

Measurement and Evaluation Perspectives On Scaling Teacher Affect with Multiple Measures

Judy R. Wilkerson
Florida Gulf Coast University
U.S.A.

Abstract

While measurement and evaluation interact closely, they are typically reported separately. In this study, the Dispositions Assessment Aligned with Teacher Standards (DAATS) scale of commitment to teaching skills is reviewed using both measurement and evaluation professional standards. Evidence of validity and reliability (measurement) and evidence of utility and feasibility (evaluation) are presented. Results indicate that (1) U.S. national pre-service teaching standards, the INTASC Principles (CCSSO, 1992), yield a construct definition that can be measured accurately and precisely holistically and by Principle; (2) a Thurstone (1928) agree/disagree scale still works well in measuring the affective domain; (3) the Bloom and Krathwohl Taxonomy (Bloom & Krathwohl, 1956) is useful in defining scaling categories; (4) different types of affective instruments can be combined into a single and useful score (5) a well-designed measurement device supports evaluation decisions that are useful, feasible, and accurate; and (6) use of Rasch measures, enhanced by qualitative analysis of constructed responses, was productive at the individual and program levels. The research points to the need for increased assessment of teacher dispositions (affective domain) as an important indicator of teachers who should be promoted or re-directed.

Keywords: Teacher assessment, dispositions, affective domain, measurement, evaluation, standards, improvement, Rasch, Thurstone, INTASC, NCATE

The Ballad of East and West

Oh, East is East and West is West, and never the twain shall meet,
Till Earth and Sky stand presently at G-d's great Judgment Seat;
But there is neither East nor West, Border, nor Breed, nor Birth,
When two strong men stand face to face, though they come from
the ends of the earth!

Rudyard Kipling

Introduction and Background

“All Children Can Learn.”


All teachers should believe that all students can learn, and all teachers should work toward ensuring that all children reach their potential –the mantra of teacher education. The truth of the matter is that not all teachers believe this. How, then, do we know what they believe and how those beliefs are translated into actions that help or harm children? Objective measurement of teacher dispositions,

the affective component of teaching, followed by effective evaluation and use of the results helps answer the question central to this research.

The literature provides many examples of single assessments of teacher affect, e.g., surveys, indices, observations, or interviews, (Richardson & Onwuegbuzie, 2003; Lund, Wayda, Woodward, & Buck, 2007; Schulte, Edick, Edwards, & Mackiel, 2004; Wasicsko, 2004; Jung & Vogt, 2006; Singh & Stoloff, 2008). Single measures, especially surveys, however, can be misleading; answers can be socially acceptable (faked).

The qualitative paradigm suggests beginning with a vignette (Stake, 1995). The vignette in figure 1 illustrates this assessment issue, using a picture prompt from a thematic apperception test and the response of a non-empathetic practicing teacher - one who agreed with the survey item "all children can learn" but would ban an impoverished child from school. The prompt asked what kind of teacher is needed for this child and what the respondent would do with him.

SRA Prompt 6: Walking to School



...If this child were in my class I would inquire about his home conditions because he does not seem to have a proper home life ... **He would not even make it onto the bus or through the front door** for that matter of our school because **he is against code**. (He has no shoes on and his shirt is un-buttoned.) ... **the law is law** and he would be against dress code.

Source: Situational Reflection Assessment (SRA) in DAATS Battery. Judy R. Wilkerson, W. Steve Lang, B. Slitkin, reprinted with permission of the authors and artist.

Figure 1. A Teacher's Viewpoint of a Child

The issue of using multiple assessments and combining them in a meaningful way for individual teacher and program improvement purposes is not evident in the literature examined. In general, measurement properties are the predominant focus in one set of journals; evaluation predominates in others. They tend to be as separate as east and west, but here the "twain shall meet." In this research, multiple assessment methods are used and reviewed from both measurement and evaluation perspectives.

Measurement vs. Evaluation vs. Assessment

Measurement. Stevens (1946) defined measurement as "the assignment of numerals to objects or events according to some rule." He developed the scales of measurement (nominal, ordinal, interval, and ratio) still in use today. Georg Rasch (1960) established the mathematical relationship between a person's ability and the difficulty of an item, demonstrating that the probability of providing a correct response was related to the ability of the respondent. Item response theory and the

family of Rasch models permit ordinal level data, including dichotomous and rating scale items, to be converted to an interval scale. This allows more appropriate use of common statistics, providing advantages over a simple raw score (count) of correct responses.

Evaluation. Educational evaluation is the process of characterizing and appraising some aspect of the educational process. Procedurally, it uses both measurement and research techniques. Stufflebeam (2001) defined evaluation as “a study designed and conducted to assist some audience to assess an object’s merit and worth” (p. 11).

Assessment. McMillan (2007) differentiated measurement from evaluation, defining assessment as a process with four sequential steps: purpose, measurement, evaluation, and use. Purpose provides the context and rationale for assessing (i.e., validity). Measurement quantifies. Evaluation judges quality based on criteria. Use makes time well spent, with results applied to decisions about learning, including diagnosis, grading, and instruction.

Standards and Accountability

Three Sets of Standards. In an era of accountability, standards are paramount. Here three standards sets are integrated.

- Content is derived from the U.S. national set of pre-service teaching standards developed by the Interstate New Teacher Assessment Consortium (INTASC) for the Council of Chief State School Officers (CCSSO). These INTASC Principles (CCSSO, 1992) are paraphrased as:
 1. The discipline taught
 2. Learning and development
 3. Diverse learners
 4. Instructional strategies and critical thinking
 5. Motivation and learning environment
 6. Communication
 7. Planning
 8. Assessment
 9. Reflective practice and professional growth
 10. Professional relationships
- Measurement is based on the *Standards for Educational and Psychological Testing*, developed by the American Educational Research Association, the American Psychological Association, and the National Council of Measurement in Education (AERA, APA, NCME, 1999). The *Standards* are divided into three parts (Test Construction, Evaluation, and Documentation; Fairness in Testing; and Testing Applications) and 15 chapters (including validity, reliability, and program evaluation). The program evaluation chapter focuses on the use of standardized tests.
- Evaluation is based on *The Program Evaluation Standards* of the Joint Committee on Standards for Educational Evaluation (Yarbrough, Shulla,

Hopson, and Caruthers, 2011). The five standards categories are utility, feasibility, propriety, accuracy, and evaluation accountability.

Validity, Proposed Uses, Utility, and Feasibility. The measurement standards begin with validity in the first chapter:

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.

Validity is, therefore, the most fundamental consideration in developing and evaluating tests. (p. 9).

The measurement standards detail the technical requirements of validity regarding decisions related to individual scores. The evaluation accuracy standards target dependability and truthfulness of representations, propositions and findings, requiring that information serve the intended purposes with valid and reliable interpretations (Standards A2 and A3). However, the evaluation Utility Standards elevate the concept of “use” to programs with Standards U5 and U6 adding the concept of relevancy. Stakeholders must find the results valuable in meeting their needs, helping them discover, interpret or revise what they do. Use should be positive, with unintended negative consequences and misuse avoided, consistent with the concept of consequential validity. The evaluation feasibility standards target effectiveness and efficiency. Standards F2 and F4 require practical and responsive procedures and effective and efficient resources.

The Need for Multiple Measures

The U.S. focus on standardized tests as the single measure of student and school success, rooted in No Child Left Behind (NCLB) legislation, has generated much concern, as summarized well by the Association for Supervision and Curriculum Development in its position paper on *The Case for Multiple Measures* (Fuller, Fitzgerald, & Lee, 2008):

No single test can tell all there is to know. As the directors of the National Center for Research on Evaluation, Standards, and Student Testing emphasize, "Multiple measures are needed to address the full depth and breadth of our expectations for student learning" (p. 2). Beyond the multiple-choice and short-answer items that are typical of current assessments, "other types of performance measures—essays, applied projects, portfolios, demonstrations, oral presentations, etc.—are needed to represent and guide students' progress" (Herman, Baker, & Linn, 2004, p. 2).

ASCD's support of multiple measures to make informed judgments about student cognitive learning and the success of education programs applies equally to the affective domain. In his seminal work on measuring attitudes, Thurstone (1928) began by conceding that attitudes are complex and cannot be wholly described by any single number. He provided the analogy of a table that could be measured by height, length, and cost. Thurstone defined attitudes to include a person's inclinations, feelings, prejudice, bias, preconceived notions, ideas, fears, threats, and convictions about any topic. While he proposed using opinions to

measure attitudes, he acknowledged the uncertainty of this practice with the respondents' potential to lie, misrepresenting real attitudes. He suggested that while behaviors, or actions, might be safer, they, too, can be misleading. While Thurstone developed the agree/disagree scaling process used in the belief scale presented herein, he also counseled that a survey, no matter how good, is inadequate.

Multiple measures, then, appear necessary; however, they can present different results for an individual. Combining the measures into a single measure (score) to diminish the impact of the inevitable problems could be a viable solution that provides for feasibility in decision-making (evaluation) while ensuring the quality of the scores (measurement).

Choices Among Affective Measures and the DAATS Battery

The Dispositions Assessment Aligned with Teacher Standards (DAATS) model (Wilkerson & Lang, 2007) suggests various strategies for assessing affect. Extensive discussion of the literature, the model, the instruments, and the results has been presented previously (Wilkerson & Lang, 2004; Wilkerson & Lang, 2006; Lang & Wilkerson, 2008; and Englehart, Batchelder, Kelly, Wilkerson, Lang, & Quinn, 2011; Wilkerson & Lang, 2011). The DAATS Battery (Wilkerson & Lang, 2006) consists of five assessments instruments, all of which measure all ten INTASC Principles. The three used in this study are asterisked.

- Beliefs About Teaching Scale (BATS): a 60 item Thurstone agreement scale*
- Experiences in Teaching Questionnaire (ETQ) includes: 10 constructed response items about prior experiences*
- Situational Reflection Assessment (SRA): 20 constructed response items, using picture prompts (Slitkin, 2007), in a thematic apperception format*
- Classroom Dispositions Checklist (CDC): 50 paired statements of positive and negative behaviors.
- K-12 Dispositions Impact (KIDS): focus group with 10 clustered prompts measuring children's perceptions.

Psychometric Challenges in Affective Measurement

The affective domain presents measurement challenges. Respondents often anticipate the "correct" answer (a socially acceptable response), providing it whether or not they believe it (Edwards, 1959; Taylor, 1961). This deficit suggests the need to take advantage of some potentially more revealing responses, such as open-ended questions, observations, and interviews of impacted populations (Wilkerson & Lang, 2007; Lang & Wilkerson, 2008). These bolder steps, however, require ratings and take time to develop rubrics and train raters. Rater effect becomes a serious challenge (Engelhard, 1994; Wolfe, 2004).

Even if time is available and rater effect is controlled through adjustments to scores, such as those provided through the Multi-Facets Rasch model, there is still a

third obstacle to overcome: respondents often see themselves differently than others see them and do not recognize the dispositions that are visible in their actions. This has been described as Johari's Window (Luft and Ingram, 1955), represented in figure 2.

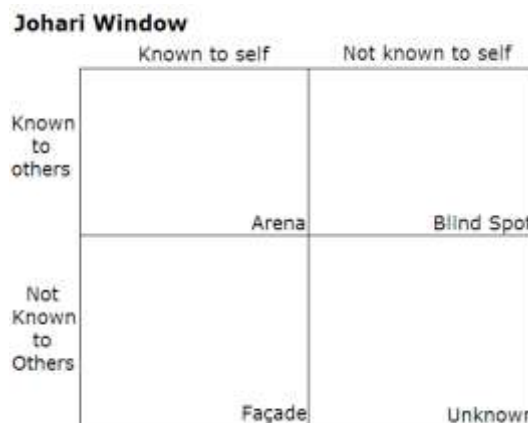


Figure 2. Johari Window. Source: Luft and Ingram, 1955

Thurstone, Likert, and Rasch

Both Likert and Thurstone scales are comprised of statements to which respondents agree or disagree. Thurstone's technique requires a dichotomous decision (agree/disagree only), while Likert provides for a rating scale, typically five-points, from strongly agree to strongly disagree with a neutral midpoint.

Thurstone suggests a labor-intensive process: clearly specifying the variable, collecting opinions and statements from many people, reducing the list to 100 items, using 300 more people to sort them, calculating item scale values, eliminating ambiguous or irrelevant items, and selecting about 20 statements graduated evenly along the scale. Likert required much less in scale development, criticizing Thurstone for the laborious process. Likert also rejected Thurstone's requirement that the results be invariant across different groups, a fundamental requirement of Rasch measurement.

Andrich (1980) compared Likert and Thurstone scaling. He discussed three aspects of the Thurstone model that are particularly relevant to Rasch theory: the importance of scale invariance for people measured, the complimentary requirement that a person's measure be independent of specific items, and the expectation of equally spaced items on a continuum. He noted that the use of the undecided category was a unidimensionality problem, since persons could respond "undecided" for various reasons - limited understanding, indifference, ignorance, or neutrality.

Andrich noted: "The [Likert] approach is popular because it is simple, it focuses directly on the attitudes of persons; empirical researchers find it satisfactory because it is theoretically undemanding" (p. 12). While Andrich concluded that

the Rasch model retains key theoretical and practical features of both, he seems to prefer Thurstone.

Drasgow, Chernyshenko, and Start (2010) presented the case for the ideal point response process over the dominance response process, concluding that the ideal point response process is better for feedback purposes. They, too, warned against faking.

Roberts, Laughlin, and Wedel (1999) examined the relationship between Likert and Thurstone agreement scaling, recommending the Thurstone scale when extreme positions (e.g., high/low levels of commitment) are of interest:

... the Likert procedure may falter for individuals who hold extreme attitudinal positions when responses result from some type of ideal point process. This is because the Likert procedure is functionally a cumulative model of the response process, and as such, it is not always compatible with responses from an ideal point process. In contrast, the Thurstone procedure is functionally an unfolding model, and thus, it does correspond to the situation in which responses follow from an ideal point process. Due to this correspondence, the Thurstone procedure does not suffer from the degraded validity exhibited with the Likert method when individuals with extreme attitudes are measured (pp. 229-230).

Research Purpose

To date, three studies of the DAATS battery have been reported (Wilkerson & Lang, 2011). The second study, conducted in 2008 (Lang & Wilkerson, 2008), measured 335 persons (predominantly pre-service) on three instruments (BATS, ETQ, and SRA), calibrating them in a single scale. A Cronbach's alpha of .96, real person separation of .90, and real item separation of .98 were reported.

Although there was connectivity in that second study, it was limited. Of the 335 respondents, 261 completed BATS, 135 completed ETQ, 46 completed half of the SRA items and 105 completed the other half. Consequently, most reporting was at the instrument level not a total measure. In this study of 190 respondents, 147 (77%) completed all three instruments, with limited missing responses within the instruments. Consequently, total scores and sub-scores on the INTASC Principles, not just the instruments, could be calculated.

Rubrics also have been substantially refined since 2008, so it was important to determine if the category structure had improved. In 2008, steps ordered appropriately, but there were problems with outfit in multiple categories (>2.0). Case studies were also used to support construct validity, but person measures by instrument limited the utility of the results.

The twofold purpose of this study was:

1. To determine if the 2008 results could be replicated and improved through enhanced scoring rubrics, a more diverse sample of students (undergraduate, master's level, and advanced graduate students), and better connectivity (higher completion rate for all three instruments).

2. To model and describe the integration of measurement and evaluation standards in the review and use of assessment instruments.

The primary significance of this study is to illustrate the potential for using a mix of a quantitative and qualitative analysis of multiple instrument types to provide rich data for identifying high and low levels of affect. High quality data can yield a rich source for improving teachers and teacher education in a critical area often under-assessed – teacher dispositions.

While this research applies specifically to teacher's measured commitment to the standards-based teaching skills, the methods illustrated apply equally to all professions requiring personnel committed to skills-based practice.

Research Questions

The overarching design question driving DAATS instrument development is: What is the measured level of commitment of teachers to the standards-based skills of teaching? Additional questions drive individual or combined instruments, e.g., Do teachers say they believe in these skills? How do teachers describe experiences and react to scenarios in ways that reveal their commitment to these skills? In this study, two specific research questions were posed:

1. What are the psychometric qualities of three DAATS instruments when combined into a single decision-making measure?
2. To what extent do measurement and evaluation standards support the use of the DAATS battery?

Method

Sample

The three DAATS instruments (BATS, ETQ, and SRA) were administered to a total of 190 students enrolled in colleges of education at two public universities in Florida. Respondents included 92 undergraduates, 49 master's level, 10 alternative certification, 19 advanced graduate (Ed.S. candidates), 3 other, and 17 with unknown student status. There were 106 females, 27 males, and 57 gender unknown; 107 Caucasian and 83 other ethnicities or unknown.

Instrument Refinement and Scoring

In this analysis, 20 items (two per Principle, all easy) were removed from BATS, based on low point-measure correlations. Previously, all items were adequate when analyzed separately. Removal of two items per Principle maintained a balanced scale (content validity) with seven items per Principle – four dichotomous and three rating scale. One minimum measure (100% correct responses) was kept, “all children can learn.” The belief scale, BATS, is dichotomous (agree/disagree); the reflection and questionnaire (SRA and ETQ) are subjective, scored using the Bloom and Krathwohl (1956) affective taxonomy.

Figure 3 provides operational definitions of the taxonomic levels. For assessment purposes, “unaware” was added.

Taxonomic Levels	Typical Teaching Behaviors at Each Taxonomic Level
Unaware	<ul style="list-style-type: none"> • Has not considered the skill in any meaningful way. • May be opposed to the skill.
Receiving	<ul style="list-style-type: none"> • Recognizes (is aware of) importance. • Is beginning to think about it. • May provide a promise to use it without evidence of having used it.
Responding	<ul style="list-style-type: none"> • Is emotionally ready to do something and makes an attempt. • Gives a little extra effort, as time permits, to comply. • Can easily be distracted from application. • Has a beginning level of commitment or satisfaction.
Valuing	<ul style="list-style-type: none"> • Accepts worth and derives definite satisfaction from it. • Feels a need and would commit continuing time and effort. • Tolerates and may expect interferences.
Organization	<ul style="list-style-type: none"> • Plans, organizes, and schedules to ensure success with it. • Determines inter-relationships among knowledge and skills. • Adapts other aspects to fit it. • Is uncomfortable with interferences or lack of time to finish.
Characterization	<ul style="list-style-type: none"> • Sees the skill as the center or driving force of all work. • Helps others to see the skill’s importance, lobbying for it. • Integrates everything with it.

Figure 3. Definition of Taxonomic Levels Adapted from Bloom and Krathwohl

Rasch Measures

Instruments were calibrated using the Andrich rating scale model (Andrich, 1988) of item response theory and Winsteps software, version 3.71 (Wright & Linacre, 1998; Linacre, 2011). In the case of dichotomous, the Andrich model is equivalent to the Rasch model. Items were combined into a single scale that included both dichotomous items (BATS) and rating scale items (ETQ and SRA). A linear transformation of the traditional mean of zero and scale of one was used, providing a mean of 50 and a scale of 10. This linear transformation avoided negative numbers and decimals, facilitating user interpretation. See Lang & Wilkerson (2008) for results of separate instrument calibrations.

Measurement-Related Results

Variable Map

The variable (or Wright) map from Winsteps is provided in Figure 4 and illustrates the distribution of person commitment (left) and item difficulty (right). At the bottom are the least committed persons and the easiest items. At the top are the most committed persons and most difficult items. The cluster of more difficult “x” items is from ETQ and SRA; “D’s” represent the easier BATS items.

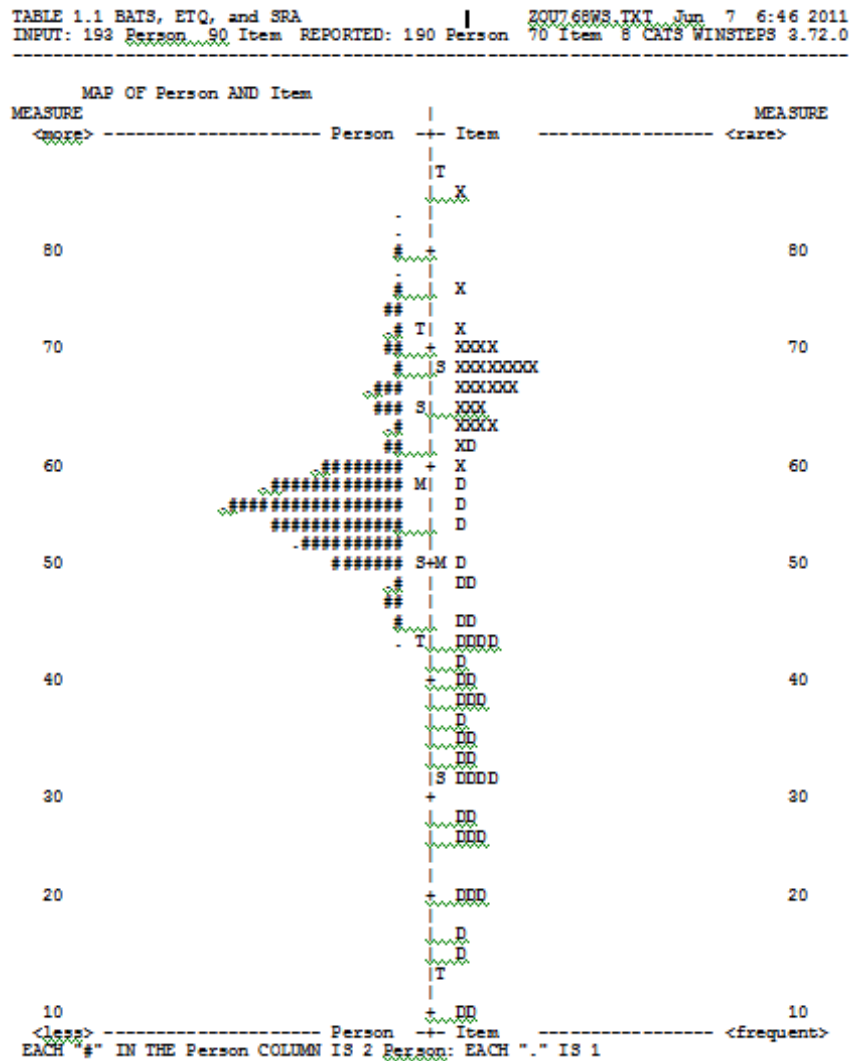


Figure 4. Variable Map of BATS, ETQ, and SRA Combined

The difficult items from SRA all pose a dilemma faced by teachers. Examples are: recognizing the confidentiality problem in publicly posted grades, accepting responsibility to work collegially with all teachers (including one who is dressed inappropriately and might be shunned), and accepting/helping all students (including one student making an impolite hand gesture). Less difficult items from ETQ required respondents to write reflectively. Examples include strategies to keep abreast of their field, teach complex concepts, and work toward improvement. Easy items from BATS tapped the utility of brainstorming in critical thinking lessons, reflecting on growth, and using multiple measures for assessment - all values taught are in class and which candidates should "know."

The sample included advanced undergraduates and graduate students, resulting in limited low person scores and many easy items. Within the

distributions of persons and items, there are relatively few gaps, indicating good construct coverage and a reasonable distribution of persons nonetheless.

Descriptive, Fit, and Reliability Statistics

Tables 1 and 2 provide descriptive, fit, and reliability statistics for items and people. Note that the mean for items was set to 50 and that the mean for persons is somewhat higher at 58. Ranges are strong: 11 to 84 for items 43-83 for persons. The standard deviation for items is almost two logits (18.4) and for people about one logit (10.9).

Table 1
Mean Statistics for Items

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Mean	216.6	173.4	50	1.96	0.99	0	1.05	0.2
S.D.	80.6	18.4	18.22	1.31	0.13	1.4	0.27	1.6
Max.	358	186	83.82	7.11	1.43	3.5	2.05	3.5
Min.	77	85	10.82	0.9	0.64	-3.7	0.53	-3.7

Table 2
Mean Statistics for Persons

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
Mean	79.6	64	57.94	2.33	0.97	-0.2	1.03	0
S.D.	24.6	10.9	7.07	0.87	0.26	1.2	0.65	1.2
Max.	145	70	82.84	7.55	1.69	3	5.18	4
Min.	18	10	43.39	1.9	0.51	-2.7	0.23	-1.9

Fit is expressed in a mean square (MNSQ) statistic – the chi-square statistic divided by its degrees of freedom. The expected value should be close to 1.0. Values >1.0 are considered underfit, introducing “noise”, or another source of variance in the data. Values <1.0 indicate over-prediction which can inflate other statistics. Infit is a t standardized information-weighted mean square statistic, more sensitive to unexpected behavior affecting responses near the person’s level; outfit is more sensitive to unexpected behavior on items far from the expected level. Fit in the range of .5-1.5 is considered productive. Fit >2.0 degrades measurement; fit of 1.5-2.0 is unproductive, fit <.5 is overly predictable and potentially misleading. Z-Standardized (ZSTD) reports the statistical significance of the MNSQ and typically should not exceed 1.96 (Linacre, 2011).

Mean fit statistics near the expected ranges of 1.0 for mean squares and .0 for standardized z’s are evident. Of the 70 items, only three exceeded the 1.5 outfit MNSQ (the highest was 2.05), and none exceeded 1.5 in infit. No items fell below .5 infit or outfit with the lowest at .64. These means, then, are not impacted by

extreme scores. Just 11 respondents (6%) had an outfit MNSQ between 2.02 and 5.18. Modelling requirements are met.

Reliability and separation statistics are acceptable, with Cronbach's alpha (KR-20) estimated at .96, typically considered high because of missing data. This is consistent with previous results. The person reliability of .87 indicates satisfactory separation of about three levels (separation = 2.67). Item reliability and separation are good at .98 and 7.63.

Tables 3 and 4 present descriptive statistics and reliability for each instrument and each INTASC Principle in the combined analysis. Model reliability for each instrument ranges from .85-.95. Instrument differences are statistically significant ($p < .01$); low BATS sub-scores can be especially revealing. Most differences between Principle means are not statistically significant.

Table 3
Results by Instrument

Item Count	Mean Measure	S.E. Mean	S.D.	Median	Model Sep.	Model Reliability	Instrument
70	49.17	2.33	19.37	50.78	5.98	0.97	All 3
40	35.24	2.21	13.8	36.24	3.17	0.91	BATS
10	64.9	0.87	2.6	64.32	2.42	0.85	ETQ
20	69.15	0.97	4.24	68.7	4.41	0.95	SRA

Table 4
Principle Means

Item Count	Mean Measure	S.E. Mean	S.D.	Median	Model Separation	Model Reliability	Principle
70	49.17	2.33	19.37	50.78	5.98	0.97	**
7	50.53	6.93	16.97	49.88	7.64	0.98	1
7	48.73	6.39	15.65	37.8	8.06	0.98	2
7	45.25	10.52	25.77	58.42	3.49	0.92	3
7	47.34	9.24	22.63	56.52	6.58	0.98	4
7	49.02	6.56	16.08	51.68	7.04	0.98	5
7	50.93	6.56	16.07	45.03	8.53	0.99	6
7	50.78	6.49	15.9	60.29	7.82	0.98	7
7	46.32	8.22	20.14	39.6	7.52	0.98	8
7	48.37	9.16	22.43	53.65	7.56	0.98	9
7	54.42	7.08	17.35	49.88	9.66	0.99	10

Principal Components Analysis of Residuals and Category Structure

The Rasch principal components analysis of residuals indicates that 50.7% of the raw variance is explained by the measures (29.4% by persons and 21.3% by items). Unexplained variance in the first contrast is 3.5%, exceeding the generally accepted 2.0% rule (Linacre, 2011); however, the variance explained is just 2.5%

(well below the 21.3% explained by the items). A small multi-dimensionality issue may exist, which is not surprising given the range of expected differences among sub-constructs (Principles).

The category structure of the rating scale items (ETQ and SRA) was appropriate. The step averages increased in order, with fit statistics near the 1.0 target, as represented in table 5. This represents a major improvement from the earlier study after rubrics were refined.

Table 5
Category Structure

Categ.	Count	%	Average	Infit MNSQ	Outfit MNSQ
0	446	9	-17.46	1.02	1.02
1	1194	25	-13.52	0.96	0.95
2	1765	37	-10.38	0.98	0.99
3	1094	23	-4.15	0.95	0.97
4	190	4	4.74	0.83	0.85
5	25	1	10.95	0.89	0.9

Figure 5 represents graphically the category structure. Curves show how probable is the observation of each category for measures relative to the item measure and resembles the expected “range of hills” (Linacre, 2011).

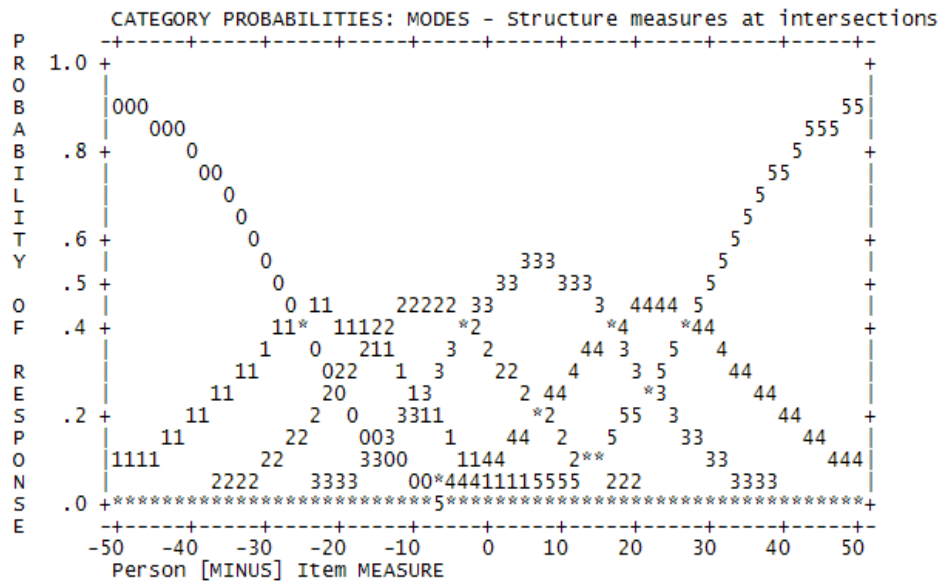


Figure 5. Category Probabilities

Validity, Reliability, and Avoidance of Bias

DAATS items were constructed to provide operational definitions of INTASC Principles, with theory driving item writing. Each Principle has seven items, ensuring domains are well represented (content validity). Judgmentally, items expected to be more difficult were more difficult; items expected to be easier were easier (construct validity). The correspondence between faculty perceptions of students and DAATS results supports construct validity (Englehart, et al., 2011; Lang & Wilkerson, 2008). Progressions from coursework to final internship and in alternative certification (Lang & Wilkerson, 2008) and progressions through degree level are evidence of predictive validity. Most preservice teachers are at the receiving or responding levels; the majority of master's level students are at the responding to valuing levels; the majority of doctoral students are at the valuing/organizing levels.

Person Principle means were relatively close to the overall mean. Collaboration (#10) was the most challenging Principle, and diverse learners (#3) easiest. Faculty confirm that candidates often prefer working alone, resisting teamwork. They are taught consistently the importance of adapting for diversity. These two principles were approximately one-half logit (and standard deviation) from the mean, evidence of construct validity. The results generally make sense. Reliability statistics are provided above. Differences in overall scores for respondents are not statistically significant between gender and ethnic categories.

Evaluation-Related Results

Utility: A Quantitative Analysis from Rasch Person Measures

The Utility Standards (U5 and U6) require relevant information for stakeholders, including data about processes and products for rediscovering, reinterpreting, or revising activities. The key question is: "Are the results useful in confirming quality and improving individual teachers and preparation programs?" The results described above on validity confirm that programs are "on track" - a relevant finding in terms of public perception and accountability. Finding students and program to target for improvement is critical to utility. Illustrative examples are presented here for teachers. Four students are represented in figures 6 (overall scores) and 7 (Principle scores).

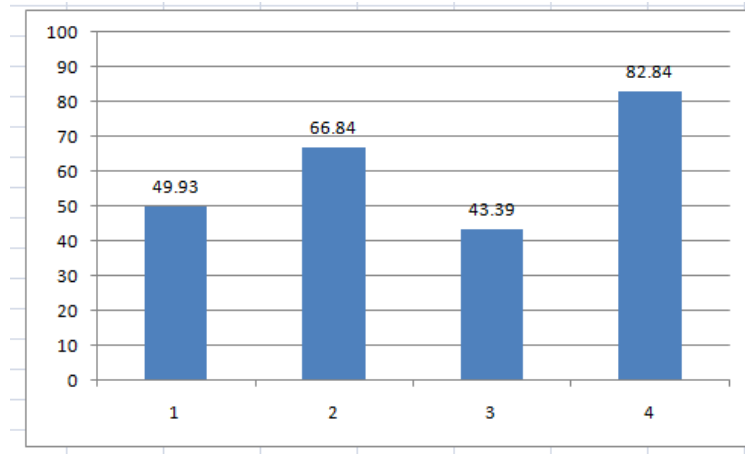


Figure 6. Total Person Measures for Four Respondents in the Sample

- Person 1 (blue): undergraduate - about 1 SD below the person mean.
- Person 2 (red): undergraduate - about 1 SD above the person mean.
- Person 3 (green): master's - about 1.5 SD below the person mean.
- Person 4 (purple): doctoral -- about 3 SD above the person mean.

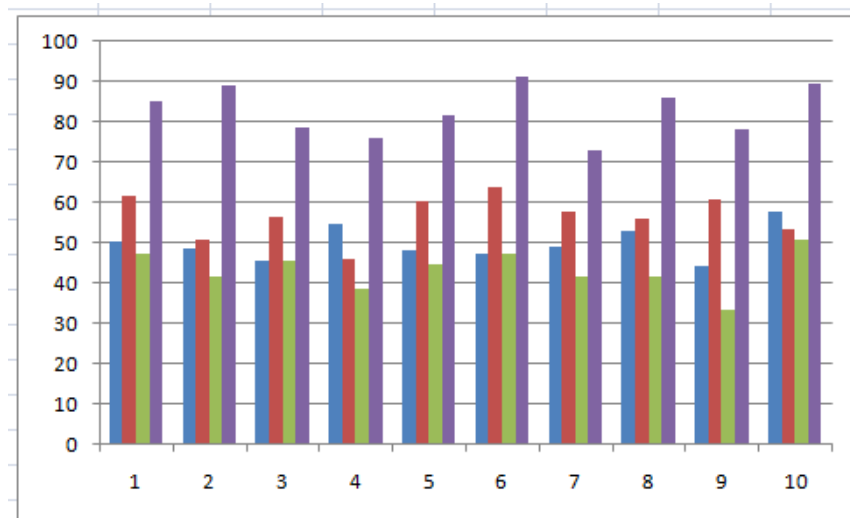


Figure 7. Principle Person Measures for Four Respondents in the Sample

Some brief (but not comprehensive) comments illustrate the process of how these data can be used for decision-making and improvement. The two undergraduate students, Persons 1 and 2 differ in their overall scores, are similar in some Principles (e.g., Principle 2 - human development), and different in others (e.g., Principle 4 - critical thinking). Person 1 could be counseled but with celebration of commitment to critical thinking. Person 2 could be celebrated in general but with the suggestion to consider commitment to critical thinking. Person 3 (master's student) shows very low commitment generally and might have limited

potential for his career path, with a particular lack of interest in continuous improvement and reflective practice (Principle 3). Person 4 (doctoral student), on the other hand, shows high commitment overall and strength in development, communication, and collegiality (Principles 2, 6, and 10); his scores bode well for his success in mentoring teachers.

Similar comparisons can be made for programs. For example, at the standards level, the finding that collaboration (Principle mean of 54) is more challenging provides an important improvement target. Conversely, pre-service teachers are relatively well committed to assessment (Principle mean of 42), a cause for celebration. At the item level, discussions related to confidentiality of grades (measure=83), embracing and helping teachers who appear unprofessional (measure=75), working with misbehaving students (measure=72), and other individual items of high difficulty could be fruitful. See Englehart (2011) for a more detailed discussion of individual and program improvement.

Utility: A Qualitative Analysis from Individual Responses

Individual responses, too, are revealing with extreme responses helping to identify teachers needing to receive help or teachers able to provide help. Figures 8-10 provide paired examples (one ETQ and two SRA) of high and low level responses on INTASC Principles 4 (instructional strategies and critical thinking), 2 (human development), and 5 (motivation and environment) respectively.

ETQ Question 4

Describe an experience in which a lesson on something difficult or complex was not going as planned. Perhaps you had asked the students to “think” about something (or used a similar instruction including “imagine”, “consider”, etc.) What did you expect to happen? What thoughts crossed your mind at that moment? Did you make any changes in your plans or interactions to help them “think”? If so, what were they? If not, why not?

I expected the student to find the answer in my situational question. Unfortunately the student came up with a solution that I did not expect. **My final thought was oh well. I can't get the students to get all the answers in two minutes.** No changes. I had tried several ways for the student to find the answer.

I constantly evaluate the way I teach. If a lesson is not going as planned I **always try various modalities** in every lesson so I can easily switch to a new method. However at the end of a lesson I **always do a plus/delta** with my students. This helps me **pinpoint** parts in the lesson that need to be reworked or parts that the students really liked.

Figure 8. Contrasting Responses for ETQ Questions 4 (Principle 4)

SRA Prompt 3



Failure: What is this child thinking and how did she get to this point? What should the teacher do?

Source: Situational Reflection Assessment (SRA) in DAATS Battery. Judy R. Wilkerson, W. Steve Lang, B. Slitkin, reprinted with permission of the authors and artist.

She is thinking that the material is too hard and she does not know how to appropriately ask for help. She needs someone to step back and teach her how to ask for help and then teach her how to do the problems. **No one ever taught her that and she has made do with throwing tantrums to get her way.**

This student looks very frustrated. **She does not realize she can do the work. The teacher needs to back up the learning until she gets to a place where the child can be successful.** The teacher needs to continue at that level of learning for a while until the student feels confident enough to take a step forward in the learning and risk doing something a little more difficult. It is important for teachers to **provide many opportunities** for perfect practice and a lot of "We Do" learning opportunities to **make sure that the student is ready to move on."**

Figure 9. Contrasting Responses for SRA Prompt 3 (Principle 2)

SRA Prompt 10



Alone: What happened to this child in the classroom? What would you say to him and to any children standing around and looking at him?

Source: Situational Reflection Assessment (SRA) in DAATS Battery. Judy R. Wilkerson, W. Steve Lang, B. Slitkin, reprinted with permission of the authors and artist.

He was reprimanded for something that he shouldn't have done. I'd tell the children to move on as he is acting in a childish way and **he needs to serve his punishment properly and that's not the way to do it.**

This child might have been **ostracized by his peers** and does not have friends in the classroom. I would have the student get up and have the class form a circle facing each other. **They will have to look at every single person and quietly think of one nice and positive thing about that person.** Then I will have them go to their desks and write each students' name on a piece of paper. Next to the students name they will then **write the one nice comment about that student.** I will collect them and **compile a list for each student and hand it out to them the following day. I would work on classroom strategies to build trust and friendship among the students.**

Figure 10. Contrasting Responses for SRA Prompt 10 (Principle 5)

Evaluation Standard U8 addresses consequences, including responsible use of data. These disturbing responses demonstrate poignantly that there are teachers who see children in a negative light. The most important question raised in this research is: “What is the impact on student learning of a teacher whose first impression of a child (or parents and colleagues or a skill) is negative?”

Feasibility

Implementing an agreement scale is relatively easy. It can be administered and scored electronically, using resources practically and efficiently (Evaluation Standards F2 and F4). ETQ and SRA are more complex and time consuming, presenting feasibility issues that are unacceptable to many potential users of this technique because of the time needed to develop rubrics, train raters, and rate responses. Rater reliability is an issue that can be controlled through the Multi-Faceted Rasch Model; however, the major feasibility issue is the time. For those committed to a balance between the cognitive and affective domains, it is feasible; for the faint of heart it is not.

Conclusions

This study established two purposes: (1) replicating and improving prior psychometric results of combining DAATS instruments and (2) modeling and describing the integration of measurement and evaluation standards in the use of assessment instruments. Some general conclusions are listed, followed by conclusions related to the two research questions:

1. The INTASC Principles provide a useful construct definition that can be measured holistically and by Principle.
2. The Thurstone agree/disagree scale contributes to the identification of strongly and weakly committed teachers.
3. The Bloom and Krathwohl affective taxonomy works in assessment, yielding proficiency levels with a credible category structure.
4. Combining affective instruments using different methods into a single Rasch scale overcomes weaknesses inherent in the instrument types.
5. A well-designed measurement device leads to useful, feasible, and accurate evaluation decisions.
6. A qualitative analysis of individual constructed response items enhances Rasch score interpretations, making them more useful for evaluation at the individual and program levels.

Question #1: Psychometric Properties of the Scale

Question #1 asked what the psychometric qualities of the DAATS instruments are when they are combined into a single measure. Using the Rasch model, 70 items (40 Thurstone agree/disagree and 30 rating scale) yielded mean fit

statistics of approximately 1.0, Cronbach's alpha estimated at .96, person reliability and separation of .87 and 2.67, and item reliability and separation of .98 and 7.63.

To ensure an appropriate sampling of the content domain and balance in the instruments, item elimination was based on both statistical and theoretical criteria, resulting in seven items assessing each of ten INTASC Principles. Means for individual Principles were near the overall mean but ordered in a logical way. Each Principle had a model reliability $>.92$ and separation >3.49 .

All categories in the rating scale sequenced correctly and yielded fit statistics between .83 and 1.02. A principal components analysis of residuals indicated that the scale accounted for 50.7% of the variance with minor indication of a dimensionality issue (3.5% unexplained variance in the first contrast).

These results indicate that the psychometric properties of the scale are adequate for valid and reliable decisions about teachers. These results are consistent with positive results from the 2008 study, while showing major improvements in the rating category structure.

Research Question #2: Combined Application of Measurement and Evaluation Standards to the Scale

Question #2 asked if measurement (AERA, APA, NCME, 1999) and evaluation standards (Yarborough, et al., 2011) supported the use of the DAATS battery, with "East meeting West." Regarding measurement Chapter 1 and evaluation standards A2 and A3, empirical and judgmental evidence of construct, content, and predictive validity was provided; the results made sense. A review of utility, particularly evaluation standards U5 and U6, demonstrated the quality and type of data that can be retrieved for individuals and programs. Results for four students were compared, with decisions regarding data use provided. Specific suggestions for counseling were identified, and the leadership potential of two contrasting graduate students was compared. Results related to programs at both the item and Principle level were also provided. These strengths and weaknesses were unknown prior to administration of the DAATS instruments. Examples were also provided to contrast low and high responses for three items, demonstrating the extent to which rich diagnostic data are available beyond the quantitative Rasch measures (Standards U5 and U6).

Regarding feasibility (evaluation standards F2 and F4), the instruments are responsive and effective, but issues of efficiency and practicality impede progress by those who are not willing to spend time scoring.

Limitations

The chief limitation in this study is the lack of respondents at a beginning (low) level. Students entering a teacher education program should be assessed to increase the variability of person measures. While previous studies on DAATS instruments included other states, these data were collected only in Florida and may not be generalizable elsewhere. The INTASC Principles were just re-written, so

instruments need revision for the 2011 update. While most Principles (now renamed *Standards*) are similar, there are also changes that will impact items.

Significance and Future Directions

Teacher dispositions are typically under-assessed. If, as professional educators, we agree that beliefs impact action, and negative beliefs about teaching and children can be harmful, then it is important to measure objectively what teachers believe about professional skills and the children they teach. Teachers who refuse to work collaboratively with colleagues may not integrate content into a cohesive curriculum from room to room (Principle 10). A teacher who sees a struggling child as “acting up” (figure 9) may devalue that child’s struggle with learning and can harm that child. A teacher who rigidly adheres to rules that do not take into account poverty and individual circumstances (figure 1), may provide little opportunity for each child to succeed and all children to learn.

The primary significance of this study was to illustrate the potential for using a mix of a quantitative and qualitative analysis of multiple instrument types to provide rich data for identifying high and low levels of affect. Why, though, is that important in this and other professions?

Knowing what teachers believe is the first step in promoting some to leadership roles while redirecting others to professional development or other jobs. Objective, high quality measurement can make this possible, if the decisions are used in meaningful ways and consequential validity is monitored. Given the difficulties inherent in affective measurement, accompanied by the general acceptance of the need for multiple measures, the use of single measures, such as surveys, should be avoided. Thurstone’s warnings made nearly a century-ago are still valid.

While this research applies specifically to teachers’ measured commitment to the standards-based skills of their profession, the methods illustrated apply equally to all professions that require personnel committed to using defined critical skills.

Regarding DAATS research, the next studies will (1) examine rater effect using the Multi-Faceted Rasch model, (2) add the dispositions checklist and child focus group (CDC and KIDS) to the scale to refine diagnostics incorporating the behavioral dimension, and (3) identify patterns of beliefs and actions of teachers measured to have both high and low commitment to the skills of teaching and to children.

References

- American Educational Research Association, American Psychological Association, and National Council of Measurement in Education (1999). *Standards for educational and psychological testing*.
- Andrich, D. (1980). Using latent trait measurement models to analyze attitudinal data: A synthesis of viewpoints. Paper presented at the 50th anniversary of the Australian Council for Educational Research (Invitational Conference

- for the Improvement of Testing in Education and Psychology, University of Melbourne, Melbourne, Victoria).
- Andrich, D. (1988). *Rasch Models for Measurement*. Newbury Park: Sage.
- Bloom, B. S., & Krathwohl, D. R. (1956). *Taxonomy of education objectives: the classification of educational goals, by a committee of college and university examiners*. New York: Longman, Green.
- Brown, A. & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance vs. ideal point controversy. *Industrial and Organizational Psychology, 3* 489-493.
- Council of Chief State School Officers (1992). *Model Standards for Beginning Teacher Licensing, Assessment, and Development: a resource for state dialogue*. Washington, D.C.: Author. Retrieved April 24, 2005, from http://www.ccsso.org/projects/Interstate_New_Teacher_Assessment_and_Support_Consortium/Publications
- Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational psychology: Perspectives on Science and Practice, 3*, 465-476.
- Edwards, A. L. (1959). Social desirability and the description of others. *Journal of Abnormal and Social Psychology, 59*, 434-436.
- Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.
- Englehart, D. S., Batchelder, H. L., Jennings, K. L., Wilkerson, J. R., Lang, W. S., Quinn, D. (2011). Teacher dispositions: Moving from assessment to improvement. *The International Journal of Educational and Psychological Assessment, 10*(1).
- Fuller, D., Fitzgerald, K., & Lee, J-S. (2008). The case for multiple measures. *INFO Brief, 52*. Washington, D.C.: Association for Supervision and Curriculum Development. Retrieved on May 31, 2011 from <http://www.ascd.org/publications/newsletters/policy-priorities/winter08/num52/full/The-Case-for-Multiple-Measures.aspx>
- Krathwohl, D., Bloom, B., & Masia, B. (1956). *Taxonomy of educational objectives. Handbook II: affective domain*. New York: McKay.
- Jung, E., Rhodes, D., & Vogt, W.P. (2006). Disposition assessment in teacher education: A framework and instrument for assessing technology disposition. *The Teacher Educator, 41*(4), 207-233.
- Lang, W. S., & Wilkerson, J. R. (2008, March). *Measuring teacher dispositions with different item structures: An application of the rasch model to a complex accreditation requirement*. Paper presented at the American Educational Research Association, NY: NY, 1-50.
- Lang, W. S., & Wilkerson, J. R. (2006, March). *Measuring teacher dispositions systematically using INTASC Principles: Building progressive measures of dispositions*. Paper presented at the American Association of Colleges of Teacher Education, San Diego, CA. 1-10.
- Linacre, J. M. (2011). A User's Guide to WINSTEPS® Ministep Rasch-Model Computer Programs: Program Manual 3.71.0. winsteps.com.

- Luft, J., & Ingham, H. (1955). The Johari window, a graphic model of interpersonal awareness, *Proceedings of the western training laboratory in group development*. Los Angeles: UCLA.
- Lund, J., Wayda, V., Woodard, R. & Buck, M. (2007). Professional dispositions: What are we teaching prospective physical education teachers? *The Physical Educator*, winter, 38-47.
- National Council for Accreditation of Teacher Education (2002). *Unit standards*. Washington, DC: retrieved March 11, 2010 at <http://www.ncate.org/institutions/standards.asp?ch=8>
- McMillan, J. H. (2007). *Classroom Assessment: Principles and Practice for Effective Standards-Based Instruction*. Boston, MA: Pearson Education, Inc.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.
- Richardson, D., & Onwuegbuzie, A. (2003). Attitudes toward dispositions related to the teaching of pre-service teachers, in-service teachers, administrators, and college/university professors. (ERIC Document Reproduction Service ED482689)
- Roberts, J. S., Laughlin, J. E., & Wedel, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, 59, 211-233.
- Saunders, W. (1994). *The program evaluation standards: A guide for evaluators and evaluation users* (1st ed.). Thousand Oaks, CA: Sage.
- Schulte, L., Edick, N., Edwards, S., & Mackiel, D. (2004). The development and validation of the Teacher Dispositions Index. *Essays in Education*, 12, Winter 2004. Retrieved April 15, 2005 from <http://www.usca.edu/essays/vol12winter2004.html>
- Singh, D. K., & Stoloff, D. L. (2008). Assessment of teacher dispositions. *College Student Journal*, 42(4), 1169-1180.
- Slitkin, B. A. (2007). Situational Reflection Assessment (SRA) art. *Dispositions Assessments Aligned with Teacher Standards (DAATS) Battery*. Unpublished instrument.
- Stake, R. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103 (2684): 677-80.
- Stufflebeam, D. (Spring, 2001). Evaluation Models. *New Directions for Evaluation*, 89, 1-106.
- Taylor, R. L., & Wasicsko, M. M. (2000). Dispositions to teach. On-line at http://www.nku.edu/~education/educator_dispositions/resources/The_Dispositions_to_Teach.pdf
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529-554.

- Wasicsko, M. M. (2004). The 20-minute hiring assessment. *The School Administrator Web Edition*. Retrieved April 24, 2005 from http://aasa.org/publications/sa/2004_10/wasicsko.htm.
- Wilkerson, J. R., & Lang, W. S. (February 2011). Standards-based teacher dispositions as a necessary and measurable construct. *The International Journal of Educational and Psychological Assessment*, 7(2), 34-54.
- Wilkerson, J. R., & Lang, W. S. (2007). *Assessing teacher dispositions: Five standards-based steps to valid measurement using the DAATS model*. Thousand Oaks: Corwin Press.
- Wilkerson, J. R., & Lang, W.S. (2006). *Dispositions Assessments Aligned with Teacher Standards (DAATS) Battery*. Unpublished instrument.
- Wilkerson, J. R., & Lang, W. S. (December 2004). A Standards-driven, task-based assessment approach for teacher licensure or certification with potential for college accreditation. *Practical Assessment, Research, and Evaluation*, 9(12). Retrieved December 11, 2010 from <http://www.pareonline.net/getvn.asp?v=9&n=12> (34 pages)
- Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46, 35-51.
- Wright, B. D., & Linacre, M. (1998). *WINSTEPS*. Chicago: MESA Press.
- Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

About the Author

Judy R. Wilkerson, Ph.D., is a Professor of Research and Assessment at Florida Gulf Coast University. She and has co-authored two books on teacher assessment in the cognitive and affective domains and presents regularly at international and U.S. professional meetings in these areas, on assessment of student learning outcomes, and accreditation and accountability. E-mail: jwilkers@fgcu.edu