. To participate, you will be asked to complete a survey online. The survey should take less than 30 minutes to complete.

2. Participants volunteered for the pilot study by responding to the advertisement.
3. An email was sent to the participants with the online link to the survey.
4. Once the participant viewed the link, they first were directed to read and type in their initials for the informed consent form.
5. Next, participants were given instructions to complete the instrument.

6. Once the participant completed the instrument, they were ask to select submit.

Analyses

Data were scaled using the Rasch Rating Scale model, which is appropriate for Likert-type responses such as those from the TMT (Wolfe and Smith). That model depicts the probability (πnix) that a specific respondent n will rate a particular item i with a specific rating scale category x. In the model, £cn represents person n's trust level, £_i represents item i's endorsability, and £nj represents the rating category threshold between category x and category x+1. The TMT instrument contains three different 5-point rating scales. These rating scale model includes a threshold difficulty parameter (t_k) that depicts the difficulty of moving from one scoring category to another on a polytomous item. It is assumed that the distance between each category threshold is constant across items within the same rating scale. The *WINSTEPS* (Linacre, 2002) software package was used to estimate the parameters in the model based on observed item responses. Parameter estimates are reported on a single linear continuum in logistic odds ratio units (logits) (Smith, 2000).

A principal component analysis was performed to evaluate the dimensionality of the data. The Rasch rating scale model assumes unidimensionality and the measures are scale to that single dimension. Conducting a principal component analysis of the residuals (i.e., the difference between the Rasch model expected outcomes and the observed responses to each item) reveals whether dependencies between items exist beyond those assumed by the unidimensional model. In this validation study, each residual component's eigenvalue was interpreted using Kaiser's criterion, which suggests that only components with an eigenvalue greater than 1.00 should be retained (Kline, 2005). Components with four or more loadings with absolute values greater than .60 were selected as being sufficiently reliable for interpretation (Kline, 2005), and items with absolute loadings greater than .30 were interpreted as indicators of a particular component. Each dimension was then interpreted based on apparent similarities in the content of items that load on that component. Based on these results, items were grouped into subscales for further analysis.

Evaluation of the functioning of the rating scale was conducted to determine whether respondents utilized the rating scale of the TMT as intended. Specifically, we evaluate the suitability of the rating scale structure for responses to the TMT items using guidelines presented by Linacre (Linacre, 2004). These guidelines indicate that each rating scale category should contain a minimum of 10 observations, the shape of each rating scale distribution should be smooth and unimodal, the average respondent measure associated with each category should increase with the values of the rating scale categories, the unweighted mean squared fit statistics should have values less than 2.0, the category thresholds should increase with the values of the rating scale categories, there should be coherence (agreement) between the model-based expected category for each respondent-by-item combination and the observed ratings (referred to here as *coherence*, and measured as the % of observed ratings that are within the model-based expected category and vice versa), and that adjacent category thresholds should be at least 1.4 logits apart and no more than 5 logits apart.

Evaluation of item quality focused on point-measure correlations (i.e., the correlation between item responses for each respondent and the estimated measure for each respondent), which indicate the degree to which are analogous to the item score-total score correlations. Point-measure correlations over .5 are considered good, while .3 or .4 require scrutiny of the item. In addition, we considered mean-squared item fit indices, which indicate the degree to which the scored responses of individual respondents are consistent with the expectations of the Rasch model (Wolfe and Smith). Item fit indices were calculated for using WINSTEPS computer software (Linacre, 2002). Generally, values of item fit greater than 1.4 indicate misfit between the observed data and the model-based expectations. Items that did not meet these requirements were scrutinized for problems in wording, format, or item position that may cause it to function in a way that is different from the other items.

The reliability of the measures for each of these subscales was computed via WINSTEPS (Linacre, 2002) and is computed as one minus the ratio of the average error variance for the parameter estimates (i.e., the mean squared error) divided by the variance of the estimated parameters. As is true for coefficient alpha, this quantity represents the proportion of observed variance that is accounted for by true variability between respondents.

Results

Dimensionality

The principal components analysis provided strong evidence for a multidimensional structure of data from the TMT, which lead to the creation of three subscales. Evaluation of the eigenvalues from a principal component analysis of the residuals from the Rasch model via Kaiser's rule indicated six potential subscales, but only three of these contained a sufficient number of items with strong loadings and were logically consistent with substantive theory. Table X1 displays the eigenvalues for the component common to all items and each of the first five residual components. This table also identifies the items that exhibited absolute loadings greater than .30 on each residual dimension. Examination of the item text for the first three residual components suggests that the instrument includes items related to three constructs: Trust in Technology, Trust in Physician, and Trust in the Healthcare System. Hence, the remaining analyses focused on measures from items grouped into these subscales.

Table X1

Principal Component Analysis Summary

Component	Eigenvalue
Common	29.02
Residual 1	5.67
Residual 2	3.02
Residual 3	2.53
Residual 4	2.11
Residual 5	1.93

Note: The common component is the component accounted for by the unidimensional scale via the Rasch model. The residual components were extracted by conducting a principal component analysis of the residuals between model-based expectations and observed ratings.

Examination of the text of the items associated with the first residual factor reveals that this factor is bipolar (i.e., contains items with both positive and negative loadings). Items that exhibit high positive loadings on this factor are associated with Trust in Medical Technology whereas items that exhibit high absolute negative loadings on that factor are associated with Trust in Physician. The second component is also bipolar. The second component contains items with positive loadings concerning reliability, confidence, trust, and dependability. It also contains items with negative loadings concerning skepticism and suspicion about health care such as doubt, distrust, and lies. This finding is consistent with literature that states that trust and distrust are theoretical opposites and not separate constructs. The third component contained positively loading items related to privacy, disclosure and intent (e.g., the belief that medical records are not kept private or devices misinforming imply an intention to deceive).

Rating scale analysis

Linacre's first guideline requires each rating scale has at least 10 observations, and the second column of Tables X2 through X4 confirm this. In addition, this column indicates that there is a single peak in the distribution of ratings—Linacre's second criterion. The third guideline requires that the average measures associated with each rating category increase with the rating scale categories. The third column of the table indicates that the measures did increase for each rating scale. The next guideline requires unweighted fit statistics less than two, and the fourth column of each table indicates that this requirement is met in each rating scale. The next guideline requires difficulty of exceeding a threshold between two adjacent rating categories increases with the values of the rating categories. Additional requirements for these indicates concerns their spacing—they should be spaced at least 1.4 logits and less than 5.0 logits. These criteria are generally satisfied, although the mid-level category thresholds are somewhat close together in Tables X3 and X4. Finally, Linacre's guidelines indicate that the observed rating should imply the measure and vice versa. The final two columns of these tables indicate the percentage of expected ratings that were consistent with the observed rating (sixth columns) and the percentage of observed ratings that were consistent with the expected rating. All three tables indicate that there are minor problems with the coherence of the ratings for the model, particularly in the lowest rating category.

Table X2 Trust rating scale for Items 1-53

Category	Count	Measure average	Unweighted MNSQ	τ_k	Coherence M->C[a]	Coherence C-> M[b]
1	33	-0.92	1.32	(-4.51)	0%	0%
2	393	-0.35	1.13	-2.1	57%	26%
3	744	0.33	0.84	-0.27	39%	52%
4	1579	1.13	0.99	2.05	70%	76%
5	126	1.67	1.02	-5.07	0%	0%

Table X3 Trust rating scale for Items 54-63

Category	**Count**	**Measure average**	**Unweighted MNSQ**	τ_k	**Coherence M->C**[a]	**Coherence C-> M**[b]
1	43	-1.21	1.44	NONE	0%	0%
2	352	-0.44	0.89	-3.1	63%	28%
3	607	0.45	0.82	-0.43	40%	60%
4	1008	1.05	1.07	0.26	64%	69%
5	143	1.77	0.94	3.26	100%	2%

Table X4 Trust rating scale for Items 64-73

Category	**Count**	**Measure average**	**Unweighted MNSQ**	τ_k	**Coherence M->C**[a]	**Coherence C-> M**[b]
1	12	-1.21	1.44	NONE	0%	0%
2	175	-0.44	0.89	-3.1	63%	28%
3	229	0.45	0.82	-0.43	40%	60%
4	312	1.05	1.07	0.26	64%	69%
5	67	1.77	0.94	3.26	100%	2%

Item Quality

Point measure correlations and item fit indices indicated that several items did not function as intended. Eleven items were flagged based on these item analyses (see Table X5).

Table X5

Items flagged for further evaluation

Item	Target	rpm(scale)	MSw
5	machine more than MD	0.27	1.71
33	MT is unfailing.	0.13	1.35
36	familiar with MT before trust	0	1.9
36	familiar with MT before trust	0.04	2.03
37	MD doesn't care	-0.31	1.92
37	MD doesn't care	-0.33	1.64
45	I trust that MDs will tell tr	0.28	1.44
46	MD may not keep info private	0.37	1.48
47	I can tell a MD anything.	-0.32	2.22
64	they do experiments w/o permission	0.34	1.75
64	they do experiments w/o permission	0.23	1.74

*Red indicates factor one, blue is factor two, and green is factor three.

Reliability

The TMT produce high reliability across the various subscale. Specifically, the person separation reliability indices for the three subscales were .85, .67, and .67. These reliability coefficients are sufficiently high for most research purposes.

Conclusion

The TMT requires several revision before it is considered a high quality instrument. Eight items were flagged for further evaluation. To increase person reliability, persons will be tested with more extreme levels of the factor (high and low) and the test will be lengthened (Linacre, 2002). Sample populations that have more experiences with health care systems can achieve larger variations in the level of the trait. To increase item reliability, more participants will be used in subsequent samples.

Because items 20-41 look like they are only really picking up four points, the neutral (point 3) will be removed and re-piloted. Section three, items 31-41 look like it might be functioning as a dichotomous item, therefore these items will be re-piloted as such. Section four items 42-73 appear to be functioning appropriately for the 5 point rating scale, so these items will be left alone. Linacre's rating scale guideline number six says that the rating should imply the measure and vice versa, which is determined with the coherence statistic. Linacre's guideline was not met, because of several low coherence statistics. This low statistic most often occurred for category one on the rating scale, it is likely that the model predicted person to choose a one or two and they chose to opposite. This low coherence statistic may not be very important.

A major weakness of the instrument is its failure to match the proposed theoretical model in terms of dimensionality. Future research should be conducted to determine the relationship of the factors to more clearly develop the construct trust in medical technology.