CONSTRUCT VALIDATION OF ATTITUDE TOWARD ARTS AND ART EDUCATION SURVEY FOR PRESERVICE ART TEACHERS USING RASCH MODEL TECHNIQUES

Dr. Khaleda Zaid al-Kaylany
Lecturer (Ph.D.), Department of Curriculum and Instruction, Hashemite University (JORDAN)  
khaleda@hu.edu.jo

DOI: 10.7813/2075-4124.2012/4-5/B.10

ABSTRACT

The purpose of this study was to validate the attitudes toward arts and arts education survey, preservice teachers form, using Rasch model analysis techniques. The survey consists of (40) items distributed to four dimension or subscale (Societal, personal, professional and economical). The sample of the study was all students enrolled in the preservice art teacher preparation program at the Hashemite University (110 students, mainly females). Rasch analysis program (e.g., Winsteps 3.65) was used to calculate in-Fit and Out-Fit mean square statistics to evaluate how well each item fits the model. Fit statistics refers to whether the items are of sufficient quality to interpret the outputs in interval scale, and whether each item contributes to the measurement of only one construct. The values of in-Fit MNSQ of all survey items except seven (items 14 and 110 from societal dimension; items 23, 27 and 28 from personal dimension; and item 35 from professional dimension) fall within the accepted range for fitting the model. Based on the results of fit statistics, it can be concluded that overall the data fit the model very well and the items contribute to the measurement of one construct, i.e. attitude toward arts and art education. Further validation study using classical test theory and confirmatory factor analysis recommended.

Key words: Teacher attitudes, Attitudes toward Arts, Art Teacher education, Rasch analysis, instrument validation

1. INTRODUCTION

During the past decade, the arts have been increasingly included in per-service and in-service professional development programs for school teacher education in Jordan, as a response to the educational reform movement. Little is known, however, about teachers’ attitudes toward the arts in education or the applications of arts processes in their teaching practice (Orrick, 2004). Attention to the attitudes of pre-service and in-service primary teachers toward the arts of fundamental importance to research on primary education. However, progress in this field of research has been slow due to the poor definition and conceptualization of the construct of primary teachers’ attitude toward the arts. This poor theoretical background has led to the use of a multitude of different concepts and measurement instruments.

Rasch measurement theory (RMT) refers to a family of statistical models and techniques used to assess the quality of tests and questionnaires, and to construct true interval-scale measures from the raw scores obtained from such instruments (Sick, 2008a). RMT can be said to embody a theory of psychological measurement, in that it provides a set of prescriptive criteria for judging the degree to which measurement has been successful. It can also play an important role in the process of construct validation, in that a set of test or questionnaire items constitute the instrument designer’s empirical definition of the construct (Sick, 2008b). A Rasch analysis can be used to evaluate the degree to which responses to these items conform to what would be expected if the items were indeed measuring a single, coherent psychological attribute.

Rasch measurement differs from Classical Test Theory (CTT) in several appreciable ways. First, CTT is primarily concerned with total scores. Total scores, the number of items answered correctly, serve as the sole indicator of a person’s level of ability or knowledge, and all items are treated as equal contributors to the total score (Sick, 2009a). That is, difficult items are not weighted more highly than easier items when estimating levels of knowledge or ability. Moreover, equal differences in total scores are treated as delineating equal ranges of ability. In CTT, the primary indicator of test quality is reliability, the overall capacity of the test to define levels of knowledge or ability consistently. Reliability estimates are usually derived by comparing the total scores of one half of the test to the total scores of the other half, either directly, or indirectly using formulas such as Cronbach’s alpha that approximate the mean correlation of all possible split-half combinations (Sick, 2008b).

A second difference is that CTT is primarily descriptive and sample dependent. It provides a description of the data from a single administration of a test, including the reliability of that administration, and the difficulty of individual items in terms of the percentage of test takers who answered them successfully.

Baku, Azerbaijan | 73
Bond and Fox (2001), among others, have highlighted three common measurement issues that threaten validity and reliability in traditional quantitative research methods. First, researchers often mistakenly assume that Likert-scaled items are at the interval level of measurement and that item endorsements mean the same thing for all respondents. Second, that all of the items on the inventory are tapping into or measuring all levels on the latent construct of interest. And lastly, that multiple factors influence item response patterns, including those of the individual (e.g., being too tired), as well as those associated with a single item or group of items (e.g., poor wording or confusing directions).

Historically, there has been no systematic way to explore these issues using traditional psychometric techniques (Bond & Fox, 2001). The Rasch model, allows researchers to address each of these issues at both the person and item level. Ideally, the Rasch measurement model is best utilized during the initial development and validation of any new scale or inventory (Sampson & Bradley, 2003; Wright & Stone, 1979), however, there are times when Rasch can be used to improve the reliability and validity of existing measures (e.g., Pomeranz et al., 2008) and, when appropriate, in developing a shorter form of an existing measure (e.g., Cole, Rabin, Smith, & Kaufman, 2004). As Bradley and Sampson (2005) write "Rasch analysis begins at the level of measurement, providing diagnostic information on the quality of the measurement tool, in addition to yielding a more comprehensive and informative picture of the construct under measurement as well as the respondents on that measure" (p. 12). Originally introduced by George Rasch (1960, 1980), the Rasch model assumes that the relationship between an individual's ability level and an item's designed difficulty can be mathematically modeled as a probability. As ability increases, so does the probability of answering items correctly. Conversely, as ability level decreases, so does the probability of answering items correctly. This relationship is the fundamental principal that underlies the Rasch measurement model. When measuring personality, attitudinal or psychological constructs, however, the concept of "ability" can be viewed as "endorsability" (Fox & Jones, 1998).

By examining the psychometric and measurement properties of the Attitude towards art and art education survey from a Rasch measurement perspective, the present study was an attempt to accomplish two objectives: first, to illustrate how the Rasch measurement model can assist art education researchers in their pursuit of increased precision and accuracy in personality and psychological measurement; and secondly, to provide additional empirical support to the measurement and psychometric value of an established art teacher attitudes assessment survey. Through use of fit statistics, variable maps, and person and item reliability estimates, this study hope to provide empirical support, from a measurement perspective, for unidimensionality, hierarchical ordering and range, and overall item and person reliability.

2. SAMPLE

Research participants were 110 late adolescent Jordanian public university students, enrolled in the Art Teacher preparation program, (103 female, 7 male) with mean ages of 19.3 (females) and 19.2 years (males), and where ( 38 1st year students, 29 2nd year, 14 3rd and 29 at 4th year).

3. INSTRUMENT

The instrument used to gather data in this study is the Teachers attitude toward arts and art education survey developed by Al-Kharoby (2003) based in literature survey and a conceptual framework. The survey has been widely used in scholarly research in Arab world environment. The survey consists of (40) items distributed to four dimensions or subscales: Societal, personal, professional and economical, with 10 items for each dimension. Each statement in the survey is scored on a 5-point Likert-type scale as follows: (1) strongly disagree; (2) disagree; (3) occasionally; (4) agree; and (5) strongly agree for the positive items, and the scale reversed for negative items. According to Al-Kharoby (2003) the survey has acceptable reliability coefficients (0.86), ranging from a low of 0.70 for societal dimension to a high of 0.77 for the personal and economical sub-scales. The validity was verified using face validity procedures.

4. DATA ANALYSIS, FINDINGS AND DISCUSSION

Rasch analysis program (e.g., Winsteps 3.65) was used to calculate in-Fit and Out-Fit mean square statistics to evaluate how well each item fits the model. Fit statistics refers to whether the items are of sufficient quality to interpret the outputs in interval scale, and "whether each item contributes to the measurement of only one construct" (Bond and Fox, 2001; Sick,2009b; Linacre & Wright, 1999-2001).

Prior to interpretation of the item and person Logit (position) scores from a Rasch analysis, appraisal of whether the data fit the model reasonably well is required. Table 1 presents overall information about whether the data showed acceptable fit to the model.

The mean infit and outfit for person and item mean squares are expected to be 1.0. For these data, they are all around 1. The mean standardized infit and outfit are expected to be 0.0. Here they are 0 for persons and 0 for items.
The standard deviation of the standardized infit is an index of overall misfit for persons and items (Bode & Wright, 1999). Using 2.0 as a cut-off criterion, both persons (standardized infit standard deviation = .44) and items (standardized infit standard deviation = .20) show little overall misfit. Here the data evidence acceptable fit overall.

The next overall statistic we look at is called separation, the index of spread of the person positions or item positions. For persons, separation is 3.25 for the data at hand (real), and is 3.56 when the data have no misfit to the model (model). This suggests we have measured persons on a continuum rather than a dichotomy.

Separation equals 1.0 or below means that items may not have sufficient breadth in position. In that case, we may wish to reconsider what having less and more of the trait means in terms of items agreed or disagreed with, and on revision, add items that cover a broader range. Item separation for the present case is 5.57, an even broader continuum than for persons. It is typical to find larger separation values for items than for persons, a function of the fact that we often have a small number of items and a larger number of people, here 40 items and 110 people. Separation is affected by sample size, as are fit indices and error estimates. With larger sample sizes, separation tends to increase and error decrease.

The person separation reliability estimate for these data is .91. We are given a conceptual analog to person reliability in item reliability—which estimates internal consistency of persons rather than items (here, .97).

Note that the mean for items is 0.0. The mean of the item Logit position is always arbitrarily set at 0.0, similar to a standard (z) score. The person mean here is .50, which 40 persons suggests these items were easy, on average, for persons to agree with but were fairly well matched to the perceptions of the sample. The persons would have a higher level of the trait than the items do. If the person mean were -1, -2, or +1 or +2, we would consider the items potentially too hard or too easy for the sample and might seek a different test.

Table 2 displays how the response scale was used. For these data, the response scale was 1 (strongly disagree) to 5 (strongly agree) for the positive items and the reverse for the negative ones. Table 2 lists the step Logit position, where a step marks the transition from one rating scale category to the next, e.g., from a 4 to a 5. "Observed Count" is the number of times the category was selected across all items and all persons. "Observed Average" is the average of Logit positions modeled in the category. It should (and does) increase by category value.

Persons responding with a "1" have an average measure (-.95) lower than those responding with a "2" (average measure = -.37), etc. There is no substantial misfit for categories 1 through 5. Step calibration is the Logit calibrated difficulty of the step. These values are expected to increase with category value, which they do. The step standard error is a measure of uncertainty around the step calibration. Another view of step function is the Thurstone threshold, which is the location of the medians—where the point of observing the categories below equals the probability of observing the categories equal to or above that 16 point. The transition points between one category and the next are the step calibration values from Table 2. For these data, all categories are being used and are behaving according to expectation.

At this point in our analysis we know rating scale categories were used appropriately and the data overall fit the model. We then proceed to examine persons and item fit to see how individual items functioned and how individual persons responded to the items.

Table 3 displays the items in order of worst to best fitting. Entry number is the item's location in this scale of 40 items. Raw score is the total number of "points" the item got across the entire sample. Count tells us that of 110 participants, all responded to all items, not surprising since the instrument was individually administered. Measure is the logit position of the item, with error being the standard error of measurement for the item. No definitive rules exist regarding what is considered acceptable and unacceptable fit but some suggestions for acceptable fit are as

Baku, Azerbaijan | 75
follows: (1) Mean square (infit or outfit) between .6 and 1.4, (2) mean square (infit or outfit) between .8 and 1.2 (Bode & Wright, 1999), (3) mean square less than 1.3 for samples less than 500, 1.2 for 500–1,000, and 1.1 if n > 1,000 (Smith, Schumacker, & Bush, 1995), (4) standardized fit (infit or outfit) between -2 and +2, (5) standardized fit between -3 and +2, and (6) standardized fit less than +2 (Smith, 1992). Infit is a weighted goodness-of-fit statistic, where unexpected responses to items close to the person’s logit position are weighted more heavily than unexpected responses to items far away from the person’s level (information laden).

Outfit is un-weighted and so is sensitive to extreme unexpected responses (outlier sensitive). The score correlation is the correlation between item score and the measure (as distinct from a total score), and so is an item discrimination index. It should be positive.

### Table 3. Item Fit Statistics in Order by Misfit

<table>
<thead>
<tr>
<th>ENTRY</th>
<th>TOTAL SCORE</th>
<th>COUNT</th>
<th>MEASURE</th>
<th>INFIT</th>
<th>OUTFIT</th>
<th>FIT-MEASURE</th>
<th>EXACT MATCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>200</td>
<td>110</td>
<td>0.95</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>18</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>19</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>20</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>21</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>22</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>23</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>24</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>25</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>26</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>27</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>28</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
<tr>
<td>29</td>
<td>256</td>
<td>120</td>
<td>0.88</td>
<td>0.51</td>
<td>1.47</td>
<td>5.51</td>
<td>0.95</td>
</tr>
</tbody>
</table>

For these data, it appears that the item in the 17th, 18th, 19th, 20th, 21st, 22nd, 23rd, 24th, 25th positions does not fit well with the rest of the scale as its mean square and standardized infit and outfit values exceed all recommendations. It noticeable that all these items were negatively worded items, scale item development literature reveals that this sometimes arise from the inclusion of both negative and positively worded items (Spector, VanKatwyk, Brannick, and Chen, 1997). The influence of careless responding to reverse-worded items during factor analysis is also discussed by Woods (2006), and the author posits that these items may cause researchers to falsely reject the uni-dimensional assumption of IRT models. In scale revision, we might consider rephrasing or deleting these items.

Based on the results of FIT statistics, it can be concluded that overall the data fit the model very well and the items contribute to the measurement of one construct, i.e. teachers’ attitude toward arts and art education.

Figure 1, the map of persons and items displayed. The distribution of person logit positions is on the left side of the vertical line and items on the right. Each “X” represents two persons in this figure. “M” marks the person and item mean, “S” is one standard deviation away from the mean, and “T” is two standard deviations away from the mean. Those at the upper end of the scale agreed with more items and agreed more strongly. The distribution of person logit positions is positively skewed. In Figure 1, we can see that there are numerous persons whose position is below where items are measuring—there are no items that match these persons’ levels of the trait very well. We see that the items cover a range of -1 to +1 logits in difficulty, narrower than the range of about -5 to +25 for persons. As such, on scale revision an attempt to write easier and harder items to extend the range of the trait measured is needed.
There has been limited research on the validity and reliability of the attitudes toward arts and art education for use in Arabic-speaking countries especially Jordan. An instrument with validated constructs and known psychometric qualities is needed to advance knowledge in this important area and to use across cultures. In order to investigate this relatively unexplored area of research, the primary goal of this study is to establish a valid and reliable Arabic version of the attitude toward art and art education survey for use in Jordan, using Rasch analysis techniques.

**Fig. 1. Item-Person Map**

5. CONCLUSION AND RECOMMENDATIONS
Advantages of the Rasch method were highlighted. One was the ability to see a fuller picture of assessment from the individual perspective (e.g., a tired test taker) and item perspective (e.g., a confusing item), which allows a practitioner the confidence that the assessment is measuring the latent construct of interest (e.g., attitudes toward arts and art education). Also mentioned were the ability of Rasch to inform a test developer about the item and test taker reliability and validity as well as developing a shorter version of an existing measure, when appropriate. The study explained how to understand and interpret the key outcomes of Rasch, explaining the thresholds which an item should be considered inappropriate to be included in the final measure (e.g., infit statistics; variable maps) or when an item can be eliminated because it accounting for the same variance in the latent construct as another item (i.e., outfit statistics).

Findings revealed that the items in the 17th, 18th, 4th, 10th, 13th, 40 and 25th positions (i.e. items 14 and 110 from societal dimension; items 23, 27 and 28 from personal dimension; and item 35 from professional dimension) does not fit well with the rest of the scale as its mean square and standardized infit and outfit values exceed all recommendations. It noticeable that all these items were negatively worded items. In scale revision, we might consider rephrasing or deleting these items.

Based on the results of FIT statistics, it can be concluded that overall the data fit the model very well and the items contribute to the measurement of one construct, i.e. teachers’ attitude toward arts and art education.

Further validation study using other Rasch techniques and confirmatory factor analysis is highly recommended.

REFERENCES
Copyright of International Journal of Academic Research is the property of International Journal of Academic Research and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.