Differential Item Functioning on a Measure of Perceptions of Preparation for Teachers, Teacher Candidates, and Program Personnel

Courtney Donovan, University of Colorado-Denver

Kathy E. Green, University of Denver

Kent Seidel, University of Colorado-Denver

Abstract

Core competencies essential for effective teaching were identified via a literature review and a review of standards for teacher education, and vetted by state groups with interests in teacher education. Survey items based on these competencies asked teacher candidates, graduates, and teacher education program faculty how well the program prepared teachers. The 41 items common to surveys of the three groups were submitted to Rasch analysis to determine dimensionality, scale use, targeting, reliability, and, of particular interest, invariance. Results suggested two dimensions were captured by the 41 items, entitled "knowledge, skills and behavior in promoting student achievement," and "resource use, academic language, and numeracy," with reliability of person separation of .94 and .73, respectively. Use of the 0-4 response scale was appropriate for both dimensions. Items were relatively easy to agree with for both scales, with person means of 1.24 and 0.57. Differential item functioning was found for respondent group and also for extent of program involvement but not for sex or for route to certification. The paper provides a discussion of implications of results for program evaluation.

Keywords: teacher effectiveness, Rasch

Teachers are the most important within-school factor in improving student achievement (Ferguson, 1991; Goe, 2007; National Research Council, 2010; Hanushek & Rivkin, 2010; Sanders, Wright, & Horn, 1997; Wenglinsky, 2002). Research supporting this finding has been made possible through improved assessments, P-12 standards, data systems, and statistical analyses such as growth and value-added modeling, as well as legislation requiring attention to formerly neglected subgroups of students. Good teachers improve student achievement, and poor teachers impact students negatively, probably for years (Sanders & Horn, 1998; Sanders & Rivers, 1996). This result propels research and policy to the next step: how do we ensure that all teachers are good teachers, and how do we support all teachers to develop the "sophisticated expertise" (Darling-Hammond & Bransford, 2005, p. 3) that defines excellent teaching?

A grant supported by the Institute for Education Sciences was proposed and received to develop assessments of preparation of teachers with linkages to the effects on K-12 student achievement. As one of the early steps in this study, surveys were created to assess perceptions of preparation from the perspectives of teacher candidates, recent graduates of teacher preparation programs, and faculty members who taught in the teacher preparation programs. The purpose of this paper is to examine whether the items common to surveys of perceptions of preparation of these three groups functioned in an equivalent manner, meaning they were invariant. This allowed us to assess whether the structure of the survey created in this project reflected a consistent variable across these three groups. As a preliminary to an analysis of differential item functioning, items were subjected to a Rasch analysis to examine dimensionality, scale use, item fit, and targeting.

The grant began with the creation of Core Competencies (CCs) or competencies considered essential for effective teaching. The survey examined here is based on the final CCs. To identify Core Competencies (CCs), documents regarding national teacher standards were examined. These included: The Interstate Teacher Assessment and Support Consortium (InTASC: Council of Chief State School Officers, 2011); The National Council for Accreditation of Teacher Education (NCATE), which is now the Council for the Accreditation of Educator Preparation (CAEP, 2016); The National Board for Professional Teaching Standards (NBPTS, 2016); The Teacher Education Accreditation Council (TEAC: 2016); and the exam elements of the Praxis II, which is a national teacher certification test. In all, 16 sources were analyzed and combined into a matrix. The teacher preparation content was selected by two criteria:

1. Policy licensure and accreditation restrictions are calling for these CCs in order to teach;
2. Programs are required to provide some evidence of how these CCs are incorporated into their program to achieve accreditation / licensure approval.

This initial mapping identified 12 potential CCs, each of which appeared in at least three of the 16 national or state sets of licensure/accreditation standards and policy recommendations. In order to focus the study, the initial 12 CCs were narrowed based on existing research and whether the CC is likely to be taught in the program (rather than being a selection criterion), is variable among programs, is observable, and is regularly employed in schools. The 12 potential CCs were grouped into 8 CCs that were considered to have less overlap, with vignettes written for each with 5-6 descriptors that would form the basis for survey items. These eight areas became: demonstrating mastery of and pedagogical expertise in content taught; managing the classroom environment; developing a safe, inclusive, respectful environment for a diverse population of students; planning and providing instruction; designing and adapting assessments, curriculum and instruction; engaging student in higher order thinking and expectation; supporting academic language development and English language acquisition; and reflection and professional growth. These CCs with their descriptors were vetted throughout the research team and through various state groups with interests in teacher education. As a result of this vetting, a *ninth* CC was added: supporting literacy and numeracy across the curriculum. This is the first time these core competencies have been constructed based on national standards and other important documents considering themes important to training effective teachers. See Hartnett-Edwards et al. (2013) for more detail on CC development. Details of these core competencies with descriptors can be found in Appendix A.

These nine CCs were reflected by 4-5 items each, with a common core of 41 items on surveys of the three groups. Surveys for each group differed slightly in wording, but 41 items were identical, with an identical response scale, across the groups.

The surveys, thus, were based on an extensive review of documents, a statewide community review process, and extended project team discussions. They were, however, surveys fielded for the first time in 2012-2013 and as such, no information was available regarding whether the CCs functioned as unique measures or whether the entire measure could be captured by one underlying dimension. Further, no information was available on whether surveys would measure similar constructs for all groups. Briggs et al. (2013)

analyzed data from two of the three surveys and concluded that different approaches to examining dimensionality yielded different conclusions about program effects.

It is vital to understand a teacher's perspective on their teacher preparation program, given the high teacher turnover rates and that one third of U.S. teachers are in their 1st-5th year of teaching (Haedden, 2014). Darling-Hammond (2006) found a relationship between teachers' perceptions of their teacher preparation program and their effectiveness as teachers. Darling-Hammond (2006) notes that while a teacher's feelings towards their preparation may not mirror their actual classroom practices, their preparation is correlated with the teacher's self-efficacy, which happens to be correlated with student achievement. The definition of outcomes in teacher education programs and the ability to measure this correlation is fundamental to aid with reform and policy in teacher education (Cochran-Smith, 2001). This work is even more vital for current educational administration who seek to support their current teachers, which begins with understanding their preparation.

The present study examined structure of the common set of 41 items for surveys from the three groups of respondents with the purpose of examining whether the items common to surveys of perceptions of preparation of these three groups functioned in an equivalent manner. This analysis provides an exploration of the constructs we created and a way to verify whether these constructs were the same across groups. Questions that directed the study were:

1. Is the measure unidimensional or are there multiple dimensions across the CC's? Are the dimensions clearly definable?
2. Is the rating scale of 0-4 consistently used?
3. What measurement gaps and redundancies exist along the subscale continuum, indicating the need for adding or deleting items?
4. Is any potential bias seen for specific items; are respondents answering differently based on groupings? Specifically, is differential item functioning found for sex, certification route, involvement with the program, and respondent group (candidate, graduate, program personnel)?
5. Is any potential bias seen for subscale scores; are respondents answering differently based on groupings? Specifically are there differences in subscale scores by sex, certification route, involvement with the program, or respondent group?

METHOD

Participants

Characteristics of three groups of participants are detailed in Table 1. Not all variables were collected for all participants, in part due to confidentiality concerns. Most candidates and graduates responding were young, white females from a traditional teacher education program. Most faculty members responding had full-time involvement with the program. Responses were received from 296 candidates, 648 graduates, and 501 program faculty members.

Table 1. *Description of the Samples*

	Candidate		Graduate		Personnel	
Variable	n	%	n	%	N	%
SEX						
Male	39	18.4%	82	18.3%		
Female	173	81.6%	366	81.7%		
AGE						
Mean (SD)	27.3 (6.95)		31.52 (9.03)			
ETHNICITY						
White	179	89.9%	383	90.1%		
Nonwhite	20	10.1%	42	9.9%		
DEGREE/ PROGRAM TYPE						
Bachelor's	112	41.3%	267	44.9%		
License only	60	22.1%	159	26.7%		
Master's	60	22.1%	70	11.8%		
Dual-Degree	39	14.4%	99	16.6%		
CERTIFICATION ROUTE						
Alternative	80	27%	192	29.6%		
Traditional	215	73%	446	68.8%		

POSITION IN PROGRAM			N	%
Full-time			159	33.4%
Part-time, regular			99	20.8%
Part-time, limited			79	16.6%
Mentor or Lead Teacher			139	29.2%

[a] *Age of participants ranged from a low of 20 to a high of 63, M = 30.33, SD = 8.84*

Instrument

The survey, as described above, was created via literature review and a comprehensive analysis of sources of standards for teacher preparation, to define eight competency areas (Hartnett-Edwards, Seidel, Whitcomb, Spurlin, Anderson, Green, & Briggs, 2013), with one additional area suggested by an advisory panel. Items were written by project personnel and vetted through teacher education program directors and a regional advisory panel. After modifications based on a series of cognitive interviews, the survey was approved by a panel of deans of colleges of education in the state.

The body of the survey for teacher candidates was split into nine sections, with each section eliciting views about an area of teaching competency. In total, the survey of teacher candidates contained 111 attitude items, 41 of which reflected overall satisfaction with the program. The survey sent to graduates was divided into the same nine competency areas. In total, the body of the graduate survey contained 90 items. Both surveys also included demographic items and items regarding teacher education program characteristics. For additional details on these two surveys, see Briggs et al. (2013).

The survey of teacher education program faculty contained 51 items. One item asked about extent of involvement with the program and the remaining items asked "OVERALL, how well does the program prepare candidates to:" where the remainder of the statement was taken from the wording for the candidate and graduate surveys. As the purpose of this study was to compare item response patterns by respondent group, only items present for all three groups were retained. This resulted in 41 items that reflected the nine CCs. Table 2 provides Cronbach's alpha values by CC by respondent group with the final number of items per CC and provides a sample item from each of the nine CCs.

Table 2. *Internal consistency reliability estimates, number of items, and sample items by group and overall by CC*

Cognitive Competency	N Items	Teacher Candidates	Graduates	Program Faculty	Overall
content mastery: The teacher is able to help students understand the interconnectedness of content areas.	5	.86	.83	.80	.84
classroom management: The teacher regularly gives learners appropriate options in learning tasks.	5	.86	.85	.82	.85
safe environment: The teacher is skilled in organizing and facilitating students' work in groups.	5	.87	.85	.83	.85
planning instruction: The teacher draws from a number of sources of information, including large-scale standardized assessments and formal and informal classroom assessments, to guide decisions about instruction.	4	.85	.82	.83	.83
adapting instruction: The teacher is able to adapt assessments, curriculum, and instruction to best accommodate students with disabilities.	5	.90	.89	.87	.89
higher order thinking: The teacher sets appropriately challenging learning expectations and communicates these effectively to all students.	5	.88	.91	.87	.90

Cognitive Competency	N Items	Teacher Candidates	Graduates	Program Faculty	Overall
academic language: The teacher uses students' first language to help clarify key concepts as needed.	4	.86	.91	.88	.89
professional development: The teacher critically reflects on his/her own identity as a teacher and cultural identity as an individual.	4	.90	.89	.84	.88
supporting literacy & numeracy: The teacher understands how to support student literacy developing in reading, writing, speaking and listening, including teaching phonics when appropriate, and teaching spelling and writing conventions.	4	.84	.90	.84	.86

Procedure

The project staff generated the online surveys, consent forms, and email instructions to access the survey. This information was sent to directors of teacher preparation programs in the state. Directors of the teacher preparation programs sent a link to the survey via email to program teacher candidates with a request to complete the survey. In addition, project staff pulled publicly available district-school emails for 897 graduates which located recent programs' graduate placements in public school posts. Directors of teacher preparation programs were also sent a link to the faculty survey with a request to convey the survey to their faculty and to mentors and lead teachers associated with the program. The surveys were open from May 2012 through November 2012. Potential participants had approximately three months to respond. Qualtrics (Qualtrics.com) was used as the online survey platform; when the survey was closed, data were downloaded as an Excel spreadsheet and transferred into a statistical software package. As the survey invitations were sent by individual program directors and not by the project staff, accurate response

rate information is not available. However, response rates of surveys of program faculty ranged from approximately 20% to close to 100% for different programs.

Analyses

The Rasch model (Rasch, 1960/1980) mandates a unidimensional construct arranged in a monotonically increasing pattern along an equal interval continuum. When data fit the Rasch model, item and person estimates are created by natural log transformations of raw data odds (Bond & Fox, 2007). Rasch modeling is the subject of an extensive literature in education and the social sciences (e.g., Bond & Fox, 2007; Fischer & Molenaar, 1995; Wright & Stone, 2004). Instruments examined via Rasch analysis enable us to determine the extent to which items serve to consistently measure a single variable from easy to difficult in a monotonically increasing fashion. Rasch models comprise a family of models applicable to dichotomous, polytomous, and continuous data. The Rasch rating scale model (Wright & Masters, 1982) was used in this study as responses were provided on a 0-4 point rating scale, with the same scale steps used for all items.

Rasch analysis allows researchers to evaluate the extent to which a unidimensional scale is created by the items in the measure. Rasch fit indices are used to determine whether each item or person contributes to the measurement of a single construct by assessing the extent to which an item or person performs as expected. That is, with adequate fit difficult items are endorsed by fewer people than are easy items. Likewise, respondents with less of the measured construct (e.g., classroom management competency) endorse fewer of the "difficult" items than respondents with more of the measured construct. Fit mean square is modeled to be 1.0 when data fit the model. Additionally, a principal components analysis of residuals is used to determine whether a second factor seems to be present in the data. Linacre (2010) suggested an instrument is likely to be unidimensional if variance explained by the first dimension is substantial, the eigenvalue for the first contrast (analogous to the eigenvalue for the second factor in an exploratory factor analysis) is less than or equal to 2.0, and the variance explained by the first contrast is less than 5%.

Item and person reliability indices estimate the replicability of item placements and person ordering. Person separation identifies the number of subgroups of persons that the instrument can discriminate. Separation and reliability of separation describe reliability in different ways (Smith, 2001). Rasch reliability indices, along with Rasch estimates of item difficulty and person ability, are based on linear measures rather than raw or ordinal data and so are more suitable for subsequent parametric calculations of means and standard deviations (Merbitz, Morris, & Grip, 1989). Separation should exceed 2.0 for an

instrument to be useful (e.g., Gauggel et al., 2004). Higher values of separation represent greater coverage of the construct along a continuum.

Finally, Rasch analysis can identify gaps in the construct continuum by identifying items and persons that are not well targeted. An item is said to be "targeted" when there is a sufficient number of persons at an ability level comparable to the item's difficulty such that the item's difficulty can be accurately estimated. A person is said to be targeted when there are items with difficulties comparable to the person's ability level. Where items and persons are not well targeted, they have larger error estimates. These gaps provide feedback on how well the instrument is actually measuring what it is supposed to measure within given ranges of the measure and also what might be done to further improve it.

RESULTS

Research Question 1: Is the measure unidimensional or are there multiple dimensions across the nine CC's? Are the dimensions clearly defined?

Dimensionality, Overall Fit, and Separation. Coherent item groupings were identified by using item fit statistics and principal components analysis of residuals. Data from all three respondent groups were combined for all analyses. Initial analysis with all 41 items indicated the measure was potentially multidimensional (1st contrast eigenvalue = 2.7, indicating the possibility of more than one dimension in the data). We identified potential subscales by selecting groupings of items that underfit the Rasch model and then refining item sets. Briefly, items that underfit (infit or outfit mean squares >1.30) were deleted individually until no further items evidenced underfit. The remaining items formed the first dimension. All items that underfit were then analyzed separately to see if a coherent second dimension was feasible. In this manner, two dimensions were empirically identified. The first dimension comprised 33 items and the initial 8 underfitting items comprised the second dimension.

Table 3 shows the number of items, overall fit values, dimensionality, separation, reliability of person separation, and person mean for these two subscales. The two subscales were entitled "knowledge, skills and behavior in promoting student achievement (skills)" and "resource use, academic language, and numeracy (resource use)." No third subscale emerged as all items were used in scales 1 or 2. Mean square fit (infit and outfit) have expected values of 1.0 if the data fit the model. Infit is weighted by the distance between item and person location while outfit is an unweighted index. Both are transformations of

chi-square statistics. For both samples and both scales, infit and outfit mean squares were close to 1.0, indicating adequate overall fit of data to the model.

Table 3. *Dimensionality, Item Fit, and Separation*

Index	Scale 1	Scale 2
Number of items	33	8
Overall Mean MNSQ Infit	1.00	.99
Variance to Measure	49.1	47.0
Eigenvalue of the First Contrast	2.6	2.0
Real Person Separation (non-extreme cases)	3.98	1.66
Real Reliability of Person Separation	.94	.73
Real Item Separation	7.56	9.35
Real Reliability of Item Separation	.98	.99
Cronbach's Alpha	.99	.95
Person Mean	1.24	.57

Item fit to the scales yielded mean square infit values of less than 1.23 for all items. For scale 1, item mean square infit values ranged from .73 to 1.22; for scale 2 from .86 to 1.22.

Research Question 2: Is the rating scale of 0-4 consistently used across the three groups and does it appear to be appropriate?

Scale Use. Figure 1 provides an example of the use of the rating scale for Subscale 1: skills. Table 4 provides category use, observed average, and step structure values by category for both subscales. There were no category inversions. Scale use was as intended and although scale category 1 was consistently the least used, there were sufficient observations to provide reasonable estimates of fit and step structure. Scale use reflected a less-to-more interpretation of the rating scale. The patterns were similar for Subscale 2 which is not displayed.

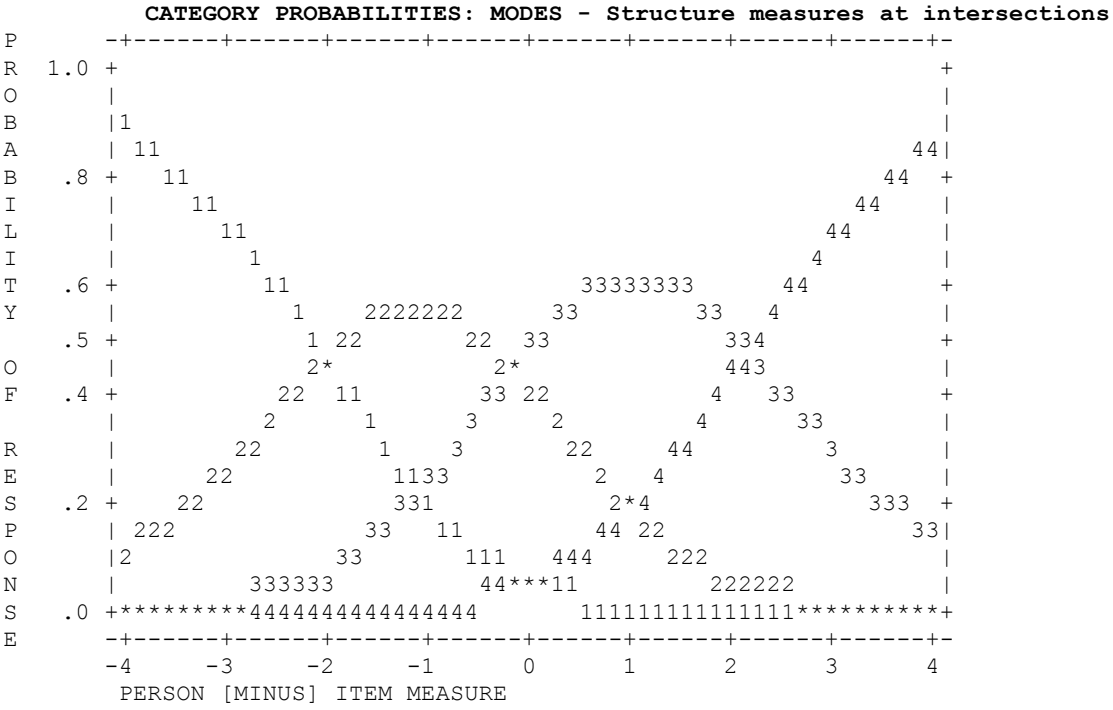

Figure 1. Rating scale use: The curves show how probable each category is to observe relative to the item measure expressed as the difference between item and person logit position. Probability of Response is the likelihood of endorsing a given rating scale category at that level of difference in person-item well-being. Intersection of adjacent rating scale categories can be seen at estimated threshold value of the higher of the two categories. For example, the threshold for category 2, or the point at which category 2 becomes a more probable response than category 1 is -2.0 logits.

Table 4. *Rating Scale Use*

Category	Observed Percentage	Observed Average	Infit MNSQ	Step Structure
Scale 1				
1	5%	-1.45	1.09	(-3.21)
2	19%	0.02	0.97	-1.12
3	42%	1.16	0.95	1.05
4	34%	2.64	1.01	-3.34
Scale 2				
1	10%	1.3	1.02	(-2.79)
2	22%	-0.18	0.97	-0.95
3	39%	0.76	0.98	0.85
4	28%	1.81	1.01	-2.97

Research Question 3: What measurement gaps and redundancies exist along the subscale continuum, indicating the need for adding or deleting items?

Targeting and Construct Coverage. Figures 2 and 3 display the item-person maps for Subscale 1 (skills) and Subscale 2 (resource use). This map provides the side-by-side positioning of persons and items with category responses to items indicated. Figure 1 shows items to be somewhat easy to agree with for this sample, and there were some persons whose position on the trait was not adjacent to any response category to any item at the lower and upper scale extremes. The person mean for Subscale 1 was 1.24. Targeting of items for Subscale 2 (Figure 3) shows good coverage of person positions, with a person mean of .57. For Subscale 1, there were numerous items at one position, indicating items may be redundant. For Subscale 2, items were more dispersed. If these subscales were to be revised, some items at similar positions might be deleted and replaced with either very easy or more difficult items to extend construct coverage.

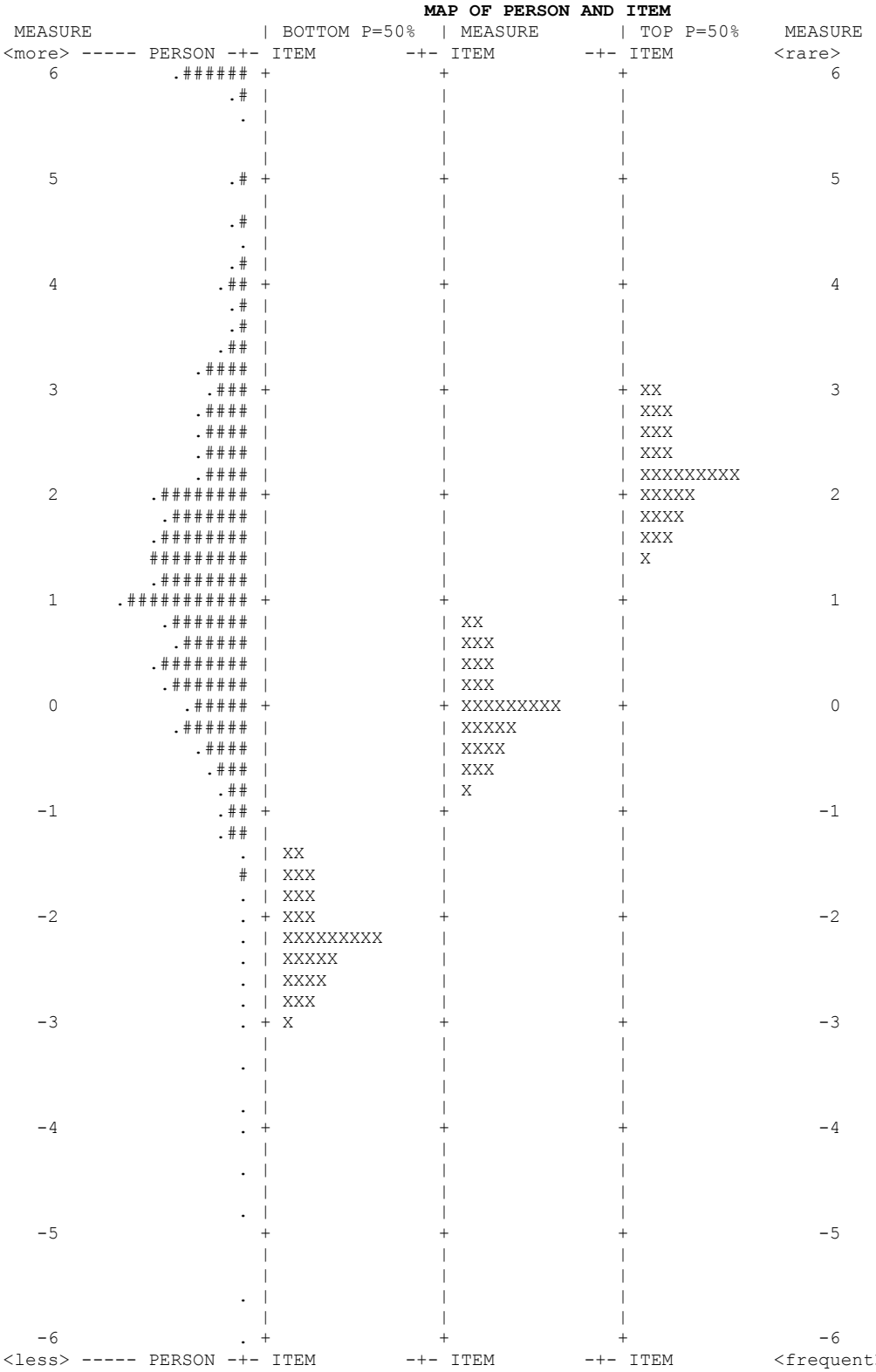

Figure 2. Map of person and items for Scale 1. Each "#" in the person column is 7 persons: Each "." is 1-6 persons; "X" indicates position of an item at the lowest, mean, and highest rating position.

```
                                    MAP OF PERSON AND ITEM
 MEASURE               | BOTTOM P=50%  | MEASURE       | TOP P=50%    MEASURE
<more> ----- PERSON -+- ITEM        -+- ITEM        -+- ITEM         <rare>
    4          .####### +              +               +                  4
                    .# |              |               |
                     . |              |               |
                     # |              |               |
                       |              |               |
                       |              |               |
                       |              |               |
    3              .## +              +               +                  3
                     . |              |               | X
                    .# |              |               |
                   ### |              |               |
                       |              |               |
                     . |              |               |
             .######## |              |               | X
    2                . +              +               + X                2
                     # |              |               |
             .######## |              |               | XX
                     . |              |               | X
             .######## |              |               | X
                   .## |              |               | X
              .####### |              |               |
    1                . +              + X             +                  1
          .########## |              |               |
                     . |              |               |
         ############# |              |               |
                .###### |              |               |
                  .#### |              | X             |
                .###### |              | X             |
    0             .#### +              +               +                  0
                .###### |              | XX            |
             .######### |              | X             |
                     . |              | X             |
                .###### |              | X             |
                    .## | X            |               |
                    .## |              |               |
   -1               .## +              +               +                 -1
                     .# |              |               |
                     .# |              |               |
                      . | X            |               |
                     .# | X            |               |
                      . |              |               |
                        | XX           |               |
   -2               .# + X            +               +                 -2
                        | X            |               |
                        | X            |               |
                      . |              |               |
                      . |              |               |
                        |              |               |
                      # |              |               |
   -3                 . +              +               +                 -3
                        |              |               |
                      . |              |               |
                        |              |               |
                      . |              |               |
                        |              |               |
                        |              |               |
   -4                 # +              +               +                 -4
<less> ----- PERSON -+- ITEM        -+- ITEM        -+- ITEM         <frequent>
 EACH "#" IN THE PERSON COLUMN IS 8 PERSON: EACH "." IS 1 TO 7
------------------------------------------------------------------------------
```

Figure 3. Map of person and items for Scale 2. Each "#" in the person column is 8 persons: Each "." is 176 persons; "X" indicates position of an item at the lowest, mean, and highest rating position.

Table 5. *Differential item functioning by position in program, respondent group, certification route, and program involvement*

POSITION IN PROGRAM

Item	Full-time Faculty	Part-time, but Regular	Part-time, involved in limited courses	Mentor or lead teacher, limited	Difference in Logit Position	*p*
2B: effective time use	-.59		.25		-.85	.002
3E: collaborates with larger community	1.17			.61	.56	.001
4B: knowledge of development research	.23	-.41			.63	.001
4C: uses variety of instructional activities	-.77			-.06	-.71	.002
7A: helps develop academic language	.63	-.16			.79	.001
7A: helps develop academic language	.63			.05	.58	.006
7C: feedback on use of academic language	.75			.12	.63	.003

RESPONDENT GROUP

Item	Candidates	Graduates	Program Personnel	Difference in Logit Position	*p*
2C: organizes to work in groups	.12		-.48	.60	.0001
2C: organizes to work in groups		.20	-.48	.68	.0001

Item	Candidates	Graduates	Program Personnel	Difference in Logit Position	*p*
3A: mutually respectful relationships	-.65		-1.20	.55	.001
3A: mutually respectful relationships		-.68	-1.20	.51	.0001

7B: practice academic language	.31	-.20		.51	.001
8B: strengths and weaknesses of assessment tasks	-.44		.32	-.76	.0001
8D: reflects on interactions with community	-.85		-.33	-.52	.001
9A: literacy development	.24	-.45		.70	.001
9A: literacy development	.24		-.45	.69	.001

CERTIFICATION ROUTE

Item	Traditional	Alternative	Difference in Logit Position	*p*
2A—routines and rules for classroom	-.92	-.40	-.52	.001

INVOLVEMENT WITH THE TEACHER EDUCATION PROGRAM

Item	Full-time Faculty	Part-time, but Regular	Part-time, involved in limited courses	Mentor or lead teacher, limited	Difference in Logit Position	*p*
2A--routines and rules for classroom	-1.24		-.29		-.95	.0001
2A--routines and rules for classroom	-1.24			-.64	-.60	.006

RESPONDENT GROUP

Item	Candidates	Graduates	Program Personnel	Difference in Logit Position	*p*
7D—language in context	.85	-.09		.94	.0001
7D—language in context	.85		-.07	.78	.0001
9C—numeracy development	-.59		-.05	-.54	.002

Research Question 4. Is any potential bias seen for specific items; are respondents answering differently based on groupings? Specifically, is differential item functioning (DIF) found for sex, certification route, involvement with the program, and respondent group (candidate, graduate, program personnel)?

Invariance. Invariance of item positions was assessed for four variables: sex, certification route (traditional or alternative), program involvement of faculty (with four categories), and group (candidate, graduates, program personnel). Table 5 provides logit positions for items with differential functioning by group and subscale. DIF was considered substantial if the Welch's t-test for difference in logit positions between groups was statistically significant ($p < .01$) *and* if the difference in logit position exceeded .50.

Scale 1. No DIF was found for sex. Six items evidenced DIF for the variable of regular involvement with the teacher education program, with seven differences found. Full-time faculty perceived items concerned with academic language development, knowledge of the research about human development, and community collaboration as more difficult to agree that preparation was good than mentors/lead teachers, and perceived items about effective use of instructional time and the variety of instructional activities easier to agree with than part-time faculty or lead teachers. Five items evidenced DIF by respondent group, with nine differences. Differences were most pronounced in item position between program faculty and teacher candidates, with program faculty overall tending to view items as easier to agree with than other groups.

Scale 2. No DIF was found for sex. One item evidenced DIF for certification route, with those from a traditional program responding most positively to the item regarding their preparation to set up routines and rules for the classroom. The same item also evidenced DIF for program involvement, with full-time faculty perceiving preparation to set up routines and rules for the classroom as easier to agree with. Finally, two items evidenced DIF by respondent group, with candidates finding setting language objectives for use of the English language as more difficult to agree the program prepared them for than either graduates of program faculty. Candidates agreed more than program faculty that they were prepared to promote student numeracy development.

Research Question 5. Is any potential bias seen for subscale scores; are respondents answering differently based on groupings. Specifically, are there differences in subscale scores by sex, certification route, involvement with the program, or respondent group?

Relationships with Background Variables. Table 6 provides descriptive information about the distribution of logit person scores for Scale 1 and Scale 2; both were relatively normally distributed. One- and two-way analyses of variance were conducted to assess effects of variables on scale 1 and 2 logit person scores. Statistically significant differences were found between respondent groups for scale 1, $F(2,1442) = 27.94$, $p < .001$, $\eta^2 = .04$. Using the Games-Howell *post hoc* test, differences were found at $p < .01$ between candidates (mean = 1.53) and graduates (mean = .75) and between graduates and program faculty (mean = 1.41). Statistically significant differences were found for Scale 2 as well, $F(2,1442) = 25.82$, $p < .001$, $\eta^2 = .04$. Using the Games-Howell *post hoc* test, differences were again found at $p < .01$ between candidates (mean = .80) and graduates (mean = .06) and between graduates and program personnel (mean = .87).

Table 6. *Description of the distribution of Scales 1 and 2*

Index	Scale 1	Scale 2
Mean	1.14	.60
Median	.77	.27
Standard Deviation	1.83	1.49
Skewness	.69	.62
Kurtosis	2.49	2.11

No significant main effect was found for sex or the interaction of sex with certification route in a 2x2 ANOVA for either Scale 1 or Scale 2. However, a significant main effect of certification route was found for Scale 1, $F(1, 656) = 6.61$, $\eta 2 = .01$, with a higher mean logit position for alternative (mean = 1.65, sd = 2.23) than for traditional (mean = 1.16, sd = 1.98).

No statistically significant effect on person logit position mean was found for level of program involvement for scale 1, $F(3, 472) = 1.41$, $p = .24$, or for scale 2, $F(3, 472) = .92$, $p = .43$.

DISCUSSION

The survey was created from an extensive literature review and content expert reviews of documents pertaining to teacher standards that guide teacher preparation programs. This yielded eight themes which we named "core competencies" (CC) that are essential for effective teaching; a ninth CC was added regarding numeracy. The survey was created based on these nine CC's with 4-5 questions for each CC. The purpose of this study was to explore the construct of the survey and verify consistency in its use across three groups: teachers, teacher candidates, and university program personnel. The survey demonstrates multidimensionality; two factors were found named Skills and Resource Use. The final survey was fairly consistent across groups, but some important differences and variances were found across the three groups.

This sample found the items on both scales easy to agree with, with most giving a rating of 3 or 4. Both scales had good person coverage, which shows variation in how persons fell along the item scale, meaning these are good scales that cover a large range of person responses. On the other hand, item coverage was not well spread and had several redundancies, especially for the Skills scale. Items falling at the same position could be revised in order to spread the items apart. Items that extend the scale in a positive direction would be very beneficial as the scale is not covering this part of the sample as well as hoped.

Invariance was tested for sex, certification route, and involvement with the program. Both scales showed no differential item functioning (DIF) for sex or certification route; all groups within these variables responded to items in a generally similar manner. This finding was somewhat surprising and adds to the mixed literature around certification routes (Sass, 2011). DIF was found for program involvement of faculty. Full-time faculty members of the program had a harder time agreeing that preparation was good than mentor/lead teachers in the field. This is interesting as both groups are preparing teachers within the program, but with very different roles and insights. Several studies refer to a disconnect between classroom learning and field experiences, which appears to be present here with faculty not just the teacher candidates (Darling-Hammond, 2009; Zeichner, 2013). This could be explored further considering coursework experiences versus student teaching experiences.

There were also differences in item position by respondent group. This analysis was the focal point of this study. Candidates found 'setting language objectives for use of the English language' harder to agree with than graduates or program personnel. Graduates and program faculty have more experiences with setting these objectives than candidates who have not taught their first year yet, so this is potentially an experience issue. Additionally, candidates found 'promote student numeracy development' easier to agree

with than program faculty. This may also be an issue of experience level, but it is interesting that candidates across the state agreed more about feeling prepared for math goals than language goals. This is particularly interesting for current educational leadership as it can help them aid their teachers in professional development opportunities. In general, items showed relatively little DIF across respondent groups, with some exceptions as noted.

Differences for demographic variables on subscale scores were found for both scales. Graduates of the program were more negative towards their program preparation on both scales than both the candidates (had not yet graduated) and program faculty. Graduates are teachers who were in their 1st-5th year of teaching, so this is possibly due to the influences of real-life teaching. Teacher candidates may feel that their program preparation was sufficient, but it is hard to actually know until they are in a full-time teaching setting. No differences were found for sex or for program involvement. There was a statistically significant difference found for certification route (traditional versus alternative), with teachers who went through an alternative program rating their preparation higher on the Skills scale but not the Resource Use scale. This is interesting as teachers in alternative programs are put right into the classroom and learn along the way, while traditional programs focus on learning first and then classroom experiences.

It was not surprising that the final scale was multidimensional, but somewhat interesting in that there were only two factors, not the nine expected CC's. Each of the nine CC's were validated separately through factor analysis (Briggs et.al, 2013) and extensive expert and document reviews in the creation process. These CC's create a useful framework for understanding what teacher candidates should know and be able to do. This study combined all the survey items for all CC's and found two overarching factors. When considering the items that fell into each dimension, the two factors were named Skills and Resource Use. This shows that while there were nine overarching ideas for effective teaching to occur, demonstrated by this sample, it really comes down to whether or not the teacher has the skills needed and can use resources appropriately and creatively.

Based on these analyses, teacher preparation programs and even professional development personnel need to evaluate their current programming to consider what aspects are Skills-related and what are Resource Use-related. This in no way means that we ignore the nine core competencies (Appendix A), but this adds a new way to evaluate teacher development programs. Asking which skills a teacher needs and what aspects of their program teaches students how to use resources creatively and effectively could improve the program and may lead to more effective teachers.

REFERENCES

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd Ed). Mahwah, NJ: Lawrence Erlbaum Associates.

Briggs, D. C., Circi, R., Seidel, K., & Green, K. (2013). *Challenges in measuring core competencies in teacher preparation programs*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, May 1, 2013.

Cochran-Smith, M. (2001). Constructing outcomes in teacher education: Policy, practice and pitfalls. *Education Policy Analysis Archives, 9*. http://epaa.asu.edu/ojs/article/view/340/466

Council for the Accreditation of Educator Preparation (CAEP). (2016). Accessed from www.caepnet.org/standards/introduction, May 27, 2016.

Council of Chief State School Officers. (2011, April). Interstate Teacher Assessment and Support Consortium (InTASC) model core teaching standards: A resource for state dialogue. Washington, DC: Author.

Darling-Hammond, L. (2006). Assessing teaching education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57,* 120-138.

Darling-Hammond, L. (2009, February). *Teacher education and the American future.* Charles W. Hunt lecture. Presented at the annual meeting of the American Association of Colleges for Teacher Education, Chicago.

Darling-Hammond, L., & Bransford, J (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do.* San Francisco, CA: Jossey-Bass.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28*(2), 465-498.

Fischer, G. H. & Molenaar, I. W. (Eds.) (1995). *Rasch models. Foundations, recent developments, and applications.* New York, NY: Springer.

Gauggel, S., Böcker, M., Heinemann, A. W., Lämmler, G., Borchelt, M., & Steinhagen-Thiessen, E. (2004). Patient–staff agreement on Barthel Index scores at admission and discharge in a sample of elderly stroke patients. *Rehabilitation Psychology, 49*, 21–27.

Goe, L. (2007, October). *The link between teacher quality and student outcomes: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality.

Haedden, S (2014). *Beginners in the classroom: What the changing demographics of teaching mean for schools, students, and society.* New York, NY: Carnegie Corporation.

Hanushek, E. A., & Rivkin, S. G. (2010, May). *Using value added measures of teacher quality. Brief 9*. Urban Institute, National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Hartnett-Edwards, K., Seidel, K., Spurlin, M., Anderson, S., Green, K., & Briggs, D. (2013). *An exploration of novice teacher core competencies: Relationship to student achievement and depth of preparation.* Presented at AACTE Annual Meeting, Orlando, FL, February 2013.

Linacre, J. M. (2010). A user's guide to winsteps ministep 3.70.0: Rasch-model computer programs. Chicago, IL: Winsteps.

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70*(4), 308.

National Board for Professional Teaching Standards (NBPTS). (2016). Accessed from http://www.nbpts.org/national-board-standards.

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, IL: The University of Chicago Press.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Research progress report. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Sass, T. R. (2011, December). *Certification requirements and teacher quality: A comparison of alternative routes to teaching.* National Center for Analysis of Longitudinal Data in Education Reform, working paper 64.

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2,* 281–311.

Teacher Education Accreditation Council (TEAC). (2016). Accessed from http://www.teac.org/accreditation/updates-to-the-guide.

Wenglinsky, H. (2002). The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10*, 1-30.

Wright, B. D., & Masters, G. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (2004). *Best test design.* Chicago, IL: MESA Press.

Zeichner, K. (2013). Rethinking the connections between campus courses and field experiences in college- and university-based teacher education. *Journal of Teacher Education, 61,* 89-99.

APPENDIX A

Core Competencies (CC) with Descriptors

CC 1. Demonstrating mastery of and pedagogical expertise in content taught:

a) The teacher deeply understands the content that s/he teaches. This include knowledge of central concepts, tools of inquiry, and specialized character of the discipline being taught.
b) The teacher understands typical ways that students' progress in learning content, as well as common misunderstandings and how to uncover and address these in teaching, and instructional practices important to the discipline being taught.
c) The teacher's understanding for both content and learners enables him/her to draw on students' real world interests and experiences to makes learning relevant for all students, and to connect students' background and contextual knowledge with new materials being taught.
d) The teaching is able to help students understand the interconnectedness of content areas.
e) The teacher works with library, media, and other resource specialists to integrate information/technology literacy skills into curriculum and instruction.

CC 2: Managing the classroom environment:

a) The teacher sets up routines and rules for the classroom that helps students work together and focus on learning. S/he is proactive in managing behavior, using appropriate interventions when needed.
b) The teacher uses time effectively, plans for learning experiences so that time is not lost in transitions and gives targeted support to students who need extra help.
c) The teacher organizes the classroom learning environment so that students can easily work in groups of varying size, see display boards and other full-group materials, and access learning materials when needed.
d) The teacher regularly gives learners appropriate options in learning tasks.
e) The teacher integrates and uses technology to maximize student learning, and appropriately supplements textbooks and other standard curriculum materials to add to the classroom learning experience.

CC 3: Developing a safe, inclusive, respectful environment for a diverse population of students:

a) The teacher maintains a classroom build on mutually respectful relationships with students and among students. This includes strategies to help students from different cultures interact positively with each other.
b) The teacher is skilled in organizing and facilitating students' work in groups.
c) The teacher maintains a classroom environment that promotes social development and group responsibility.
d) Cultural inclusiveness is supported through structured classroom talk, curricula, and instructional experiences which connect learning to students' lives and interests within and outside of school.
e) The teacher works collaboratively with families and significant adults in the lives of their students to foster healthy relationships among students, parents, and the larger community.

CC 4. Planning and providing instruction:

a) The teacher draws from a number of sources of information, including large-scale standardized assessments and formal and informal classroom assessments, to guide decisions about instruction.
b) The teacher has knowledge of current research about how students' social, emotional, physical, and cognitive developments influence learning, and current research on effective practices.
c) The teacher uses a variety of instructional activities that guide students to not only summarize or recall information, but to also apply, synthesize, interpret, and/or evaluate materials in order to deepen understanding.
d) The teacher effectively incorporates homework and projects; their completion, grades and feedback provide students with increased learning time and the teacher with a tool for monitoring students' progress over time.

CC 5: Designing and adapting assessments, curriculum and instruction:

a) The teacher is able to adapt assessments; curriculum and instruction to best accommodate individual differences among students.
b) The teacher is able to adapt assessments; curriculum and instruction to best accommodate students with disabilities.
c) The teacher provides appropriate social/emotional, academic, and other supports to reach challenging and/or seemingly unmotivated students. S/he

acknowledges and builds on any emotional responses to the content as opportunities to support learning.

d) The teacher is able to adapt assessments, curriculum, and instruction to best accommodate students with disabilities who are from culturally and/or linguistically diverse backgrounds.

e) The teacher provides proactive, clear and constructive feedback to families about student progress and work.

CC 6: Engaging student in higher order thinking and expectation:

a) The teacher sets appropriately challenging learning expectations and communicates these effectively to all students.

b) The teacher models and encourages students to reflect on and assess their own learning, asking them to explain, "how they know what they know" or "how they solved a problem of task."

c) The teacher encourages students to engage with challenging material. The teacher works with students to help them understand the importance of the work and to assess their own ability to be successful.

d) The teacher pays careful attention to all students' learning so that s/he can give feedback to students to guide their learning. The feedback given has important properties: it is descriptive, specific, relevant, timely, and constructive. It enables students to guide their own work and thereby increase their active involvement.

CC 7: Supporting academic language development and English language acquisition:

a) The teacher helps all students develop academic language by appropriately modeling language and conventions typical for the content area/discipline, providing explicit instruction in language and ways of expression that are used in the discipline.

b) The teacher provides opportunities for students to practice academic language of content areas in listening, speaking, reading, and writing.

c) The teacher's feedback for students includes a focus on improving their appropriate use of academic and other language in learning tasks and assessments.

d) The teacher sets specific language objectives for instruction, and provides opportunities for use of English language in the context of learning new content.

e) The teacher uses students' first language to help clarify key concepts as needed.

CC 8: Reflection and professional growth:

a) The teacher uses multiple formal and informal sources of evidence about what students know and can do in order to evaluate and critically reflect on the impact of his/her teaching.
b) The teacher is aware of the strengths and weaknesses of his/her assessment tasks.
c) The teacher critically reflects on his/her own identity as a teacher and cultural identity as an individual.
d) The teacher works to reflect on and improve his/her interactions and relationships with students, other educators, and families and community.

CC9: Supporting literacy and numeracy across the curriculum

a) Teachers understand how to support student literacy development in reading, writing, speaking and listening, including teaching phonics when appropriate, and teaching spelling and writing conventions.
b) Teachers use instructional strategies to develop students reading comprehension of different genres and texts, including teaching students to write in a variety of genres, and help foster students oral (speaking and listening) and written responses to literature.
c) Teachers demonstrate knowledge of mathematics and understand how to promote student development in numbers and operations, algebra, geometry and measurement, and data analysis and probability, including teaching mathematical problem-solving processes.
d) The teacher helps students make connections among mathematics/numeracy and other subjects, as well as teaching connections among mathematical ideas within math subjects (e.g., connections among geometry, algebra, and trigonometry).