IOMW XII – 2004

Measuring Teaching Ability with the Rasch Model by Scaling a Series of Product and Performance Tasks

Judy R. Wilkerson and William Steve Lang
University of South Florida
St. Petersburg, FL

wilkerso@tempest.coedu.usf.edu
wslang@tempest.coedu.usf.edu

IOMW XII – 2004

Measuring Teaching Ability with the Rasch Model by Scaling a Series of Product and Performance Tasks

Judy R. Wilkerson and William Steve Lang

ABSTRACT

Rasch measurement can provide a much needed solution to scaling teacher ability. Typically, decisions about teacher ability are based on dichotomously scored certification tests focused on knowledge of content or pedagogy. This paper presents early developmental work of a partial credit teacher ability scale of 42 tasks (performances and products) with 348 rated items or criteria. The tasks and criteria are aligned with national and state standards for expected teacher knowledge, skills, and dispositions. These tasks are being used in two-thirds of Florida school districts and are spreading to colleges of education. Over time there will be many variations in both tasks and criteria, but here we focus on the initial system.

Introduction

In the United States we are witnessing the growth of traditional standardized tests as a panacea intended to solve the teacher shortage crisis in most states by allowing states to admit untrained teachers into the profession. The most highly publicized test was developed using 3P IRT (ABCTE, 2004) and is based on a computer delivered multiple-choice assessment accompanied with video and audio presentations of item content. Comments from the ABCTE website (http://www.abcte.org/) illustrate the potential controversy on this topic:

- "American Board exams are not easy – but that is what makes them a valid indicator of teacher quality."
- "Once the field tests are completed, automated test assembly techniques will be used to construct the initial test forms that are parallel in content and statistical characteristics."
- "American Board examinations are based on rigorous standards in professional teaching and subject area knowledge. To maintain the integrity of the exam during development, these standards are available to state leaders and other certification experts to review after signing a non-disclosure agreement."

Most alternative certification programs being developed across the country provide very short training programs with little to no assessment of teacher skills. Typically, under trained teachers do a poor job for a year or two and then often leave as unsuccessful beginners (Darling-Hammond, 2003). U.S. teacher educators are, therefore, faced with serious challenges to demonstrate the quality of the graduates they prepare, and school districts are faced with the challenges of trying to become teacher trainers.

Neither the U.S. teaching profession nor the accreditors have realized the need for objective measurement to help them accomplish their goals and preserve the profession. This is largely a function of what Stiggins bemoans as assessment illiteracy (2000). They are

satisfied, at best, with ordinal scales for poorly constructed criteria on ill-defined tasks, or, at worst, with counting papers in portfolios constructed without regard to any form of psychometric consideration (Wilkerson and Lang, 2003). Florida, however, provides some hope for serving as a model.

We originally developed a set of 42 tasks designed to measure specific skills through both observation (performance) and product development. These instruments measure teacher ability on two parallel sets of standards – the Pre-professional Florida Educator Accomplished Practices (FEAP's) and the Interstate New Teacher Assessment and Support Consortium (INTASC) Principles.

The tasks were created under contract to the Florida Department of Education as the core of the State's alternative certification assessment system. A list of tasks is provided in Appendix A. Our directive was to provide for an effective measurement system for these new recruits, regardless of the background or training method provided locally for the candidate's entry into teaching. In a standards-based accountability-focused environment, we decided to use both state and national standards to help us define the construct and the Rasch model to scale it (Linacre, 2003).

We expect that over time the tasks will be expanded in many directions – new criteria, edited criteria, variations in the criteria, more tasks at the entry level, and tasks at advanced stages of teacher development. We are in the process of constructing an ability scale based on the Professional and Accomplished FEAP's that may serve as a career ladder in Florida and we are circulating an initial ruler that we have dressed up with graphic text boxes to help the uninitiated understand the potential of Rasch. Reactions are surprisingly positive to date.

Instruments

Here we are considering the initial 42 tasks developed for the system. Each task is accompanied by a rating scale and a decision-making rubric. There are currently a total of 348 items (criteria) spread among the tasks, which are judgmentally and empirically showing different levels of difficulty. We are looking at sub-scales by standard, as well. There are already variations on these instruments as redesigned for teacher preparation programs in the colleges. We expect the task bank to become very large, the initial set of tasks are the core of the system so that variations are generated as substitutes, improvements, or extensions on the original set.

Initial Evidence of Validity of the FACP Assessment System: Judgmental Approach

The assessment system uses both judgmental and empirical methods to validate these instruments. To date, four studies have been conducted that yield strong support for construct validity. We have previously summarized three in a recent on-line article in *Practical Assessment, Research, and Evaluation* (Wilkerson and Lang, 2004b), and the fourth was conducted in April, 2004 through expert panel review. In this latter review, again results were obtained to support construct and content validity. About 15 districts nominated judges to attend. All but one of the judges had direct experience in working with FACP teachers using this system. The judges reviewed the tasks to identify areas needing clarification or modification (including both the directions and the rubrics), to confirm the criticality of each task, and to confirm the decision-making structure (cut scores) for each task. Each team was

composed of three members from different districts, and they reviewed the tasks from two or three Educator Accomplished Practices sets with the following results:

- Suggestions for improvement were made in the instructions for about half of the tasks. Most of these were small wording changes such as a deletion or substitution.

- Judges confirmed the quality of the rubrics and found that for about two-thirds of the tasks, none of the criteria should be eliminated. Minimal editorial suggestions were made.

- The cut score decision-making process was unanimously supported with only a few suggestions to be harsher rather than more lenient in making the cut between "demonstrated" and "partially demonstrated".

- Reviewers were asked to identify alternative tasks to be used as substitutes for the FACP tasks. All review teams indicated that there were no alternatives. All reviewers supported keeping all tasks.

Setting and Validating Standards

The FACP system does not lend itself to standard setting methods as traditionally described. We have some parallel to Stone's (2001) Objective Standard Setting. Stone's Stage One is criterion development. We also gathered a set of professionals and judges to operationalize the standards into visible, clear descriptions of the FEAP and INTASC standards. We also concentrated on simple and objective language and asked professionals to define the criterion from some different prospective and experiences.

Stone's Stage Two relates to the concept of mastery. In our case, we conducted visualization exercises to describe the "minimally competent" or "unacceptable" teacher candidate on each described construct. A variety of discussions and audiences were involved.

We departed significantly from Stone in Stage Three. Stone creates a concept of Mastery (or professional competence to us) and obtains a judge's view into a proportion to equate to a logit. Our system started with each teacher candidate given three tries over as long as a three year period of inservice probationary teaching to complete the tasks to a level of competence. We started with three as an arbitrary number, but we asked professional assessors and mentors working with examinees in the field how many tries constitutes "too harsh", "too lenient", or "just right" a standard for the system. This study confirmed the approach.

It seems confounded to have judges estimate proportions for one, two, or three tries, particularly when we expect all candidates to complete all tasks at an acceptable level in order to continue to certification. Instead, we plan to gather evidence from assessors in the field as they use the tasks, and report empirically differences in the number of tries and task measures. At another review, judges should be able to recommend a number of tries allowed. This underscores the difficulty of standard setting with traditional measures in longitudinal performance assessment.

Initial Evidence of Validity of the FACP Assessment System: Empirical Approach

The effort in Florida depends on both judgmental and empirical analysis, and this is recommended for all institutions and districts developing an assessment system. Some references are included here for interested readers. The Rasch model of Item Response Theory has been chosen as the measurement model for this system, in part because the model is robust with regard to missing data and accommodating different item types in a test (Wright & Panchapakesan, 1969). Given the on-going nature of assessment in this system, with teachers completing the "test" over a two-year period rather than in one sitting, the robust nature of the model for missing data is extremely important for on-going diagnostic and remediation purposes. Also, the tasks are clearly using different item structures (observations, products, etc.) as measures. Another advantage of the choice of the Rasch model is the ability to detect and correct rater effects in judged assessment (Myford & Wolfe, 2003).

There are two ways to establish empirical construct validity that are useful for measures in systems such as this. One is the operationalization or functioning reality of the measures, which Trochim (2002) calls Translation Validity and consists of a blend of face validity and content validity. This approach asks the basic question of whether or not the numbers are working in different situations as expected to support the definition of the construct. Additional empirical evidence includes many descriptive analyses that use measures resulting from the tests as part of convergent and discriminant validity studies such as multitrait-multimatrix. The choice of the Rasch model for item analysis is also useful for this purpose.

Bond & Fox (2001) stated that, "In his American Psychological Association (APA) presentation, *Construct Validity: A Forgotten Concept in Psychology?,* Overton (1999) detailed the importance of Fisher's (1994) claim that the Rasch model is an instrument of construct validation." (p. 192). Fisher (2001) later describes the internal statistical analysis of a test as necessary to establish construct validity separately from content validity. Linacre (1996) describes the comparison of Rasch and the true score models for various correlational studies that would be typical of convergent and discriminant validity studies. Linacre demonstrates the advantages of the Rasch model as opposed to a true score model for applications similar to the performance system described here.

Early results using the Rasch procedure with the Florida performance tasks support empirical evidence of construct validity. Figure 1 provides the sample logistic ruler, calibrating the items from 5 of the 42 tasks in the current performance system. Even at these early calibration stages, where sparse data remain largely unconnected, it is possible to confirm that items that were expected to be more difficult are being scaled as more difficult and items expected to be easier are being scaled as less difficult. One example is demonstrated in Figure 6 from task 01A – Unit Exam. Instructors' experience indicates that teachers have less difficulty in making ESE and ESOL accommodations on tests (criterion 5, coded 01A05) than on matching test items to instructional content (criterion 1, coded 01A01). Further, the lack of gaps in the scale of items supports the adequacy of coverage of the domains, an indication of content validity interpreted as construct validity by Trochim (2002).

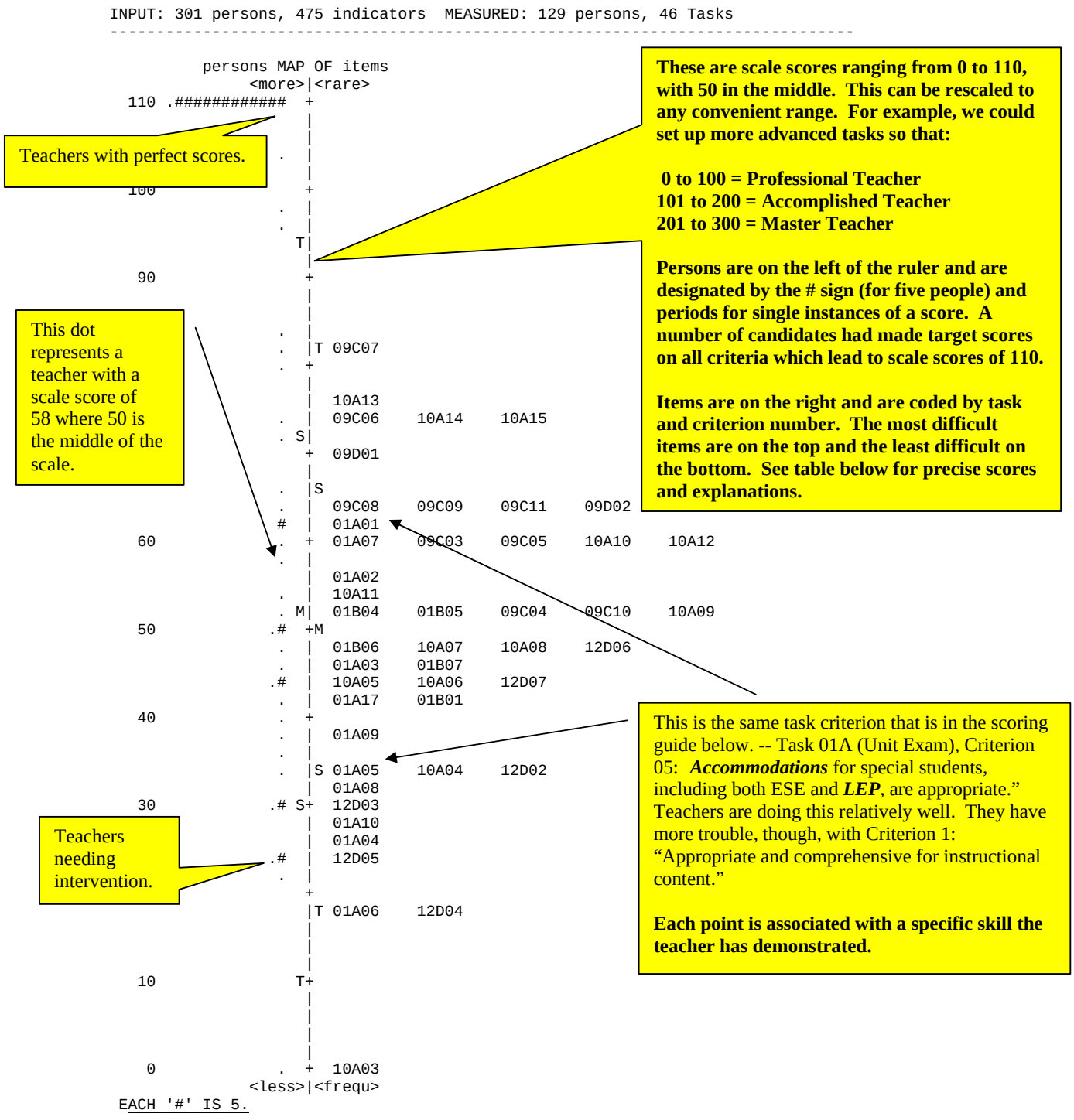
The successful calibration of items onto an interval level scale (logistic ruler) is an important step for any number of future criterion-related validity studies. The complete

discussion of construct evidence from Rasch analysis is beyond the purpose of the current article, but useful statistics would include coherence (Lopez, 1996), separation (Linacre, 1996), fit (Bohlig M., et al., 1998), discrimination (Engelhard, G., 1994), and principal components analysis (Linacre, 2003). In our example, all these statistics were consistent with predicted construct validity but are not reported here. For an alternate classical treatment of psychometric properties, even though we did not find it as useful for our application, see Ingersoll & Scannell (2002).

The ruler in Figure 1 also provides evidence of a simpler and more often overlooked component of validity: operational functionality. An assessment is only as good as the ability to report information that is practical and informative to the user. Percent correct or percentile rank results accompanied by a cut score are weak for these purposes. In the example below, even those who are not statisticians can quickly see that most teachers have mastered the tasks, but that a few are lacking. Outliers among both persons and items are readily observable. Gains on the measures, prerequisite ordering of tasks, gaps and redundancy of items, specific diagnosis of person weaknesses, and the interaction of different tasks are graphically visible. A few points are demonstrated in the callout on Figure 3 to illustrate.

Figure 1

Logistic Ruler (Scale Scores for Items and Persons) on Skill-Based Tasks as Presented to School Districts and College of Education Deans ("marketing tool"):



Data on Items:

```
INPUT: 301 persons, 475 items  MEASURED: 129 persons, 46 items, 3 CATS      3.47
--------------------------------------------------------------------------------
person: REAL SEP.: 2.24  REL.: .83 ... item: REAL SEP.: 1.94  REL.: .79
```

Analysis

.

Sample

At present, we have analysed data on 1326 teachers on 348 criteria. This changes daily as data comes into an on line tracking tool step up for the state.

Analysis

The rating scale for each item is "target" (essentially mastery), "acceptable," (minimally correct and fixable), or "unacceptable" (essentially incorrect or incomplete). We are using a partial credit model (Stone, 2003):

$$\ln\left(\frac{P_{nik}}{(1-P_{nik})}\right) = B_n - D_i - F_k$$

Results

We are had initial problems with a halo effect and the initial data are sparse, but rater training is beginning to improve judgements. There is much missing data and some lack of connectivity. A plan to use a FACETS model to analyze judge effects is being considered, but will have to await more data that identifies specific assessors within each school district.

The preliminary results are presented in Figure 1, in the format we have used to convince Floridians of the utility of the model. For presentation to general audiences, we have accompanied Figure 1 with a sample Task (Figure 2) indicating the criteria associated with the annotated ruler for five of the 42 tasks scaled. We advise them, however, that the scale is moderately precise, but still inaccurate and will remain so until more and better data are collected. We present the ruler in this format in this proposal in order to share a presentation approach that seems to be working well for us. Those with no prior exposure to IRT seem to understand in a flash what we are proposing for the State.

Following this "marketing tool" version of the ruler are summary statistics based on the complete data set starting with Table 1 (Winsteps 3.1). In this table, the initial values are moderate, but given the lack of data variation and connectivity, we expect this to continue to improve. In fact, only half the initial persons with some ratings are included in the table and some items still do not calibrate even though almost 2000 persons will eventually be rated on all items.

Figure 2

Sample Task Rubric Showing Items (Criteria) Scaled and Explained to Districts and Deans

Task 01A: Unit Exam
Scoring Rubric

Rating Scale Key: T= target; A= acceptable; U = unacceptable

Decision for A.P. 1 (Assessment) on this Task (check one):

- ❑ Demonstrated: 0-3 ratings are flawed; none are unacceptable.
- ❑ Partially Demonstrated: 4 or more ratings are flawed; none are unacceptable.
- ❑ Not Demonstrated: 1or more ratings are unacceptable.

These criteria are noted on the ruler on page 2 above.

Element	#	Criterion for "target" rating	Rating
Overall exam	1	Appropriate and comprehensive for instructional content.	__ T __ A __ U
	2	Test map is accurate and properly formatted.	__ T __ A __ U
	3	Items address knowledge/comprehension, application/analysis, and synthesis/evaluation.	__ T __ A __ U
	4	Directions are clear.	__ T __ A __ U
	5	***Accommodations*** for special students, including both ESE and ***LEP***, are appropriate.	__ T __ A __ U
Individual items	6	Items are appropriate for instructional outcomes.	__ T __ A __ U
	7	Items are appropriate for development ***and linguistic*** level of students.	__ T __ A __ U
	8	Items are written at the specified taxonomic levels.	__ T __ A __ U
	9	Items are clear, ***free from bias***, and formatted correctly.	__ T __ A __ U
	10	Key, rubric, or sample answers are correct.	__ T __ A __ U

Table 1

TABLE 3.1 MasterTasks USF PATS, ACP FL, SOUTHERN ZOU426ws.txt Jun 21 12:03 2004
INPUT: 1885 persons, 348 items MEASURED: 1326 persons, 348 items, 3 CATS   3.49

SUMMARY OF 743 MEASURED (NON-EXTREME) persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	151.7	53.2	2.40	.55	.95	-.2	.96	-.2
S.D.	194.5	66.1	1.41	.24	.40	1.9	.78	1.9
MAX.	1019.0	343.0	6.49	1.06	2.79	4.4	9.90	6.0
MIN.	5.0	3.0	-3.20	.12	.00	-9.2	.00	-9.2

REAL RMSE	.63	ADJ.SD	1.26	SEPARATION	2.01	person RELIABILITY .80
MODEL RMSE	.61	ADJ.SD	1.27	SEPARATION	2.10	person RELIABILITY .82

S.E. OF person MEAN = .05

MAXIMUM EXTREME SCORE: 583 persons
LACKING RESPONSES: 559 persons
VALID RESPONSES: 15.4%

SUMMARY OF 345 MEASURED (NON-EXTREME) items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	326.7	114.5	.00	.36	1.03	.0	1.05	.1
S.D.	127.8	45.8	.79	.17	.33	1.6	.83	1.0
MAX.	644.0	229.0	1.72	1.04	2.24	3.5	9.43	3.6
MIN.	66.0	23.0	-3.86	.13	.20	-8.4	.02	-3.7

REAL RMSE	.42	ADJ.SD	.67	SEPARATION	1.58	item RELIABILITY .71
MODEL RMSE	.40	ADJ.SD	.68	SEPARATION	1.73	item RELIABILITY .75

S.E. OF item MEAN = .04

MINIMUM EXTREME SCORE: 3 items
UMEAN=.000 USCALE=1.000

For the interest of those familiar with Rasch analysis, we reproduce the Winsteps misfit in Table 2 below. It shows eight items with an overfit (less than -2.0) or misfit (over 2.0). Given a pool of 425 calibrated items, Type I error predicts about 21 items in this range, so 23 items as misfitting is not alarming. Regardless, we have already begun to examine individual items for revision and these items are the first to receive attention.

Most importantly, we have begun the process of looking at individual Keyform results from Winsteps to see if results make sense in terms of the visualized construct and person characteristics. This should eventually add to our construct validity evidence in addition to developing face confidence in the system by users. A typical sample Keyform which we have presented is displayed in Figure 3.

In Figure 4 we have included the Expected Score ICC for the system from Winsteps for data collected to date. This is consistent with our observations that the system has some assessor difficulties with the easier tasks due primarily to rater effects. We are examining item directions, scoring rubrics, and judge training to address identified issues.

Table 2

TABLE 10.1 MasterTasks USF PATS, ACP FL, SOUTHERN ZOU426ws.txt Jun 21 12:03 2004
INPUT: 1885 persons, 348 items MEASURED: 1326 persons, 348 items, 3 CATS 3.49

person: REAL SEP.: 2.01 REL.: .80 ... item: REAL SEP.: 1.58 REL.: .71
items STATISTICS: MISFIT ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	items
273	398	134	-3.86	.53	1.92	1.6	9.43	2.1	A .13	10A04
319	264	89	-.64	.58	1.01	.2	4.88	2.7	B .02	11B15
125	362	124	-1.02	.37	2.05	2.5	4.56	2.6	C .11	05A02
316	260	89	.23	.38	.94	.0	4.52	3.6	D .15	11B12
348	67	23	-.29	.72	1.01	.3	4.04	1.9	E .00	01B04
72	343	116	-.91	.46	1.15	.5	3.93	2.2	F .02	02B01
318	266	90	-.39	.50	.98	.1	3.84	2.5	G .08	11B14
220	192	66	-.18	.46	.92	.0	3.54	2.1	H .29	08D01
174	432	147	-.53	.36	1.41	1.1	3.06	2.3	I .20	07A02
126	351	124	.03	.26	2.13	3.5	3.02	2.8	J .13	05A03
250	139	48	.21	.53	1.31	.7	2.90	1.7	K .20	09B08
218	193	68	.56	.34	1.15	.5	2.73	2.2	L .23	08C09
339	131	46	.71	.40	1.15	.5	2.66	2.1	M .09	12B01
254	138	48	.61	.42	1.35	.9	2.46	1.8	N .13	09B15
31	183	66	1.20	.30	1.46	1.5	2.43	2.3	O .28	01C07
146	340	115	-.94	.48	.81	-.3	2.41	1.4	P .25	05D04
325	279	98	.68	.27	1.32	1.2	2.35	2.5	Q .11	11D02
74	338	116	-.17	.33	.95	.0	2.25	1.7	R .14	02B03
35	195	66	-.91	.65	2.24	1.7	.33	-.2	S .27	01C11
36	195	66	-.91	.65	2.24	1.7	.33	-.2	T .27	01C12
185	224	77	.04	.42	.92	.0	2.24	1.5	U .26	07B06
221	545	188	-.47	.25	1.27	1.1	2.22	2.0	V .30	08D02
244	257	88	-.40	.41	.94	.0	2.21	1.4	W .23	09B02
295	607	210	.16	.22	1.39	1.7	2.16	2.6	X .18	10C04
184	222	77	.35	.37	1.32	.9	2.12	1.5	Y .23	07B05
182	228	79	.32	.37	1.02	.2	2.11	1.5	Z .24	07B03
208	180	66	.54	.28	.39	-3.3	1.81	1.3	.63	08B05
86	416	153	1.06	.17	1.60	3.3	1.41	1.4	.32	03A05
87	417	152	.94	.18	1.56	3.0	1.38	1.2	.32	03A06
209	176	66	.84	.27	.48	-2.8	1.52	1.0	.64	08B06
299	479	167	.53	.22	1.50	2.1	1.40	1.2	.22	11A04
5	621	228	.27	.14	1.40	2.9	1.39	1.3	.34	01A05
154	426	157	1.23	.17	1.34	2.1	1.39	1.4	.36	05D12
90	478	174	.96	.17	1.37	2.3	1.07	.4	.36	03A09
7	627	228	.14	.15	1.33	2.3	1.06	.3	.41	01A07
282	403	144	-.45	.25	.59	-2.0	1.06	.3	.63	10A13
BETTER FITTING OMITTED										
198	357	134	.47	.18	.26	-7.3	.86	-.3	.72	08A05
197	343	131	.65	.18	.31	-6.9	.76	-.7	.74	08A04
44	225	77	.02	.42	.70	-.6	.34	-.9	z .33	01D05
75	342	115	-1.43	.58	.70	-.4	.23	-.6	y .25	02B04
41	224	77	.18	.39	.68	-.8	.39	-.9	x .35	01D03
100	411	151	.74	.18	.68	-2.3	.54	-1.6	w .57	04A03
73	344	116	-1.14	.51	.67	-.6	.20	-.9	v .28	02B02
101	354	136	1.19	.17	.66	-2.6	.67	-1.3	u .62	04A04
42	215	77	1.08	.27	.65	-1.5	.42	-1.5	t .46	01D04
43	226	77	-.17	.45	.64	-.7	.23	-1.1	s .35	01D05
25	276	98	-.07	.27	.64	-1.6	.36	-1.2	r .54	01C01
260	299	103	-.86	.33	.63	-1.2	.36	-.8	q .44	09C05
45	205	77	1.67	.22	.62	-2.2	.59	-1.3	p .52	01D06
269	303	106	-.52	.29	.53	-2.1	.59	-.5	o .52	09D03
206	184	68	.64	.27	.49	-2.7	.43	-1.1	n .64	08B03
205	184	68	.64	.27	.48	-2.8	.36	-1.4	m .65	08B02
232	211	71	-1.58	.76	.47	-.7	.10	-.4	l .35	08D13
203	359	135	.50	.18	.32	-6.5	.43	-2.0	k .73	08A10
207	183	67	.51	.28	.42	-3.1	.26	-1.6	j .65	08B04
196	344	130	.54	.18	.31	-6.6	.39	-2.2	i .74	08A03
202	365	136	.40	.18	.24	-7.5	.35	-2.3	h .74	08A09
194	345	130	.50	.18	.22	-8.1	.34	-2.4	g .76	08A01
201	362	136	.50	.18	.34	-6.2	.24	-3.1	f .73	08A08
204	189	68	.23	.30	.30	-3.6	.18	-1.7	e .65	08B01
219	191	64	-2.22	1.04	.29	-.7	.02	.0	d .34	08C10
200	364	136	.43	.18	.26	-7.3	.24	-3.0	c .74	08A07
195	347	130	.44	.18	.20	-8.4	.16	-3.6	b .76	08A02
199	363	135	.38	.18	.20	-8.3	.14	-3.7	a .75	08A06

Figure 2

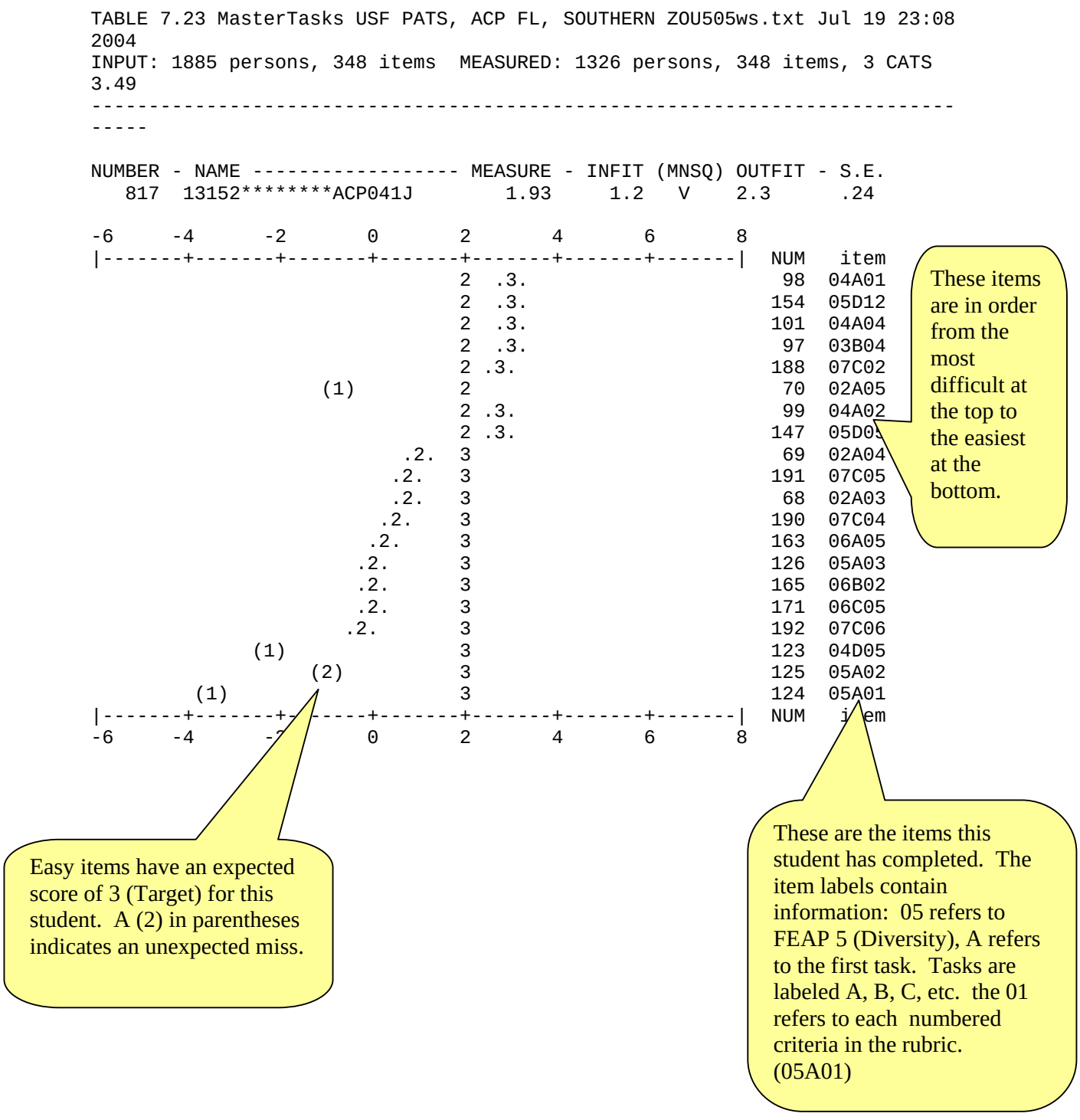
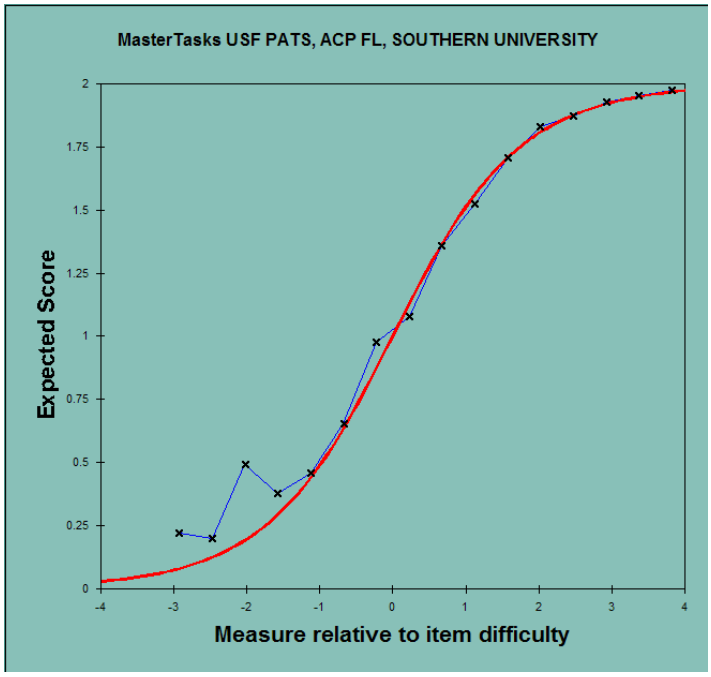


TABLE 7.23 MasterTasks USF PATS, ACP FL, SOUTHERN ZOU505ws.txt Jul 19 23:08 2004
INPUT: 1885 persons, 348 items MEASURED: 1326 persons, 348 items, 3 CATS 3.49
NUMBER - NAME ---------- MEASURE - INFIT (MNSQ) OUTFIT - S.E.
817 13152********ACP041J 1.93 1.2 V 2.3 .24
-6 -4 -2 0 2 4 6 8
NUM item
98 04A01
154 05D12
101 04A04
97 03B04
188 07C02
70 02A05
99 04A02
147 05D05
69 02A04
191 07C05
68 02A03
190 07C04
163 06A05
126 05A03
165 06B02
171 06C05
192 07C06
123 04D05
125 05A02
124 05A01
These items are in order from the most difficult at the top to the easiest at the bottom.
Easy items have an expected score of 3 (Target) for this student. A (2) in parentheses indicates an unexpected miss.
These are the items this student has completed. The item labels contain information: 05 refers to FEAP 5 (Diversity), A refers to the first task. Tasks are labeled A, B, C, etc. the 01 refers to each numbered criteria in the rubric. (05A01)

Below in Figure 3 is the Expected Score ICC which seems appropriate except at the lower ability levels.

Figure 3



We have chosen Table 3.2 in order to examine item rating structure. This fits our judgmental expectations, since we know that raters have been overly humanistic and reluctant to use the "unacceptable" point on the scale. So, we expect an "unacceptable" to be a bit lower on the scale and anticipate the difference as being due to a known rater effect. The middle rating is not as meaningful to judges as we hoped, but we anticipated this as there was considerable discussion as to the operational definition and language for this category.

```
TABLE 3.2 MasterTasks USF PATS, ACP FL, SOUTHERN  ZOU426ws.txt Jun 21 12:03 2004
INPUT: 1885 persons, 348 items  MEASURED: 1326 persons, 348 items, 3 CATS   3.49
--------------------------------------------------------------------------------


SUMMARY OF CATEGORY STRUCTURE.  Model="R"
+-------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ|| MEASURE | MEASURE|
|-------------------+------------+------------++---------+--------+
|  1   1     949   0|   .48   .22|  1.16  1.49||  NONE   |( -1.74)| 1 Unacceptable
|  2   2    3935   2|  1.67  1.80|   .93   .97||    -.41 |    .00 | 2 Flawed
|  3   3   34633  14|  3.41  3.40|   .95  1.00||     .41 |(  1.74)| 3 Target
|-------------------+------------+------------++---------+--------+
|MISSING  216818  85|  2.26      |            ||         |        |
+-------------------------------------------------------------------
AVERAGE MEASURE is mean of measures in category.

+--------------------------------------------------------------------------
|CATEGORY    STRUCTURE   |  SCORE-TO-MEASURE   | 50% CUM.| COHERENCE|ESTIM|
| LABEL    MEASURE  S.E. | AT CAT. ----ZONE----|PROBABLTY| M->C C->M|DISCR|
|-----------------------+---------------------+---------+----------+-----+
|   1      NONE          |( -1.74) -INF    -.97|         |  67%  15%|     | 1 Unacceptable
|   2        -.41    .04 |    .00   -.97    .97|    -.69 |  47%  29%|  .86| 2 Flawed
|   3         .41    .02 |(  1.74)   .97  +INF |     .69 |  91%  97%| 1.06| 3 Target
+--------------------------------------------------------------------------
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?

        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      ++--------------+--------------+--------------+--------------++
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |                                                             |
B   .8 +1111                                                    3333+
I      |    1111                                            3333     |
L      |        111                                      333         |
I      |           111                                333            |
T   .6 +              11                             33              +
Y      |                111                       333                |
    .5 +                   11                    33                  +
O      |                     111    22222    333                      |
F   .4 +                    2222**22     22**2222                    +
       |              22222      11     33      22222                |
R      |           2222             11*33             2222           |
E      |      22222               333 111                22222      |
S   .2 + 22222                 333       111                 22222 +
P      |2                   3333            1111                   2|
O      |               33333                    11111               |
N      |     3333333333                              1111111111     |
S   .0 +33333                                                  11111+
E      ++--------------+--------------+--------------+--------------++
       -2             -1              0              1              2
                      person [MINUS] item MEASURE

1 = Unacceptable
2 = Flawed
3 = Target
```

We have also analysed judgmentally the results of the initial calibration and found that the items we expected to be more difficult are more difficult and items that we expected to be less difficult are less difficult. Predictably, teachers seem to have the most trouble aggregating and using assessment data and organizing instruction that accounts for individual needs while working with large groups of children. The ten most difficult and least difficult items, along with their measures follow:

Most Difficult Items:

01C10: Portfolios are used effectively (for student self-assessment). (Measure = 1.72)

01D06: Strategies include work with families, colleagues, and possibly the community (Measure = 1.67)

01A17: Needs for improvement in the test are identified and appropriate. (Measure = 1.64)

01A14: Weaknesses in items are identified, including those associated with validity, reliability, bias, and scoring (Measure = 1.59)

07C01: The teacher has counted the number of responses correctly and classified the clusters appropriately for each student. (Measure = 1.52)

01A13: Differences for performance of subgroups is identified as needed. (Measure = 1.49)

01D14: Analysis (of assessment data in student's cumulative folder) is complete and readable. (Measure = 1.35)

04A01: There is at least one objective for each level in Bloom's taxonomy, and they are classified correctly according to Bloom's taxonomy (Measure = 1.35)

01A16: Strategies for correction of learning by individuals and the class as a whole (modifications to instruction) are identified and appropriate. (Measure = 1.33)

01B05: Reasonable levels of proficiency are defined for decision-making for each criterion (on the scoring form for the alternative assessment created). (Measure = 1.29)

Least Difficult Items:

01E03: Post-assessments (including both traditional and alternative) provide valid data on progress of students toward learning the outcomes. (Measure = -1.69)

01E01: Unit plan with outcomes and pre and post-assessments as well as copies of the assessment instruments are included in the folder. (Measure = -1.75)

03A12: The mentor has assisted in the development of the (professional development) plan, and the principal has approved it. (Measure = -1.84)

05C08: The teacher has made appropriate provision for this student in terms of time and circumstances for work, tasks assigned, communication, and response modes. (Measure = -1.96)

05D15: Accommodate or make provisions for needs of individual students --OBSERVED. (Measure = -2.08)

08C10: The sample (from one student) indicates that the teacher helped students improve their writing skills (Measure = -2.22)

10B01: The teacher adapts plans to incorporate study and test-taking skills as needed. (Measure = -2.53)

12C03: The teacher has justified all ratings (of computer software) appropriately. (Measure = -2.90)

11C01: The teacher is familiar with the Florida laws regarding abuse, including requirements for reporting abuse. (Measure = -2.99)

10A04: The units follow a logical sequence and are hierarchical where needed. (Measure = -3.86)

Conclusions

The importance of this work is in its potential to provide a major application for Rasch measurement in a statewide assessment program that is professionally aligned with standards using tasks that provide job related performance ratings that can realistically measure the construct. Initial data indicate that the teacher ability scale will provide a practical and accurate way of measuring teacher ability for teachers entering the profession regardless of preparation route. We hope eventually to use vertical and horizontal equating to scale ability for use in a career ladder with alternatives within the steps of that ladder.

This long term project also offers considerable potential for teacher effectiveness research. By placing teacher performance on a ruler, the measure of a perceived complex construct can be examined in a scientific way that has frustrated and confused many researchers as they wrestle with value added models based on complex structural equation modelling as illustrated in a recent topic issue of the *Journal of Educational and Behavioral Statistics*. As in the well-known case of Lexiles (Wright & Stone, 2004), the research process should become more clear as effective and calibrated measures are available.

Appendix A

FEAP #1 and INTASC #8: Assessment

- 01A: Unit Exam/ Semester Final Assessment
- 01B: Alternative Assessment
- 01C: Classroom Assessment System
- 01D: Case Study of a Student Needing Assistance
- 01E: Demonstration of Positive Student Outcomes

FEAP #2 and INTASC #6: Communication

- 02A: Written Communication from the Teacher
- 02B: Evaluation of Video-Taped Teaching
- 02C: Interaction between Teacher and Students

FEAP #3 and INTASC #9: Continuous Improvement

- 03A: Professional Development Plan
- 03B: School Improvement Team Involvement

FEAP #4 and INTASC #4: Critical Thinking

- 04A: Questioning Using a Taxonomy
- 04B: Lesson(s) to Teach Critical and Creative Thinking
- 04C: Portfolio of K-12 Student Work
- 04D: Critical Thinking Strategies and Materials File

FEAP #5 and INTASC #3: Diversity

- 05A: A Demographic Study of Your Students and a Plan to Meet Their Needs
- 05B: Documentation of Diversity Accommodations
- 05C: Individual Planning for Intervention
- 05D: Observation for Diversity

FEAP #6 and INTASC #9: Ethics

- 06A: Analysis of Slippery Situations
- 06B: Multiple Jeopardies and Infraction Penalties
- 06C: Potential Infractions and Teacher Responses

FEAP #7 and INTASC #2: Human Development and Learning

- 07A: Assessing Developmental Characteristics
- 07B: Assessing Learning Modalities
- 07C: Student Attitudes about School Learning

FEAP #8 and INTASC #1: Knowledge of Subject Matter

- 08A: Interdisciplinary Unit
- 08B: Portfolio of K-12 Student Work (cont.)
- 08C: Integrating Literacy Skills in Instruction
- 08D: Integrating Mathematics Skills in Instruction

FEAP #9 and INTASC #5: Learning Environment

- 09A: Classroom Management System
- 09B: Cooperative Learning Activity

09C: Case Study on Classroom Management and Motivation
09D: A Productive Classroom Environment

FEAP #10 and INTASC #7: Planning

10A: Semester/Year Curriculum Plan and Individual Unit Plan
10B: Semester Planning Record and Analysis
10C: Comprehensive Resource File

FEAP #11 and INTASC #10: Role of the Teacher

11A: Open House and Other Professional Involvement Plan
11B: Parent/Teacher/Student Conference
11C: Kids in Crisis
11D: Case Study of a Student Needing Assistance (cont.)

FEAP #12: Technology

12A: Computer-Enhanced Instructional Delivery
12B: Computer-Enhanced Management of Instruction
12C: Resource Materials from the Web

References

American Board for Certification of Teacher Excellence (2004, June). **Retrieved from http://www.abcte.org/index.html.**

Bohlig M., Fisher, W.P. Jr., Masters, & G.N., Bond, T. (1998) Content Validity and Misfitting Items. *Rasch Measurement Transactions, 12:1,* 607

Darling-Hammond, L. & Sykes, G. (2003). Wanted: A national teacher supply policy for education: the right way to meet the "highly qualified teacher" challenge. *Education Policy Analysis Archives, 11,* 33.

Engelhard, G. (1994). Resolving the attenuation paradox. *Rasch Measurement Transactions, 8:3,* 379.

Linacre, J.M. (2003) A User's Guide to Winsteps Rasch-Model Computer Programs, Chicago.
\Linacre J.M. (1996) True-Score Reliability or Rasch Statistical Validity? Rasch Measurement Transaction 9:4 p. 455-6

Ingersoll, G. M. & Scannell, D. P. (2002). *Performance-based teacher certification: creating a comprehensive unit assessment system.* Fulcrum, Golden, CO.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement.* 4:4, 386-421.

Stiggins, R. (2000). *Specifications for a Performance-Based Assessment System for Teacher Preparation.* National Council for Accreditation of Teacher Education, Washington, D.C. Retrieved June 15, 2004 from http://www.ncate.org/resources/commissioned%20papers/stiggins.pdf

Stone, G. E. (2001). Objective standard setting (or truth in advertising). *Journal of Applied Measurement, (2)2,* 187-201.

Trochim, W. M. (2002). *The Research Methods Knowledge Base, 2nd Edition*. Available online at: http://trochim.human.cornell.edu/kb/index.htm.

Wilkerson, J., Lang, W.S., Hewitt, M., Egley, R., & Stoddard, K. (2002). *Florida Alternative Certification Program Assessment System.* Bureau of Teacher Certification, Florida Department of Education, Tallahassee, FL.

Wilkerson, J.R., & Lang, W.S. (2003). Portfolios, the Pied Piper of teacher certification assessments: Legal and psychometric issues. *Education Policy Analysis Archives, 11:45*. Retrieved December 20, 2003 from http://epaa.asu.edu/epaa/v11n45/.

Wilkerson, J., Lang, W.S. (2004a). Designing Standards-Based Tasks and Scoring Instruments to Collect and Analyze Data for Decision-Making. Workshop for annual meeting of the American Association of Teacher Educators in Washington, D.C.

Wilkerson, J.R. & Lang, W.S. (2004b). A standards-driven, task-based assessment approach for teacher credentialing with potential for college accreditation. *Practical Assessment, Research & Evaluation*, 9(12). Retrieved June 20, 2004 from http://PAREonline.net/getvn.asp?v=9&n=12

Wright, B. D. & N. A. Panchapakesan (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement*, *29*, 23-48.

Wright, B. D. & Stone, M. H. (2004). *Making Measures*, Phaneron Press, Chicago.