MASSACHUSETTS DEPARTMENT OF ELEMENTARY AND SECONDARY EDUCATION

Educator Preparation Surveys: Technical Report

May, 2017

Massachusetts Department of Elementary and Secondary Education

75 Pleasant Street, Malden, MA 02148-4906

Phone 781-338-3000

www.doe.mass.edu

MASSACHUSETTS DEPARTMENT OF ELEMENTARY AND SECONDARY EDUCATION

This document was prepared by the
Massachusetts Department of Elementary and Secondary Education
Mitchell D. Chester, Ed.D.
Commissioner

Office of Planning and Research
Massachusetts Department of Elementary and Secondary Education
75 Pleasant Street Malden, MA 02148-4906
Phone 781-338-3625 TTY: N.E.T. Relay 800-439-2370
http://www.doe.mass.edu/

Content

Executive Summary i

1. Introduction 1
2. Data Sources and Survey Administration 1
   - 2.1. Survey Stakeholder Profiles and Administration 1
   - 2.2. Survey Response Rates 2
   - 2.3. Survey Item Specifications 3
3. Item and Measure Development and Validity Framework 4
   - 3.1. Item and Measure Development 4
     - 3.1.1. Program Approval Items 4
     - 3.1.2. Performance Standards for Teachers Items 4
   - 3.2. Validity Framework 5
4. Data Analyses Procedures 5
   - 4.1. Classical Theory Methodology 5
   - 4.2. Rasch Methodology 5
5. Validity Evidence for Program Approval Criteria 6
   - 5.1. Reliability 6
   - 5.2. Descriptive Statistics 6
6. Validity Evidence for Performance Standards of Teaching Scales 7
   - 6.1. Content Validity 7
     - 6.1.1. Item Review 7
     - 6.1.2. Item Technical Quality 7
   - 6.2 Substantive Validity 8
     - 6.2.1. Rating Scale Functioning 8
     - 6.2.1. Item Hierarchy 8
   - 6.3. Generalizability 9
     - 6.3.1. Reliability Evidence 9
     - 6.3.2. Differential Item Functioning 10
     - 6.3.3. Item Invariance across surveys 11
   - 6.4. Structural Validity 12
     - 6.4.1. Rasch Dimensionality Analyses 12
     - 6.4.2. Confirmatory Factor Analyses 14
   - 6.5. External Validity 17
     - 6.5.1. Survey responsiveness 18
     - 6.5.2. Non-Extreme Person and Preparation Provider Score Distributions 18
     - 6.5.3. Measuring Group Differences 20
7. Conclusion 21
   - 7.1. Program Approval Criteria 21
   - 7.2. Performance Standards of Teaching Instruments (42 item) 21
   - 7.3. Hiring Principal PST Instruments (6 item) 21
   - 7.4. Limitations and Recommendations 22
     - 7.4.1. Study Limitations 22
     - 7.4.2. Survey and Item-level Recommendations 22

References 24

Appendices 26

Executive Summary

Introduction

As outlined in 603 CMR 7.03 (5), The Massachusetts Department of Elementary and Secondary Education (ESE) is required to administer and publicly report survey data about the quality of educator preparation in the state. ESE administered four educator preparation stakeholder surveys in the spring of 2016 to evaluate the perceptions of teacher readiness in the Commonwealth. Considering the perceptions of key stakeholders is critical to a Sponsoring Organization's (SO) continuous improvement. The purpose of this report is to provide reliability and validity evidence to support the use of ESE's teacher preparation surveys for various purposes, including programs' continuous improvement, public transparency, and as one source of evidence in the evaluation of teacher preparation providers' programs[1].

Data Sources and Survey Administration

ESE asked teacher candidates, teacher completers, Supervising Practitioners and hiring principals to respond to the teacher readiness surveys[2]:

1. **Teacher Candidate survey,** issued at the point of program completion
2. **Teacher Completer survey,** issued to individuals employed in a Massachusetts public school one year after program completion
3. **Supervising Practitioner survey,** issued to individuals who served as a supervisor to a candidate during the practicum experience
4. **Hiring Principal survey,** issued one year after program completion to principals who hired a teacher completer

The survey is bifurcated for candidates and completers that were Teacher of Record[3] before or during their preparation program experience. The profiles of the key stakeholders administered for each survey are shown in Table 1.

Table I: Profiles of Stakeholder Groups

Survey Acronym	Stakeholder Group	Teacher-of-Record during Program	Respondent Group Defined	Survey Administration Defined
NTCD	New Teacher Candidate	No	Graduate of Sponsoring Organization with no prior teaching experience	Self-report 42 PST items
TRCD	Teacher-of-Record Candidate	Yes	Graduate of Sponsoring Organization with prior teaching experience	Self-report 42 PST items
NTCP	New Teacher Completer	No	Employed in MA public school	Self-report 42 PST items
TRCP	Teacher-of-Record Completer	Yes	Employed in MA public school	Self-report 42 PST items
SPCD[1]	Supervising Practitioner	NA	Practicum supervisor of new teacher and teacher-of-record candidates	Perception 42 PST items
NTHP	New Teacher Hiring Principal	No	Principals who hired completers (teachers new to teaching) before the start of 2015-2016 school year	Perception 6 PST items
TRHP	Teacher-of-Record Hiring Principal	Yes	Principals who hired completers before the start of 2015-2016 school year	Perception 6 PST items

[1] For more information on the development process of the stakeholder surveys, please see the Development & Validation memo
[2] ESE prioritized the development and administration of surveys associated with *initial teacher licensure* as this is the largest group of program completers in Massachusetts ever year (approximately 4,500 completers annually).
[3] In 603 CMR 7.02, Teacher of Record is defined as one or more teachers who are assigned primary responsibility for a student's learning in a subject, grade, or course.

The surveys had two types of items:

1. Items that aligned to observable outcomes relative to the Program Approval Standards and Review Criteria (henceforth called program approval items), which set forth expectation for providers. Topic areas assessed were respondents' views on their overall experience with their program, coursework, field-based experiences, practicum supervision and their views on their performance assessment.
2. Items that aligned to observable practices within the Professional Standards for Teachers (henceforth called PST items), which define the pedagogical and other professional knowledge and skills required of all teachers. For most stakeholder groups (Table 1), 42 items were used to operationalize the **teacher readiness construct** and measure stakeholder perceptions of readiness in the four standards of effective practice (hiring principals responded to only 6 PST items):
    a. Standard I: Curriculum, Planning and Assessment (13 items),
    b. Standard II: Teaching All Students (18 items),
    c. Standard III: Family and Community Engagement (6 items),
    d. Standard IV: Professional Culture (5 items).

Particular weight was placed on items that correspond to one of the six essential elements identified through the Candidate Assessment for Performance (CAP). The emphasis of this report is to provide reliability and validity evidence for the use of scores from the PST items, but does include more limited reliability and validity evidence for the program approval items.

Data Analyses Procedures

ESE used Wolfe and Smith's (2007a, 2007b) adaptation of Messick's validity framework (1980, 1995) for construct validity to guide the evidence collected to evaluate the **content** (do the PST items represent and measure the teacher readiness construct?), **substantive** (do the respondents use the response options (rating scale) as the instrument developers intended them to?), **generalizability** (are the PST items measuring the construct reliably and consistently at the construct level and at the sub-scale level?), **structural** (do the PST items align with the four standards of effective practice, the hypothesized internal structure of the teacher readiness construct?) and **external** (is the instrument able to measure change in respondents' perceptions (i.e., is it responsive?)), and if teacher readiness construct scores concur with scores from instruments measuring related constructs' (e.g., evaluation scores) **validity aspects** of the teacher readiness construct. ESE used the Rasch model (Rasch, 1960) predominantly to provide validity evidence for the five validity aspects of the teacher readiness construct. ESE paid particular attention to how well the results replicated across the respondent groups; a high level of replicability provides strong evidence for the reliability and validity of the items measuring the teacher readiness construct.

Results: Program Approval Criteria Items

The indices used to measure program approval criteria are reliable and support the use of indices and item scores in assessing stakeholder views of their preparation programs.

Validity evidence supporting the use of Program Approval Criteria indices

The data presented in this section relate to teacher candidate and teacher completer perceptions of the program approval items. The analyses are restricted to assessing the reliability of the program approval criteria.

Methodology: Reliability was measured using classical test theory. Cronbach alphas are reported.

Major Findings: The Cronbach's alphas of the program approval items for each topic area were all near or above 0.8 (Table II). Cronbach Alpha values above 0.9 are desired for high stakes tests with values greater than 0.7 considered acceptable when test scores are used in relatively low stakes tests or decisions.

Conclusion: Given some topic areas are composed of relatively few items, the reliability of these topic areas are sufficient to make inferences on stakeholder perceptions of their preparation and readiness to teach. The survey items adequately measure the respondents' perceptions of overall program experience, coursework, field-based experience, supervision, and assessment. Replication of these sub-scale reliabilities across three respondent groups further supports the conclusion that the program approval items will provide consistent measures of respondents' perceptions of each topic area.

Read More: Validity Evidence for Program Approval Criteria

Table II: Cronbach's alpha of topic areas

Cronbach Alpha (# of Items)	New Teacher Candidate (NTCD)	New Teacher Completer (NTCP)	Teacher of Record Candidate (TRCD)
Overall Program Experience (9)	0.90	0.91	0.92
Course Work (3)	0.88	0.84	0.92
Field-base Experience (5)	0.73	0.79	0.79
Supervision (5)	0.83	0.82	0.90
Assessment (4)	0.89	0.92	0.91

[1]Due to a survey administration error, data is not available for the Teacher-of-Record Completer responses

Results: Performance Standards of Teaching Scales (PSTs)

The Rasch construct validity framework (Wolfe & Smith, 2007a, 2007b) established validity evidence to justify the use of the teacher readiness PST scores for each of the seven stakeholder groups. Overall, the magnitude of the scale and subscale reliabilities of the five 42 item surveys support the generalizability of score meaning designed to measure the teacher effectiveness construct. The stability and replicability of reliabilities and validity evidence across different respondent groups provides strong confirming evidence that the items are generalizable and representative of the teacher readiness construct. The validity evidence for the six item hiring principal surveys is mixed. Overall, there is sufficient validity evidence to support the continued use of the hiring principal surveys, although improvement to the Teacher-of-Record hiring principal (TRHP) survey, in particular, is warranted.

Validity evidence supporting the use of PST scale and sub-scale scores

Content and Substantive Validity

The data in this section relates to content and substantive validity of the PST items. Content validity examines the "content relevance, representativeness and technical quality" (Messick, 1995, p.745) of the PST items used as indicators of the teacher readiness construct. Substantive validity assesses whether the responses to the PST items are consistent with the theoretical framework used to develop the items.

Methodology: Item technical quality was assessed using the point-to-measure (PTM) correlations and item fit statistics. PTM specifies how well any one item relates (correlates) to the other items of the survey. If the items are purported to

measure one construct, you would expect a positive correlation between items, preferably a correlation of above 0.3. Item-fit statistics provide evidence of how well the observed response data fit the expectations of the Rasch model; if the observed data are perfect fit "expected" by the model, the mean square error statistic will equal 1.0. Typical criteria for retaining items in an instrument are mean squares that range from 0.7 to 1.3.

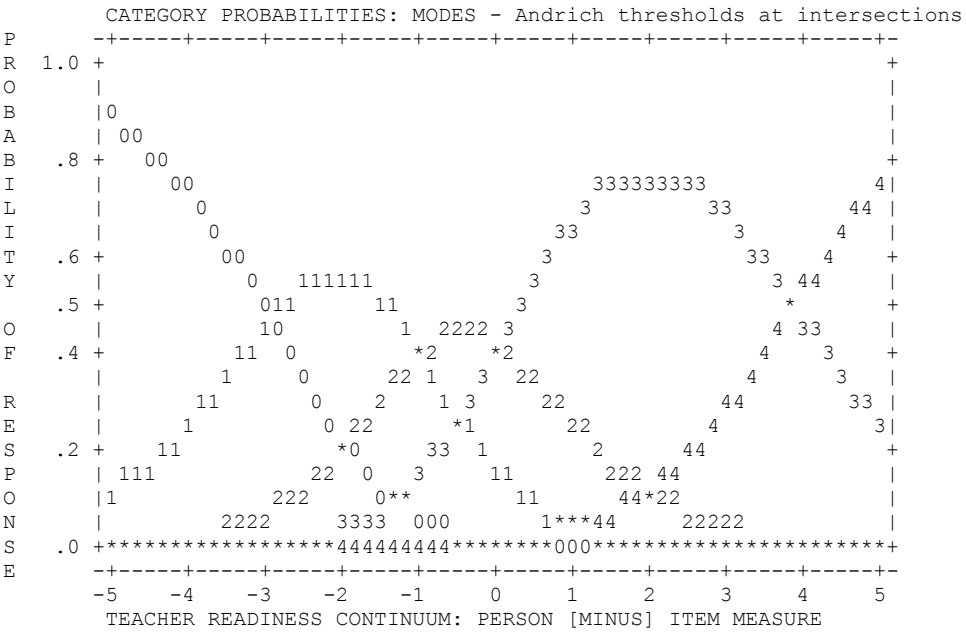
In Rasch, the expected performance of a well-functioning item is that the ordered thresholds (deltas) are monotonic; that is, increasing levels of the latent trait (teacher readiness) are associated with endorsement of more affirmative categories (e.g., strongly agree). In addition, a qualitative assessment of the hypothesized item hierarchies is used to determine if the hierarchies follow the developer's *a priori* expectations.

Major Findings: The PTM correlations were all positive and predominantly between 0.60 and 0.75 on each of the surveys. With minor exceptions, the item fit statistics were between 0.7 and 1.3, indicating that the items on each of the surveys were well-fitting. Across the seven surveys, the findings indicated that the respondents were using the rating scales as intended with category thresholds increasing monotonically with more affirmative thresholds (Figure 1, TRCD rating scale). Similarly, the item difficulty hierarchies distributed from low to high along the teacher readiness continuum according to the instrument developer's *a priori* theory across all seven surveys.

Conclusion: These findings and the replicability across surveys provide strong evidence to support the content and substantive validity aspects of the teacher readiness construct.

Read More: Validity Evidence for Performance Standards of Teaching Scales (PSTs)

Figure 1: Teacher-of-Record Candidate (TRCD) Rating Scale Function



Generalizability Evidence

This section examines the generalizability of the teacher readiness construct and its sub-scales. Reliability assesses whether the scores from a respondent(s) would be replicable if the same respondent(s) were tested again under similar conditions; that is, reliability looks at the stability or reproducibility of survey scores. Scores used in high stakes tests/decisions should be measured with low levels of measurement error or, in other words, with reliabilities of above 0.9. For scores associated with relatively low stakes decisions/tests, a reliability of 0.7 is considered minimally

acceptable. Similarly, a measure is considered generalizable when the score meaning and properties function similarly (are invariant) across multiple contexts (e.g. stake holder groups, survey forms, items) or time points.

Methodology: Reliability is measured using Cronbach alpha or the Rasch-based Person Separation Reliability. Differential item function analyses are used to assess whether the meaning and interpretability of items remain invariant across different contexts (e.g., survey forms, respondent subgroups). ESE used attenuated and disattenuated (corrected for measurement error) Pearson product moment correlations to assess the linear relationship (item invariance) between item deltas of the PST items.

Major Findings: The reliability of the construct across all surveys is above 0.9 (Table III), indicating that the level of measurement error is very low, thereby supporting the generalizability and use of the teacher readiness scores. At the standard level, the reliability of items of Standard IV was near or above 0.7 across all surveys. With the exception of TRCP survey, the reliabilities of Standard III items were all above 0.85; item reliabilities for Standard I and Standard II items were near or greater than 0.9.

With very few exceptions across all seven surveys, the 42 items (or six-item hiring principal surveys) did not exhibit DIF when respondents attending public institutions were compared to those attending private institutions. In addition, the Pearson correlations of the item deltas across the five surveys (NTCD, TRCD, NTCP, TRCP, and SP) provide strong evidence of the invariance of the 42 items across survey forms. Disattenuated correlations were all greater than 0.9 for each paired comparison.

Conclusion: Overall, the replicable reliability data support the generalizability of the PST measure. With the caveat noted for Standard IV items, the reliability data support providing program providers with scaled scores for each standard.

Read More: Generalizability

Table III: Descriptive Data and Real Person Separation Reliabilities (PSR) for Teacher Preparation Surveys

Non-Extreme Person Report	New Teacher Candidate (NTCD)	Teacher-of-Record Candidate (TRCD)	New Teacher Completer (NTCP)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)	New Teacher Hiring Principal (NTHP)	Teacher-of-Record Hiring Principal (TRHP)
Number of Respondents[1]	370 (408)	167 (179)	185 (202)	167 (186)	625 (649)	571 (628)	275 (449)
Number of Items	42	42	42	42	42	6	6
Real PSR[2]	0.95	0.97	0.96	0.96	0.96	0.94	0.80

[1]Number in parentheses includes persons with extreme measures, [2] Real Person Separation Reliability (lower bound)

Structural Validity Evidence

This section evaluates the alignment of the scoring structure to the hypothesized structure of the construct. The hypothesized structure of the teacher readiness construct is that it is represented by items that measure the four Professional Standards for Teachers (PSTs); items representing each standard form a sub-scale. Unidimensionality is a fundamental assumption of the Rasch model; the instrument items should only measure one latent trait (teacher readiness construct). This does not preclude that within the 42 items used to measure the construct that certain items

will be more correlated with some items than others. the Rasch model extracts the common variance across all respondents' response patterns to all 42 items.

Methodology: ESE used dimensionality statistics provided by the Winsteps software program to assess the structural validity of the construct. ESE also performed Confirmatory Factor Analysis (CFA) on the data from three of five surveys with sufficient data (Table IV). In CFA, a model for the internal structure of the construct is hypothesized (expected) and the observed data is tested to determine if it fits the expected model (four correlated standards (factors) of practice with the construct, teacher readiness, explaining the relationship between the four factors).

Major Findings: An analysis of the Rasch sub-scale (standards) measure correlations and the standardized residuals of the 42 item Rasch model confirmed that the items in each survey were measuring a unidimensional construct with teacher readiness explaining the relationship among and between the items. With the exception of one respondent group, the percent variance explained in stakeholder responses was over 50% (considered good for perception surveys).

A second-order factor model provided the best-fitting model in all three surveys; the data are shown in Table IV. The Root Mean Square Error of Approximation, Comparative Fit Index (CFI), Tucker Lewis Index (TLI) and factor loadings all met the criteria of a well-fitting model. This model was also significantly better fitting when compared to a four, three or one first-order factor model, respectively.

Conclusion: The Rasch and CFI provide corroborating evidence that that teacher readiness is the higher-order factor (construct) that explains the covariation among and between the items of the four standards. These results replicate across the three surveys further supporting this validity argument.

Read More: Structural Validity Evidence

Table IV: Model Fit Statistics for Second-Order Confirmatory Factor Model

Number of Items = 42	New Teacher Candidate NTCD (N = 407)	New Teacher Completer NTCP (N = 202)	Supervising Practitioner SPCD (N = 649)
RMSEA[1]	0.054*	0.065*	0.066*
RMSEA (90% CI)[2]	0.051 – 0.57	0.060 – 0.070	0.064 – 0.069
CFI[3]	0.97	0.97	0.96
TLI[4]	0.97	0.97	0.96
Factor Loadings above 0.5	YES	YES	YES

[1]RMSEA: Root Mean Square Error of Approximation point estimate; [2]RMSEA 90% Confidence Interval; [3]CFI: Comparative Fit Index; [4]TLI: Tucker-Lewis Index; *$p < 0.0005$.

The results of both the Rasch analyses and the CFA provide confirmatory evidence that the teacher readiness construct is composed of four factors (standards) with one higher-order factor (teacher readiness) explaining the relationship between these factors.

External Validity Evidence

This section examines evidence to support the external validity aspect of construct validity. External validity relates to the responsiveness of an instrument and the relationship of its scores to the scores of related or un-related external

construct measures. If an instrument is responsive, then it can be applied appropriately to measure expected group differences or individual change.

Methodology: ESE used the person strata index, H, to assess the responsiveness of the instruments. It provides the number of statistically distinct ability or endorsement groups whose centers of score distributions are separated by at least three standard errors of measurement within the sample. Unfortunately, there was insufficient data available to assess the relationship of the scores from each survey with external measures (e.g., completer evaluation scores).

Major Findings: The teacher readiness scales for six of the seven surveys are responsive and capable of measuring change on the variable and of assessing group differences. The number of statistically distinct groups (person strata) for each survey was equal to or above five (Table V). The exception was for the teacher-of-record hiring principal survey (TRHP) which could only discriminate three statistically distinct groups; the variance of the items was poor for this survey.

Conclusion: The results in this section provide limited but positive evidence for the external validity aspect of the teacher readiness construct.

Read More: External Validity

Table V: Responsiveness of Teacher Preparation Surveys

Non-Extreme Person Report	New Teacher Candidate (NTCD)	Teacher-of- Record Candidate (TRCD)	New Teacher Completer (NTCP)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)	New Teacher Hiring Principal (NTHP)	Teacher-of-Record Hiring Principal (TRHP)
Number of Respondents[1]	370 (408)	167 (179)	185 (202)	167 (186)	625 (649)	571 (628)	275 (449)
Number of Items	42	42	42	42	42	6	6
Variance Explained	45.5%	60.3%	51.9%	49.9%	51.7%	80.2%	60.2%
Responsiveness (person strata)	6	8	7	7	7	5	3

1. Introduction

As outlined by state regulation 603 CMR 7.03 (5), The Massachusetts Department of Elementary and Secondary Education (ESE) is required to administer and publicly report survey data about the quality of educator preparation in the state. The intention is to give Sponsoring Organizations (SOs) and the general public access to important information about perceptions of educator readiness in the Commonwealth. ESE invested significantly in the development of statistically valid instruments such as the educator preparation stakeholder surveys that will be used for various purposes, including programs' continuous improvement, public transparency, and as one source of evidence in program evaluations. ESE developed a suite of surveys in order to triangulate perceptions across different perspectives; ESE surveys teacher candidates, teacher completers, Supervising Practitioners and hiring principals.

The purpose of this report is to provide reliability and validity evidence to support the use of ESE's educator preparation stakeholder surveys for programs' continuous improvement, public transparency, and as one source of evidence in program evaluations. It is intended for readers with knowledge of survey development and validation using the Rasch theoretical framework, psychometrics and educational measurement. Readers should be familiar with Messick's (1980, 1995) validity framework and Wolfe and Smith's (2007a, 2007b) adaptation of this framework for Rasch-based instrument development.

This report is divided into seven sections. Section 1 introduces the validity study. Section 2 summarizes the data sources used in the study. Section 3 describes the test (survey) design, development and administration process for the surveys; the activities described in this section relate mostly to content validity. Section 4 offers a brief report of the data analyses procedures used in the development of the surveys. The validity analyses take advantage of classical test theory (CTT) and item response test theory (IRT). Section 5 provides CTT reliability evidence for the program approval criteria items. Section 6 assesses the reliability and validity of the items that align to observable practices within the Professional Standards for Teachers (PSTs). Section 7 summarizes the totality of the evidence from the validation work and discusses the study's limitations and recommendations.

2. Data Sources and Survey Administration

ESE administered four educator preparation stakeholder surveys in the spring of 2016 to evaluate the perceptions of teacher readiness in the Commonwealth.[4] ESE surveyed the following stakeholder groups[5]:

5. **Teacher Candidate survey,** issued to candidates at the point of program completion
6. **Teacher Completer survey,** issued to educators employed in a Massachusetts public school one year after completing a preparation program
7. **Supervising Practitioner survey,** issued to educators who served as a supervisor to a candidate during the practicum experience
8. **Hiring Principal survey,** issued to principals one year after hiring a preparation program completer

2.1 Survey Stakeholder Profiles and Administration

The profiles of the key stakeholder survey groups are summarized in Table 2.1. Portions of the teacher candidate and teacher completer surveys were bifurcated. All teacher candidates and all teacher completers were administered the same program approval items. The items assessing the PSTs were bifurcated. As such, ESE asked new candidates and

[4] ESE prioritized the development and administration of surveys associated with initial teacher licensure as this is the largest group of program completers in Massachusetts every year.

[5] ESE prioritized the development and administration of surveys associated with *initial teacher licensure* as this is the largest group of program completers in Massachusetts ever year (approximately 4,500 completers annually).

completers who had no prior teaching experience if their experiences in their program "prepared" them for their jobs, whereas candidates or completers who had teaching experience as teachers of record were asked if their program "improved" their ability to perform their jobs. The hiring principal survey was similarly divided such that principals who assessed the performance of a recent program completer who *was already teacher-of-record in their school* evaluated the performance growth of that teacher after they completed their preparation program. Principals who assessed the performance of recent program completers who *were not already a teacher-of-record in the school* evaluated the completers' performance relative to the other teachers in the school. The Supervising Practitioner survey was not bifurcated. ESE administered the surveys in the spring of 2016. The administration specifics for each stakeholder group are shown below in Table 2.1.

Table 2.1: Profiles of Stakeholder Groups

Survey Acronym	Stakeholder Group	Teacher-of-Record during Program	Respondent Group Defined	Survey Administration Defined
NTCD	New Teacher Candidate	No	Graduate of Sponsoring Organization with no prior teaching experience	Self-report 42 PST items
TRCD	Teacher-of-Record Candidate	Yes	Graduate of Sponsoring Organization with prior teaching experience	Self-report 42 PST items
NTCP	New Teacher Completer	No	Employed in MA public school	Self-report 42 PST items
TRCP	Teacher-of-Record Completer	Yes	Employed in MA public school	Self-report 42 PST items
SPCD[1]	Supervising Practitioner	NA	Practicum supervisor of candidates	Perception 42 PST items
NTHP	New Teacher Completer	No	Principals who hired completers (teachers new to teaching) before the start of 2015-2016 school year	Perception 6 PST items
TRHP	Teacher-of-Record Completer	Yes	Principals who hired completers (teachers-of-record) before the start of 2015-2016 school year	Perception 6 PST items

[1]Information was not available to divide Supervising Practitioners' candidates into new candidates and teacher-of-record candidates.

For more information on the different stakeholder groups who took the survey, see Appendix A-1.

2.2 Survey Response Rates

Table 2.2 displays the survey response rates. It is important to note that these surveys represent perceptions of readiness as reported by a subset of stakeholders in the state who elected to take the survey and should not be considered representative of all stakeholders engaged in educator preparation.

Table 2.2: 2016 Survey Response Rates

Survey	Stakeholder Groups	Respondent Group	Number of Respondents	Sampling Frame	Percent Response Rate
NTCD TRCD	New Teacher Candidate Teacher-of-Record Candidate	Self-reports	587	2,311	25%
NTCP TRCP	New Teacher Completer Teacher-of-Record Completer	Self-reports	388	2,064	19%
SPCD	Supervising Practitioner	Supervising Practitioner	649	2,555	25%
NTHP TRHP	New Teacher Hiring Principal Teacher-of-Record Hiring Principal	Hiring Principals	1,077	2,038	53%

2.3. Survey Item Specifications

There are two types of items in the surveys:

3. Items that align to observable outcomes relative to the Program Approval Standards and Review Criteria (henceforth called program approval items), which set forth expectation for providers.
4. Items that align to observable practices within the Professional Standards for Teachers (PSTs) (henceforth called PSTs items), which define the pedagogical and other professional knowledge and skills required of all teachers.

For teacher candidates and teacher completers, the items administered for both types of items were nearly identical, allowing ESE to triangulate evaluation data from the two perspectives. Items related to PSTs were also common to the Supervising Practitioner survey. The hiring principal survey was considerably shorter (only six items) and placed particular emphasis on items that correspond to one of the six essential elements identified through the Candidate Assessment of Performance. Table 2.3 presents the test specifications for scalable common items in each survey. Section 3 provides details on the development of these measures.

Table 2.3: Test Specification for Scalable Measures

Item Type	Topic area	Teacher Candidate	Teacher Completer	Supervising Practitioner
Program Approval Standards and Review Criteria	Overall Program Experience	9	9	NA
	Course Work	3	3	NA
	Field-base Experience	5	5	NA
	Supervision	5	5	NA
	Assessment	4	4	NA
Professional Standards of Practice (PSTs)	Standard I Curriculum, Planning and Assessment	13	13	13
	Standard II Teaching All Students	18	18	18
	Standard III Family and Community Engagement	6	6	6
	Standard IV Professional Culture	5	5	5
	Total Items Readiness Construct	42	42	42

Of note, the primary focus of this report is to provide validity evidence for the development of the measures used to assess teacher readiness in the Professional Standards for Teachers (PSTs), with more limited evidence provided for measures used to assess Program Approval Standards and Review Criteria.

3. Item and Measure Development and Validity Framework

3.1. Item and Measure Development

3.1.1. Program Approval Items

ESE used the Program Approval Standards and Review Criteria to guide item development for the program approval items. Stakeholders were asked for their views related to five primary areas: their overall experience with the program (9 items), their course work (3 items), their field-based experiences (5 items), their supervision during their practicum (5 items), and their views of the assessment used during their practicum (4 items). ESE developed items to evaluate the breadth of their experience in each topic area and to support the assessment of the reliability of stakeholder responses in each topic area or measure.

ESE used a Likert scale with five response options to rate stakeholder perceptions of their programs for each of the program approval items; coding for all items dictated that a response of "0" (*strongly disagree*) would be indicative of the lowest level of perceived readiness with a "4" (*strongly agree*) denoting the highest level of perceived readiness. Response categories scored"1", "2" and "3" corresponded to "*disagree", "neither agree nor disagree",* and "*agree"*, respectively. ESE also used this rating scale to assess stakeholder views on the program's preparation of teachers to perform the classroom practices measured by the PSTs.

3.1.2. Performance Standards for Teachers Items (PSTs)

ESE leveraged work done during the development of the Staff & Student Feedback surveys for the Massachusetts Educator Evaluation Framework. This teacher effectiveness continuum is naturally represented within the descriptors of each rubric element, indicator and standard of ESE's model performance rubric for teachers . As a result, the rubric was used to guide the development of items for Standard I (Curriculum, Planning and Assessment: 13 items), Standard II (Teaching All Students: 18 items), Standard III (Family and Community Engagement: 6 items), and Standard IV (Professional Culture: 5 items). ESE developed items using a hierarchical perspective or mindset. ESE first identified what behaviors (practices) represent proficient and exemplary practices that are relatively easy to enact within the classroom, and then identified those that represent the most difficult practices to enact. Once these practices were identified, items were developed to measure and anchor the two ends of the teacher readiness continuum. The next step in the item development process was to develop items to fill in the continuum. The final step in developing the readiness measure was to ask stakeholders how well their program providers prepared (candidate) or improved (completer) their ability to teach to these standards of practice.

Therefore, the rating scale (strongly disagree to strongly agree) and the type and level of teacher practice measured stretch the item calibrations and person distribution along the teacher readiness continuum in each standard. The item development process differentiated respondents in to high and low scorers on the teacher readiness continuum with the goal of providing programs with feedback that can help them diagnose their relative strengths and weaknesses.

A brief summary highlighting the survey development and piloting of the PSTs items is available on ESE's website. Specifically, the PST items developed for the teacher candidate survey underwent two years of piloting. Using reliability and validity information from the 2014 teacher candidate pilot, PSTs items from this survey served as a foundation for all other stakeholder surveys (teacher completer, Supervising Practitioner and hiring principal) piloted in 2015. The

psychometric analyses from all pilots informed the final form of multiple surveys used to develop the 2016 surveys (the subject of this technical report).

3.2. Validity Framework

Messick's (1980, 1995) unified concept of construct validity guided the validity analyses for the teacher readiness construct. Messick (1995, p. 741) defines validity as "an evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment." Evidence from six aspects of test validity (content, substantive, generalizability, structural, external and consequential) combine to provide test developers with the justification to claim that the meaning or interpretability of the test scores is trustworthy and appropriate for the test's intended use. More recently, Wolfe and Smith (2007a, 2007b) used Messick's validity conceptualization to detail instrument development activities and evidence that are needed to support the use of scores from instruments based on the Rasch measurement framework. Appendix B-1 takes advantage of Wolfe and Smith's (2007b, p. 205) conceptualization to summarize the validity aspects addressed in this technical report. This report primarily focuses on internal validity with some limited external validity evidence provided for the instruments.

4. Data Analyses Procedures

4.1 Classical Test Theory Methodology

Reliability (internal consistency) of both the program approval measures and the PSTs measures were assessed using Cronbach's Alpha; values above 0.9 are desired for high stakes tests with values greater than 0.7 considered acceptable when test scores result from low-stakes tests or used in low-stakes decisions.

To support the structural validity PSTs analyses, the internal structure of the readiness construct was examined using confirmatory factor analyses (CFA); more technical details on CFA are shown in Appendix C-1. Due to limited sample sizes for some of the surveys, these analyses were only performed on the new teacher candidate (NTCD), new teacher completer (NTCP) and the Supervising Practitioner (SPCD) surveys. The CFA were performed using MPLUS (Muthén & Muthén, 2010). Theoretically, the multi-factor construct will be composed of four first-order factors (Standard I, Standard II, Standard III and Standard IV factors) with the latent construct, teacher readiness, explaining the polychoric correlations among the first-order factors. However a plausible alternative for the internal structure of the teacher readiness construct is that the construct is represented by only one factor with insufficient intra-correlations among the items of each standard to provide enough "signal" to be considered separate sub-dimensions. These nested models can be compared for goodness of fit using MPLUS's DIFFTEST.

The criteria used to compare the fit of the CFA models were taken from Schumacker and Lomax (2004, p. 82, Table 5.1) The chi square goodness-of-fit test is susceptible to large sample sizes (Schumacker & Lomax, 2004; Byrne, 2012) resulting in Type I errors or false positives. As a result, this study relies on the root mean square error of approximation (RMSEA), factor loadings, fit indices (Tucker-Lewis Index (TLI) and Confirmatory Fix Index (CFI)) and modification indices to assess model fit. The criteria used to assess and compare model fit are summarized in Appendix C-2.

4.2 Rasch Methodology

Analyses using the Rasch measurement model (Rasch, 1960) and validity framework (Wolfe & Smith, 2007a, 2007b) are the primary source of reliability and validity data for the PSTs survey measures. Technical details explaining the Rasch model are provided in Appendix C-3. In the Rasch framework, the scale metric axis represents the desirable structural

properties of a Rasch scale; it is: linear, unidimensional (measures only one construct), hierarchical (items are ordered according to their difficulty to endorse) and measures a continuum of items and persons. Winsteps software developed by Linacre (2014a, 2014b) was used to perform a rating scale model analysis of the data (Andrich, 1978a; Andrich, 1978b) providing the needed information for each of the construct validity aspects. The evaluation criteria to perform a Rasch-based reliability and validity assessment for each construct validity aspect (content, substantive, generalizability, structural and external) are summarized in Appendix C-2.

5. Validity Evidence for Program Approval Criteria

The data presented in this section relate to teacher candidate and teacher completer perceptions of the program approval items. Supervising Practitioners and hiring principals were, for the most part, not asked these questions.

5.1 Reliability

The Cronbach's alpha of the program approval items are all near or above 0.8 (Table 5.1). Given some topic areas are composed of relatively few items, the reliability of these topic areas are sufficient to make inferences on stakeholder perceptions of their preparation and readiness to teach. The replication of the pattern of reliabilities provides supporting evidence that each form was suitable for use for each context (stakeholder group).

Table 5.1: Cronbach's alpha of topic areas

Cronbach Alpha (# of Items)	New Teacher Candidate (NTCD)	New Teacher Completer (NTCP)	Teacher-of-Record Candidate (TRCD)
Overall Program Experience (9)	0.90	0.91	0.92
Course Work (3)	0.88	0.84	0.92
Field-base Experience (5)	0.73	0.79	0.79
Supervision (5)	0.83	0.82	0.90
Assessment (4)	0.89	0.92	0.91

[1]Due to a survey administration error, data is not available for the Teacher-of-Record Completer responses.

5.2 Descriptive Statistics

A quick assessment of the strength of stakeholder perceptions toward their preparation was assessed using the mean and standard deviation of each program approval topic scale. Stakeholder perceptions are, on average, very positive across each topic and respondent group (Table 5.2). Respondents report that their Sponsoring Organizations are, on the whole, providing students with the content and pedagogical knowledge to be an effective educator. Respondents' course work, field-based experiences, assessment feedback and supervision have all contributed to their positive evaluation of their programs. In summary, these data suggest that there is, overall, a high level of respondent satisfaction with their Sponsoring Organizations and the preparation they received to become effective educators.

Table 5.2: Average perceptions of program approval criteria topic areas[1, 2]

Topic area (# of items)	New Teacher Candidate (NTCD)		New Teacher Completer (NTCP)		Teacher-of-Record Candidate (TRCD)	
	Mean	SD[3]	Mean	SD[3]	Mean	SD[3]
Overall Program Experience (9)	27.6	6.2	27.8	6.4	26.2	6.8
Course Work (3)	9.9	4.0	9.7	4.4	9.5	5.1
Field-base Experience (5)	16.4	3.0	16.3	3.4	15.5	3.4
Supervision (5)	17.1	3.6	17.2	3.5	17.2	3.6
Assessment (4)	12.3	3.3	11.8	3.7	11.9	3.6

[1]Due to a survey administration error, data is not available for the Teacher-of-Record Completer responses; [2]Caution should be taken in interpreting these data; Likert data are ordinal and as such do not meet the requirements of parametric tests. Mean differences should not be compared for statistical significance; [3]Standard Deviation.

6. Validity Evidence for Performance Standards of Teaching Scales (PSTs)

The majority of this report is dedicated to the validity evidence needed to support score use for the PSTs. ESE will present data for five aspects of construct validity: content, substantive, generalizability, structural and external. Appendix C-2 provides a summary of the validity criteria used in this study for each aspect of construct validity.

6.1. Content Validity

Content validity examines the "content relevance, representativeness and technical quality" (Messick, 1995, p.745) of the items used as indicators of the construct. The content validity evidence reported here predominantly focuses on the technical quality of the 2016 survey items and builds on prior work supporting the content validity of the survey items. In 2014-15, ESE developed survey items for each of the four Professional Standards for Teachers (PSTs). Items were developed for the four Standards of effective teaching that comprise the Massachusetts Educator Evaluation Model system: Standard I (Curriculum, Planning and Assessment); Standard II (Teaching All Students); Standard III (Family and Community Engagement); and Standard IV (Professional Culture).

6.1.1. Item Review

Expert review panels (educators from preparation programs, classroom teachers and ESE program staff) reviewed the items developed for each of the four standards. The items were checked for alignment (did they measure the PST it was designed to measure?), accessibility (would respondents be able to read the item and understand it?), actionability (would programs be able to use the information?) and responsiveness (would survey items elicit a range of responses?). The items remaining from this pilot review process form the basis of the 2016 surveys discussed for the remainder of this report. A brief summary of the 2014-15 pilot results is available here.

6.1.2. Item Technical Quality

Item technical quality was assessed using point-to-measure (PTM) correlations and item fit statistics. The PTM correlations and item fit statistics are shown in Appendix D-1 (New Teacher Candidate, New Teacher Completer and Supervising Practitioner); Appendix D-2 (Teacher-of-Record Candidate, Teacher-of-Record Completer) and Appendix D-3 (New Teacher and Teacher-of-Record Hiring Principal). Across the seven surveys, PTM correlations ranged from a low of 0.50 to a high of 0.90 with the vast majority of PTM correlations falling between 0.60 and 0.75; this indicates that the items are positively associated with the PSTs being used to operationalize the readiness construct.

The item fit data reported in the appendices encompass all collected data, with no removal of misfitting persons (persons that score unexpectedly high or low on an item). Across the seven surveys, there were minimal items with infit and outfit statistics that fell outside of the mean square error range of 0.7 – 1.3. Fits statistics of below 0.7 do not degrade measurement and are suggestive of redundancy in content (e.g., items 2.16, 2.17 and 2.18 on the supervising practitioner survey all had fit statistics below 0.7). Fit statistics above 1.3 indicate that the items may not measure the construct of interest; these items have additional source(s) of variance and can degrade measurement. However, for each survey, there were no more than two items with both infit and outfit statistics outside of the upper boundary. The identity of these items was also different across the surveys, suggesting that these deviations could have occurred by chance. Item 2.8 is flagged for monitoring on the "new teacher" surveys (NTCD and NTCP) with item 1.9 flagged on the "teacher-of-record" surveys (TRCD and TRCP) as both these items were misfitting in more than one survey.

6.2. Substantive Validity

Substantive validity assesses whether the responses to the items are consistent with the theoretical framework used to develop the items. Two pieces of evidence are discussed which provide support for the substantive validity aspect of construct validity; these are (1) rating scale functioning and (2) item difficulty hierarchy.

6.2.1. Rating Scale Functioning

Appendices E-1 through E-7 provide the rating scale function data for each of the seven PSTs surveys. Appendix C-2 summarizes the criteria used to evaluate the rating scale functions of each survey. Replicated across the seven surveys, the observed category averages increase monotonically with more affirmative categories. The unweighted mean square error fit statistic is less than 2.0 for all categories in each of the seven surveys. These findings support the claim that the rating scale is being used by the respondents as intended.

The Andrich step thresholds (deltas) similarly increase monotonically in six of the seven surveys; the step function is disordinal only in the new teacher candidate (NTCD) survey. In addition, ideally, the distance between adjacent step thresholds should be greater than 0.8 logits and less than 5 logits apart in a five point Likert scale. This ideal but not essential criterion was not met for three of the seven surveys (NTCD, NTCP, and TRCP). In each of these surveys, respondents were rarely using the middle category (neither agree nor disagree) resulting in this category only covering a narrow interval on the readiness construct continuum. This finding also helps to explain the disordinality observed in the NTCD survey. For some categories of the hiring principal surveys (NTHP and TRHP), the distance between step thresholds was greater than 5.0 logits. The Andrich step functioning is working relatively well in six of the seven surveys. With the exception of the hiring principal surveys, however, the data do suggest that a four category response option could capture most of the variance in responses. For the two hiring principal surveys, a six or seven category response option may be beneficial.

6.2.2. Item Hierarchy

A qualitative assessment is used to evaluate the PSTs item hierarchies along the readiness construct continuum. It assesses how well the item hierarchies correspond to the instrument developer's *a priori* theoretical expectations. Item-variable maps aid this process by placing items and persons on to the same scale metric. The maps are presented for the new teacher candidate (NTCD), new teacher completer (NTCP), and supervising practitioner PSTs items in Appendices E-8, E-9 and E10, respectively (teacher-of-record item-variable maps are not shown but are comparable).

Overall, the ordered pattern of item difficulties conforms to theoretical expectations across all five surveys. In terms of standards, items belonging to Standard III (Family and Community Engagement) were, on average, more difficult to endorse than items from Standard I (Curriculum and Assessment) and Standard II (Teaching All Students), which in turn,

were, on average, more difficult to endorse than items from Standard IV (Professional Culture). This rank ordering could be explained by the fact that new teachers have limited exposure to engaging families and communities in support of their students. Practices of Standard IV involve self-reflection by the respondent, a relatively easier task than what is required to teach in a classroom. The replication of item patterns across the three item-variable maps supports the substantive validity aspect.

In addition to looking at the overall item hierarchy, developers can assess item hierarchies within elements of the standards. Appendices E-11 and E12 provide examples of item hierarchies related to assessment practices (Standard I) and classroom management practices (Standard II). Practices associated with assessment and classroom management are hard for new teachers (Ferguson, 2010) and they are both, on average, the most difficult items to endorse in Standard I and Standard II, respectively. Within these elements, the item hierarchies met our *a priori* expectations (see Appendices E-11 and E12). For example, it is a much harder task for new teachers to effectively engage students who resist wanting to learn (item 2.5; ~0.6 logits) than it is to refocus students who have become distracted (item 2.3; ~0.20 logits) or to use classroom management techniques to keep students on task (item 2.8; ~-0.10 logits). The replication of item hierarchies within each of the four standards provides further evidence to support the substantive validity aspect.

6.3. Generalizability

A measure is considered generalizable when the score meaning and properties function similarly across multiple contexts (e.g., stakeholder groups, forms) or time points. Reliability analyses and differential item functioning (DIF) analyses are used to assess the generalizability of the measures. Similar to Cronbach's alpha, person separation reliability (PSR) looks at the stability (internal consistency) of the measures across the instruments (Schumacker and Smith, 2007) and scoring structures and helps set the boundary for the inferences made using the measures. ESE used DIF analyses to empirically test for item invariance across subgroups and surveys.

6.3.1. Reliability Evidence

Table 6.1 provides descriptive data and the PSR for the 42-item PST measure for each survey (6-item PST measures for the hiring principal stakeholder group). With the exception of the new teacher hiring principal survey (NTHP), the average scaled-score of the stakeholder groups are high (greater than 1.5 logits), indicating that, on average, the multiple stakeholders perceived the program providers are doing a good job preparing teachers for the classroom. Psychometrically, the items are not well-targeted for the person distributions; however, this is ameliorated by the robust variability in stakeholder responses (standard deviations were above 1.5 logits in all stakeholder measures) and the variance explained by each measure (greater than 45% over all surveys). With the exception of the teacher-of-record hiring principal survey (TRHP), the PSR are all above 0.9; these replicable reliabilities provide supporting evidence for the generalizability of the PST measures. Of note, is the high percentage of extreme scores (37%) in the TRHP survey, which has likely reduced the reliability of the PST measure for this group.

The reliability analyses examined the reliability of the scores derived from separate, free calibrations of the items that make up the four PST. Descriptive data and other reliability data for each standard are shown in Appendices F-1 through F-4. In Table 6.2, a summary of the PSR data is provided. For the thirteen and eighteen item Standard I and Standard II measures, respectively, the PSRs are robust and replicable, with reliabilities near or above 0.9. Similarly, the reliabilities for the six item Standard III measure are all above 0.8 (with the exception of the TRCP, where ceiling effects appear to reduce the reliability of the items overall). The reliabilities for the five item Standard IV measure were near or above 0.70, indicating less internal consistency of this measure.

Table 6.1: Descriptive Data and Real Person Separation Reliabilities (PSR) for Educator Preparation Surveys

Non-Extreme Person Report	New Teacher Candidate (NTCD)	Teacher-of- Record Candidate (TRCD)	New Teacher Completer (NTCP)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)	New Teacher Hiring Principal (NTHP)	Teacher-of-Record Hiring Principal (TRHP)
Number of Respondents[1]	370 (408)	167 (179)	185 (202)	167 (186)	625 (649)	571 (628)	275 (449)
Maximum Extreme Score	9.1%	7.7%	8.4%	9.7%	3.5%	3.2%	36.9%
Number of Items	42	42	42	42	42	6	6
Mean	2.28	2.57	1.92	1.95	2.08	-0.11	3.97
Standard Deviation	1.60	2.07	1.69	1.56	1.65	4.16	2.97
Variance Explained	45.5%	60.3%	51.9%	49.9%	51.7%	80.2%	60.2%
Real PSR[2]	0.95	0.97	0.96	0.96	0.96	0.94	0.80
Responsiveness (person strata)	6	8	7	7	7	5	3

[1]Number in parentheses includes persons with extreme measures, [2] Real Person Separation Reliability (lower bound)

Table 6.2: Person Separation Reliabilities across the Four Professional Standards for Teachers (PSTs)

Non-Extreme Person Report	New Teacher Candidate (NTCD)	Teacher-of- Record Candidate (TRCD	New Teacher Completer (NTCP)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)	New Teacher Hiring Principal (NTHP)[1]	Teacher-of-Record Hiring Principal (TRHP)[1]
Standard I (13 items)	0.87	0.91	0.87	0.88	0.90	NA	NA
Standard II (18 items)	0.90	0.94	0.92	0.92	0.93	NA	NA
Standard III (6 items)	0.86	0.87	0.87	0.59	0.87	NA	NA
Standard IV (5 items)	0.69	0.67	0.74	0.77	0.70	NA	NA

[1]The hiring principal survey has an insufficient number of items to assess reliability of the four standards.

Overall, the replicable reliability data support the generalizability of the PST standard measures. With the caveat noted for Standard IV items, the reliability data support providing program providers with scaled scores for each standard.

6.3.2. Differential Item Functioning

Respondents with the same agreement disposition (level) should have the same probability of endorsing an item irrespective of the subgroup they belong to. The DIF analyses examined and compared the functioning of the PSTs items for respondents attending public and private preparation program providers. ESE selected this particular comparison because of the relatively equal sample size in both categories. Alternative comparisons were not appropriate due to insufficient sample size in certain subgroups. ESE performed these analyses for all seven surveys. The DIF plots showing the average item deltas for each of the stakeholder groups are shown in Appendices F-5 through F-11.

When comparing public to private school respondents, the items do not function differently across the two new teacher surveys (NTCD and NTCP), the supervising practitioner survey (SPCD) and the hiring principal new teacher survey (NTHP). There was one Standard IV item (item 4.1) which exhibited DIF (0.72 logit difference) on the hiring principal teacher-of-record survey (TRHP). This item behaved well on the NTHP survey; this item should be monitored in future administrations of the TRHP survey. Three items exhibited DIF on the teacher-of-record candidate (TRCD) and teacher-of-record completer (TRCP). These items were different on each survey. The DIF analyses for these surveys were limited by the number of respondents who attend public schools. With less than 50 public school respondents in each of these surveys and with 5 response options available, it is likely that there is insufficient data to perform a reliable DIF analysis.

Where there is sufficient data, the DIF analyses support the claim that respondents with equal latent trait propensities have the same probability of endorsing any one of the survey items. These findings validate the generalizability of the PSTs items measuring the readiness construct.

6.3.3. Item Invariance Across Surveys

Item invariance across multiple contexts is an important assumption of the Rasch model (Wolfe & Smith, 2007b; Bond & Fox, 2007) and ensures that the items used to measure a construct have the same meaning and interpretation across subgroups and contexts. The 42 items used for five of the surveys (NTCD, NTCP, TRCD, TRCP and SPCD) were essentially the same with the only difference accorded to the question stem. This provides an opportunity to examine item invariance across different stakeholder groups. A free calibration of the 42 items on each of the surveys was performed separately for each survey. Appendix G-1, G-2, G-3 and G-4 compare the average item deltas for all 42 items broken out by Standard I items (13 items), Standard II items (18 items), Standard III items (6 items) and Standard IV items (5 items), respectively. Without accounting for measurement error, there were very few items that differed by more than 0.5 logits, indicating that there is a high degree of correspondence between the item deltas. To support these analyses, ESE calculated the Pearson product moment correlation coefficients (henceforth Pearson correlation) for each paired survey comparison.

Table 6.3 provides the attenuated (below diagonal) and disattenuated (above the diagonal) Pearson correlations for the five PSTs survey items. The Pearson correlations are all near or above 0.9 for each paired comparison. These results support the prior evidence that the items are largely invariant and are appropriate for use across multiple contexts. In Appendices G-5 and G-6, comparable data are shown for the two six-item hiring principal surveys (NTHP and TRHP). The results indicate only a moderate positive correlation (0.48 disattenuated) between the two sets of item parameters. The absolute difference in deltas is over 0.6 for two items (item 1.1 and item 4.1). These data suggest that the context of the survey (new teacher and teacher-of-record) is making hiring principals respond differently to the same items (caution should be used when comparing data from these two surveys). Readers should not compare results of principals who had a new teacher to those that had a teacher-of-record.

Table 6.3: Rasch Item Delta Invariance across Five Professional Teaching Standards Surveys[1]

Number of Items = 42	New Teacher Candidate (NTCD)	New Teacher Completer (NTCP)	Teacher-of-Record Candidate (TRCD)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)
New Teacher Candidate	1	0.99	0.97	0.97	0.95
New Teacher Completer	0.94	1	1.00	1.00	1.00
Teacher-of-Record Candidate	0.93	0.89	1	1.00	0.93
Teacher-of-Record Completer	0.92	0.93	0.92	1	0.93
Supervising Practitioner	0.93	0.93	0.87	0.86	1

[1]Pearson correlations observed are shown below the diagonal; Disattenuated correlations are shown above the diagonal.

6.4. Structural Validity

Structural validity evaluates the alignment of the scoring structure to the hypothesized structure of the construct. The fundamental assumption of the Rasch model is that it is used to measure one latent construct (in this study, the teacher readiness construct). If the data meet this requirement, the measures are linear, invariant and additive; equal differences on the scale translate into equal differences in the probability of getting an item right (or endorsing an item) no matter where on the scale an item is located. For the Rasch data in this study, the unidimensionality of the data are assessed by (1) an analysis of the freely calibrated sub-scale (standards) correlations and (2) an analysis of the standardized residuals and additional dimensionality data provided by the Rasch software. In addition, ESE used confirmatory factor analyses (CFA) to substantiate the Rasch-based findings. The root mean square error of approximation (RMSEA), fit indices (Tucker-Lewis Index (TLI) and Comparative Fit Index (CFI)), factor loadings and modification indices were used to support the CFA validity data. The criteria used for both methodologies to evaluate the structural validity of the teacher readiness construct are shown is Appendix C-2.

6.4.1. Rasch Dimensionality Analyses

Subscale Correlations. ESE evaluated the Pearson correlation between subscale scores (standards) for the four freely calibrated PSTs item sets. The correlations should be positive and of sufficient magnitude (greater than 0.5 but less than 0.9) to indicate that the four sub-scales are measuring distinct but related sub-dimensions of the teacher readiness construct. The results from these Rasch analyses are provided for the new teacher candidate (NTCD) in Table 6.4 (above the diagonal), with comparable results for the NTCP and SPCD surveys found in Appendix H-1.

The subscale correlations range from 0.67 (III and IV) to 0.89 (I and II). The pattern of correlations is as expected with the highest correlation between Standards I and II and the lowest between Standards III and IV. Standard I (Curriculum, Planning and Assessment) and Standard II (Teaching All Students) ask for respondents' views on the practices needed to effectively teach within the classroom environment and the scores are highly related due to this common context. In contrast, the expectation is that the association between practices embedded within Standard III (Family and Community Engagement) and Standard IV (Professional Culture) is lower; Standard III pertains to relationships outside of the school building and Standard IV pertains to self-reflection and the culture within the school building.

Table 6.4: New Teacher Candidate (NTCD) Sub-scale Correlations[1]

	PST I	PST II	PST III	PST IV
PST I		0.89	0.76	0.78
PST II	0.90		0.79	0.81
PST III	0.72	0.76		0.67
PST IV	0.80	0.81	0.62	

[1]Sub-scale correlations below the diagonal are from the four-factor confirmatory factor analysis model; Correlations above the diagonal are the person measure correlations from the four standard-based consecutive (separate) Rasch models.

Although each correlation indicates that the items for each standard are distinct, there is not sufficient signal to suggest that the construct would be better measured using a multi-dimensional model. This pattern of correlations is replicated in the NTCP and SPCD data (Appendix H-1) which provides further confirmatory evidence of the unidimensionality of the readiness construct.

Residual Analyses. If the data fit the model and the variance in responses is explained by one latent trait (teacher readiness construct), the unexplained or residual variance should be random (i.e., there is no relationship between residuals). The principal components residual analysis first removes the common variance attributable to the first dimension (teacher readiness construct). Once this variance is removed, it examines the pattern of the contrasts or unexplained variance (residuals). Of note, with the exception of the new teacher candidate survey (NTCD, 45.5%), the total variance explained across all surveys was greater than 50% (Table 6.1).

The residual data analyses suggest a second dimension across the five 42-item scales (NTCD, TRCD, NTCP, TRCP, and SPCD). Items of Standard III (Family and Community Engagement) appear to explain the non-random variance observed in the residuals. For each of these surveys, this residual variance accounted for near or slightly over 5% of the total variance explained by the model and the eigenvalue of the residual items was above 3 (both criterion fall slightly outside the cutoffs supporting unidimensionality). Similarly, the variance explained by the items of the first dimension (teacher readiness construct) was, for each survey, between 3.3-3.7 times the variance explained by the first contrast (potential second dimension), slightly below the criterion of a multiple of four (Appendix C-2) to support unidimensionality. These data suggest that items related to Standard III are possibly forming a separate dimension and the teacher readiness construct is multidimensional in nature. However, the deviance from the unidimensionality criteria (Appendix C-2) was not definitive to state conclusively that a second dimension is apparent as the data were close to be compliant in all surveys. In addition, an examination of the item fit statistics for Standard III items (Appendix D1 – Appendix D2) support the unidimensionality of the construct. The items of Standard III are well fitting (between 0.7 and 1.3) and have relatively strong item-to-measure correlations (all above 0.6).

To investigate the unidimensionality of the scales further, ESE examined the item clusters formed by the residuals for each of the 42-item surveys. In all surveys, three clusters of items were apparent in the residuals; items from Standard III formed the first cluster. The correlations between the person measures of the first cluster items (Standard III) and cluster 2 and cluster 3 items were above 0.8 and 0.9, respectively. If the residual clusters are different dimensions, ESE expected the correlations to be small and the person measures to be considerably different for the different clusters (Linacre, 2014a; Linacre, 2014b). These strong correlations between the three residual clusters across the five surveys suggest that all of the items belong to the same latent trait (teacher readiness construct). In addition, from a theoretical

perspective, family and community engagement is an important and distinct theoretical component of the teacher readiness construct, warranting its place as one of the sub-dimensions of the construct.

The explained variance of the two six-item hiring principal surveys was over 60%. There was no evidence of a second dimension in either survey; eigenvalues were less than two and cluster correlations were 1.0. Of note, the hiring principal surveys did not include any items from Standard III.

Overall, the evidence from all the surveys supports the structural validity aspect of the teacher readiness construct. The one dimension extracted by the Rasch model meets the unidimensionality assumption of the Rasch model. With all but one of the surveys explaining over 50% of the variation in stakeholder responses and the totality of the structural evidence, ESE's use of scaled scores to measure stakeholder perceptions of teacher readiness is supported and justified.

6.4.2. Confirmatory Factor Analyses (CFA)

The primary focus of these analyses was to provide confirmatory evidence that the teacher readiness construct is composed of four sub-dimensions (the four standards) with the underlying latent trait (teacher readiness) explaining the correlation between the items of the four standards (a four-factor confirmatory factor model). Another plausible model for the teacher readiness construct is that all of the items form one factor and there are no sub-dimensions within the model. These analyses were only performed for the three of the five 42-item PST surveys (NTCD, NTCP, and SPCD). CFA is a classical theory technique that uses raw Likert responses and is distinct from Rasch. The Rasch model transforms the Likert-scale raw responses into interval-level logits and extracts all the common variance across all items with the goal of forming and validating a unidimensional construct. For CFA, a model is postulated *a priori* and the observed correlational data structure is analyzed to determine if the CFA model recovers the expected structure specified by the *a priori* model. The analyses began by examining and comparing the one-factor model with a four factor model; as a result of these analyses, a three factor model was also tested (items from Standard and Standard II were combined to represent one factor). The criteria used to evaluate the CFA results are shown in Appendix C-2.

First-Order CFA Models. The model fit for the one factor model was acceptable but the model did not recover the *a priori* structure well. Table 6.5 shows the model fit for the new teacher candidate (NTCD) analyses. The RMSEA point estimate was 0.094, above 0.05 our criterion used to denote a well-fitting model. Similarly, the CFI and TLI were both above 0.9 but below 0.95, which indicates the model is good but is likely not recovering all of the relationships between the items. The factor loadings were all above 0.5, suggesting that the items are all related to the construct (data not shown). The combined data indicate that the one factor model is reasonable but is not fully recovering the structure embedded within the construct.

ESE specified a four factor model. This model recovered the structure of the observed data well. The RMSEA point estimate was 0.055, bounded by a 90% confidence interval of 0.052 – 0.059. The CFI and TLI both improved over the one-factor specification with both indices equaling 0.97. All factor loadings were above 0.5. Although the RMSEA is not below 0.05, the four factor model is a clear improvement on the one-factor model with the RMSEA close to our criterion for excellent model fit.

Appendix J-1 repeats Table 6.5 and provides the results for the other two surveys (NTCP and SPCD) for comparison purposes. The results show that the pattern of results are replicated in the new teacher completer (NTCP) and supervising practitioner survey (SPCD) supporting the primacy of the four factor model (of note, the model fit statistics for the NTCP and SPCD were very good but were slightly poorer than the NTCD analyses).

The correlations of the four factors for the NTCD four factor model are shown in Table 6.4 (below the diagonal) and again in Appendix H-1 where they can be compared to those of the NTCP and SPCD surveys. Similar to the Rasch analyses, the correlation between the items of Standard I and Standard II (0.90) is the strongest with the correlation between the items of Standard III and Standard IV the lowest (0.62). This pattern of correlations is replicated in both the NTCP and SPCD surveys. The high correlation between the first two standards suggests that the items from these two standards could be combined into one factor. A three-factor model was tested; the results are shown in Table 6.5. The model fit was good and offers a plausible competing model to the four factor model. The point estimate of the RMSEA (greater than 0.6) is not as good as the fit as the four factor model but the values of the CFI and TLI indices are comparable.

Table 6.5: New Teacher Candidate (NTCD)

Number of Items = 42	One-Factor (All 42 Items)	Four-Factor (Four Performance Standards)	Three-Factor (PST I items combined with PST II items)
RMSEA[1]	0.094*	0.055*	0.061*
RMSEA (90% CI)[2]	0.091 – 0.97	0.052 – 0.059	0.058 – 0.064
CFI[3]	0.92	0.97	0.97
TLI[4]	0.92	0.97	0.97
Factor Loadings above 0.5	YES	YES	YES

[1]RMSEA: Root Mean Square Error of Approximation point estimate; [2]RMSEA 90% Confidence Interval; [3]CFI: Comparative Fit Index; [4]TLI: Tucker-Lewis Index; *$p < 0.0005$.

ESE used the CFA DIFFTEST provided by MPLUS (Muthén & Muthén, 2010) to statistically compare the competing models to determine which model (one-factor model, three-factor and four-factor model) best recovered the teacher readiness construct data. The results are shown in Table 6.6.

Table 6.6: Confirmatory Factor Analyses Chi-Square Model Difference Test

	One-Factor (H0) vs. Four-Factor (H1)	Three-Factor (H0) vs. Four-Factor (H1)	Second-Order (H0) vs. Four-Factor (H1)
NTCD N = 407	χ^2 Value: 412.53 *df* = 6 *p*<0.0005	χ^2 Value: 81.01 *df* = 3 *p*<0.0005	Insufficient Sample
NTCP N = 202	χ^2 Value: 277.30 *df* = 6 *p*<0.0005	χ^2 Value: 77.86 *df* = 3 *p*<0.0005	Insufficient Sample
SPCD N = 649	χ^2 Value: 849.95 *df* = 6 *p*<0.0005	χ^2 Value: 159.71 *df* = 3 *p*<0.0005	χ^2 Value: 24.50 *df* = 2 *p*<0.0005

MPLUS uses a Weighted Least Square Mean Variance (MLSMV) estimator in modeling categorical data; as a result the DIFFTEST procedure adjusts the chi-square statistic to reflect that this difference value is not distributed as a chi-square. In all the comparisons, H0 is the more restrictive null model. In the second column, the unidimensional one-factor model is the more restrictive model and is nested within the four-factor model (H_1); in contrast, in the third column, the three-factor model is the more restrictive model (H0) and is nested within the four-factor model (H1). In all comparisons, we rejected the H0 null model that the less restrictive model (H1) does not significantly improve model fit.

Across all three surveys, the four-factor model significantly improves the model fit when compared to both the one – factor model and the three-factor model. This provides strong replicable evidence to support the internal structure of the teacher readiness construct which is composed of items measuring the four standards of professional teacher practice.

Second-Order CFA Model. The hypothesized structure of the teacher readiness construct is that the 42 items will load onto their respective standards (the four first-order factors) that they were intended to measure and that a higher-order teacher readiness construct explains the relationship among and between these factors. This model was tested for all three surveys (NTCD, NTCP and SPCD); however, there was only sufficient data to test if the second-order CFA model was a better fitting model than the four factor first-order factor model for the supervising practitioner data (SPCD). The RMSEA, CFI and TLI improved marginally over the data reported in Appendix J-1. The data are shown in Table 6.7 indicating that the second-order factor model was of good model fit. The standardized correlations between the first-order factors and the second-order factor are shown in Table 6.8. The pattern of correlations replicates across the three surveys supporting the hypothesis that the teacher readiness construct is the higher-order factor that explains the covariation among the four standards. The variance explained of the underlying continuous PSTs factors by the teacher readiness construct ranges from 62% for Standard IV (SPCD) to 99% for Standard II (NTCP).

Table 6.7: Model Fit Statistics for Second-Order Confirmatory Factor Model

Number of Items = 42	New Teacher Candidate NTCD (N = 407)	New Teacher Completer NTCP (N = 202)	Supervising Practitioner SPCD (N = 649)
RMSEA[1]	0.054*	0.065*	0.066*
RMSEA (90% CI)[2]	0.051 – 0.57	0.060 – 0.070	0.064 – 0.069
CFI[3]	0.97	0.97	0.96
TLI[4]	0.97	0.97	0.96
Factor Loadings above 0.5	YES	YES	YES

Table 6.8: Standardized Correlation and Percent Variance Explained between PSTs and the Second-Order Teacher Readiness Factor[1]

First-Order PST Factors	NTCD N = 407	NTCP N = 202	SPCD N = 649
Standard I (13 items)	0.93 (87%)	0.91 (83%)	0.94 (88%)
Standard II (18 items)	0.97 (94%)	0.99 (98%)	0.97 (94%)
Standard III (6 items)	0.78 (61%)	0.84 (71%)	0.74 (55%)
Standard IV (5 items)	0.84 (71%)	0.82 (67%)	0.79 (62%)

[1]Percent variance explained (squared correlation) found in parentheses

The relatively high level of variance explained by the second-order teacher readiness factor is also evident in the factor loadings of the individual items (data not shown). For the SPCD survey, the variance explained by the underlying

continuous PST items range from 48% to 78% for Standard I, 58% to 81% for Standard II, 75% to 84% for Standard III, and 79% to 86% for Standard IV. Overall, the fit statistics (RMSEA, CFI and TLI) and variance explained by each item and factor support the structural validity aspect of the teacher readiness construct. The results provide confirmatory evidence that the teacher readiness construct is composed of four factors (standards) with one higher-order factor (teacher readiness) explaining the relationship between these factors.

DIFFTEST results (Table 6.6) indicate that the second-order factor model significantly improved the model fit for the SPCD survey data; this further supports the structural validity of a higher-order teacher readiness construct. These data should, however, be treated with caution as it was not possible to replicate the DIFFTEST for other stakeholder data due to insufficient sample size.

Modification Indices. Modification indices (MIs) are used to assess whether the model has been mis-specified and examines whether the parameter constraints (e.g., assignment of an item to load onto a particular factor) are poorly constructed. The modification indices highlight some areas where the specification of the model could be improved. The "BY" statements highlight possible cross loadings of items if they were allowed to parameterize freely. For example, item 2.10 from Standard II (the candidate is able to connect students with socio-emotional problems with appropriate supports) loaded on both Standard II (Teaching All Students) and Standard III (Family and Community Engagement). If this item was allowed to cross-load, the reduction in chi-square would be significant and improve the overall model fit. Theoretically, this association makes sense as Standard III is concerned with connecting with students' families and the community to support student well-being. However, it is also an important classroom practice encouraging inclusion of all students in the classroom learning environment.

Similarly, the "WITH" statements produced point to areas where items' residual covariances are correlated. For example, the residual covariances of items that are related to English Language Learners (e.g., items 1.1 and 2.6) are highly correlated and, if this was modeled, it would improve the fit of the model. Similar findings were evident for classroom management and assessment content-related items. These findings suggest redundancy in these types of items and may be an avenue to explore to reduce survey length. However, theoretically, the content of these items form important aspects of classroom practice and capture different facets of teacher readiness in these knowledge/skill areas. The Rasch analyses also indicated that these items were not uniformly difficult to enact (their average item deltas spread along the teacher readiness continuum from low to high) and they therefore help support the responsiveness of the PST scale.

The decision was made to not respecify the models in order to keep the parameterization as parsimonious as possible. In addition, the final models for each of the surveys have good fit statistics (RMSEA near to 0.05 and TLI and CFI above 0.95) and provide supporting and replicable evidence for the structural validity aspect of the teacher readiness construct.

6.5. External Validity

This aspect of construct validity relates to the responsiveness of an instrument and the relationship of its scores to the scores of external measures. It also examines whether instrument scores can predict future scores on a criterion measure. The responsiveness of an instrument refers to "the degree to which an instrument is capable of detecting changes in person measures following an intervention that is assumed to impact the target construct" (Wolfe & Smith, 2007b, p. 222). If an instrument is responsive, it can be applied appropriately to measure expected group differences or individual change. Unfortunately, this study does not examine convergent validity (relationship between teacher readiness scores and scores from instruments measuring the same construct) or divergent validity (relationship between

teacher readiness scores and scores from instruments measuring related but distinct constructs). For example, we might expect to see a moderately strong relationship between teacher readiness scores and respondents' educator evaluation scores or student feedback survey scores (convergent validity). In contrast, it is likely that, if examined, the correlation between teacher readiness scores and teacher candidates' average student achievement scores would be lower as the two constructs are less related constructs (divergent validity). In addition, there was insufficient data to perform analyses on the relationship between preparation provider respondent scores and criterion measures (e.g., respondents' student achievement). This study restricts itself to examining survey responsiveness, distributional properties of the teacher readiness scores and measuring group differences.

6.5.1. Survey Responsiveness

The responsiveness of an instrument is measured by the person strata index, H, which provides the number of statistically distinct ability or endorsement groups whose centers of score distributions are separated by at least three standard errors of measurement within the sample. If an instrument is designed to reliably differentiate individuals into four performance levels such as in the teacher evaluation rubric (upon which the teacher readiness construct is premised), the person strata index should be above four.

With the exception of the hiring principal surveys (NTHP and TRHP), the number of person strata for all of the 5 other surveys (NTCD, TRCD, NTCP, TRCP, and SPCD) was equal to or exceeded six person strata (Table 6.1). These data show that the 42-item PST instrument is responsive across stakeholder groups and should be able to reliably measure group differences and any change in stakeholder perceptions of teacher readiness. The six-item new teacher hiring principal survey (NTHP) is able to reliably differentiate between five scoring levels, with the teacher-or-record (TRHP) survey only being able to discern three statistically distinct score groups. This latter survey had a high level of maximum scores (36%), which creates a ceiling effect on the score distribution. In addition, the spread (standard deviation) of the TRHP items was only 0.24 logits for the TRHP survey; this compares to 0.59 logits for the NTHP survey (data not shown). This suggests that the NTHP survey is behaving differently than the TRHP survey. However, with the caveat related to the NTHP survey, these data, overall, strongly support the external validity aspect of the teacher readiness construct.

6.5.2. Non-Extreme Person and Preparation Provider Distributions

The non-extreme person and preparation provider distributions for the new teacher candidate survey respondents (NTCD) are shown in Figure 6.1 and Figure 6.2, respectively. These are repeated in Appendix K-1 and K-2 along with the non-extreme person and preparation provider distributions for the new teacher completer (NTCP), supervising practitioner (SPCD) and the new teacher hiring principal (NTHP) survey respondents (Appendices K-3 through K-8). The aggregate preparation provider distributions are limited in scope by the requirement that at least 6 respondents' scores were needed to provide the average measures. For example, for the NTCD survey, there were only 22 out of a possible 80 preparation provider scores used to examine the distributional properties of the aggregate measures.

The person measure distributions of the NTCD (Figure 6.1) and NTCP (Appendix K-3) respondents are relatively normal but exhibit positive skewness. The person measure distributions of the SPCD (Appendix K-5) and NTHP (Appendix K-7) survey respondents are normally distributed. As the responsiveness data indicates (section 6.5.2), the variability in person responses (and hence their perceptions of their readiness) is robust.

Best test design proponents advocate that the person mean should be within 1 logit of the item mean (which is set to 0.00 logits in all surveys). With the exception of the NTHP survey, the item measures are not particularly well targeted for the respondent distribution measures; for 6 of the 7 surveys, the person means are above 1.0 logit, indicating overall favorable views of respondents toward their teacher readiness. However, the spread of person scores suggests that the scale is doing a good job of capturing respondents' views. For the NTCD survey responses, 95% of the respondents'

scores fall within the range of -0.85 logits and +5.41 logits. Similarly, 95% of NTCP, SPCD and NTHP scores lie between -1.40 and +5.24 logits, -1.16 and +5.32 logits, and -3.61 and +3.92 logits, respectively. These data are shown in Table 6.9.

Figure 6.1:
New Teacher Candidate (NTCD) Non-Extreme Person Measure Distribution

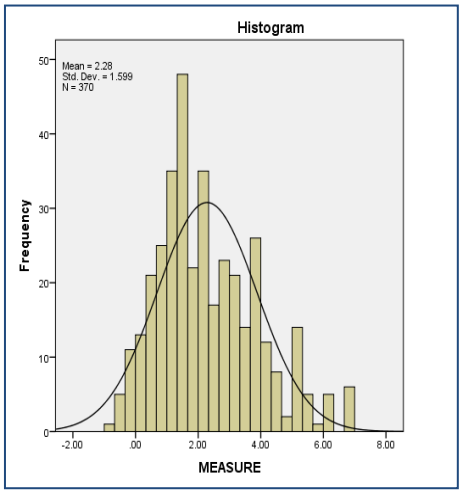
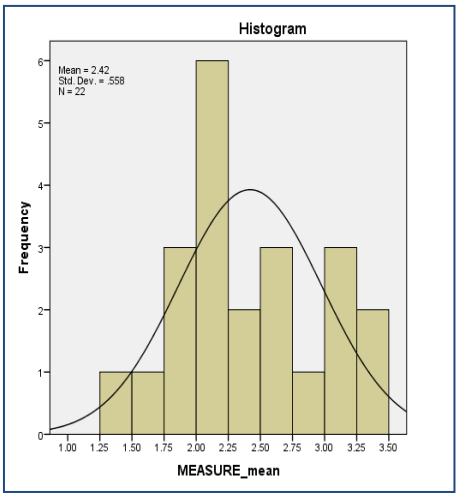


Figure 6.2:
Preparation Provider Aggregate NTCD Person Measure Distribution



[1]Preparation Providers with at least 6 respondents are shown.

Table 6.9: Non-Extreme Person and Institution Teacher Readiness Score Distributions

Score Distributions (logits)	New Teacher Candidate (NTCD)	New Teacher Completer (NTCP)	Supervising Practitioner (SPCD)	New Teacher Hiring Principal (NTHP)
Person Distribution Mean	2.28	1.92	2.08	-0.11
Number of Persons	370	185	625	571
95% Person CI	-0.85 – 5.41	-1.40 – 5.24	-1.16 – 5.32	-3.61 – 3.92
Institution Distribution Mean	2.41	2.18	2.51	-0.49
Number of Institutions	22	13	33	30
95% Institution CI	1.33 – 3.51	0.38 – 3.98	0.10 – 4.92	-3.30 – 2.32

At the preparation provider level, there is variability in the average scores for each of the stakeholder groups. Caution is required when evaluating these results. In order to obtain an adequate number of institution scores for the analyses, ESE used a cutoff of six respondents to compute the means for preparation providers (the ideal is ten respondents). Even with this cutoff, the number of institutions in the NTCP analyses was only 13. Figure 6.2 shows the person score distribution aggregated across preparation providers for the NTCD survey respondents. The preparation provider distributions for the new teacher completer (NTCP), supervising practitioner (SPCD) and the new teacher hiring principal (NTHP) survey respondents are found in Appendix K-4, K-6 and K-8, respectively.

For the NTCD survey, for example, 95% of the institution scores fall in the range of +1.33 logits to +3.51 logits. For the NTCP, SPCD and NTHP scores, 95% of the institution average scores lie between +0.38 to +3.98 logits, +0.10 to +4.92 logits, and -3.30 to +2.32 logits, respectively. These data are shown in Table 6.9. The new teacher hiring principal survey

(NTHP) seems particularly well targeted (person mean of -0.31 logits) at the person level and also highly variable at the institution level (N = 30). Interestingly, the teacher-of-record hiring principal survey respondents show less variability and respondents held more favorable views of their teacher-of-records' improvement in practice. Overall, at the institution level, the high level of variability in preparation provider average scores supports using the scores to differentiate teacher readiness at the institutional level. Future work should include an analysis to determine the reliability of scores at this aggregate level when sample sizes are increased.

6.5.3. Measuring Group Differences

Another important purpose for the PST instrument is to measure group differences. Group differences could be treatment groups or differences in respondent demographics. The essential question is, do the respondents perceive they have different capabilities due to the context of their preparation experience? For example, do respondents who attended private institutions differ in their views from those who attended public preparation providers?

Table 6.10 shows the mean difference between respondents attending public and private institutions for the NTCD, NTCP, TRCD, TRCP and the SPCD survey respondents. For three of the respondent groups (NTCD, TRCD and TRCP), there was no significant difference between the mean person scores. In contrast, new teacher completer (NTCP) respondents who attended private institutions rated their improvement significantly higher than those attending public institutions (mean difference of 0.60). This difference is of a small-to-moderate effect size (Cohen's D equals 0.36). Similarly, Supervising Practitioners who supervised candidates from private institutions rated their readiness significantly higher (0.35) than those that supervised candidates from public institutions; this difference was of a small effect size.

Table 6.10: Mean Differences in Teacher Readiness between Public and Private Preparation Providers

Number of Items = 42	New Teacher Candidate (NTCD)	New Teacher Completer (NTCP)	Teacher-of-Record Candidate (TRCD)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)
Public Mean ± SD (N, Reliability)[1]	2.17 ± 1.64 (130, 0.96)	1.53 ± 1.61 (66, 0.96)	2.68 ± 2.24 (51, 0.98)	1.79 ± 1.44 (49, 0.96)	1.88 ± 1.62 (272, 0.97)
Private Mean ± SD (N, Reliability)[1]	2.35 ± 1.57 (240, 0.96)	2.13 ± 1.69 (119, 0.96)	2.52 ± 1.99 (116, 0.97)	2.02 ± 1.79 (118, 0.96)	2.23 ± 1.65 (353, 0.97)
Mean Difference	0.17	0.60	0.18	0.23	0.35
Significance	NS	$p < 0.05$	NS	NS	$p < 0.01$
Cohen's D	---	0.36	---	---	0.21

[1]Reliability: Person Separation Reliability for group parameter estimate.

7. Conclusion

7.1. Program Approval Criteria.

The indices used to measure program approval criteria are reliable and support their use in assessing stakeholder views of their preparation programs in general and, more specifically, their views on their coursework, field-based experiences, supervision and assessment.

7.2. Professional Standards for Teaching Instruments (42-item).

The Rasch construct validity framework (Wolfe & Smith, 2007a, 2007b) established validity evidence to justify the use of the teacher readiness PST scores for each of five stakeholder groups (NTCD, NTCP, TRCD, TRCP, and SPCD). Evidence for five validity aspects (content, substantive, generalizability, structural and external) support the 42-item instruments' construct validity. The results of the technical quality, rating scale functioning and the item hierarchies of these analyses show that, regardless of the stakeholder group, the 42-item instruments exhibited excellent model fit, good rating scale functioning and item hierarchies (between and within subscales) that conform to theoretical expectations.

The results of the generalizability analyses indicated that items were invariant between stakeholder groups and across sub-groups (no DIF was present when respondents attending public institutions were compared to those from private institutions). Similarly, evidence from two different methodologies (Rasch and CFA), justify the claim that the 42 items form a unidimensional scale composed of four sub-dimensions. The signal-to-noise ratio of these four sub-dimensions is sufficient to warrant reporting four sub-dimensions scores but the strong correlation between the person measures of the four standards, and between the four first-order factors and the second-order teacher readiness factor support that that the underlying construct they measure is teacher readiness. The PSTs psychometric data reported here support the formation of a reliable, internally consistent, responsive unidimensional scale to measure teacher readiness across multiple stakeholder groups.

The replication of the results across the five instruments for each aspect of construct validity provides particularly strong evidence that the teacher readiness construct is being measured appropriately. Replication was evident in the technical quality of the items, the rating scale functioning, the DIF and structural validity analyses, and in the, albeit limited, external validity analyses.

Overall, the magnitude of the scale and subscale reliabilities of the five 42-item surveys support the generalizability of score meaning designed to measure the teacher effectiveness construct. The stability and replicability of reliabilities across forms provides further supporting evidence that the items developed are generalizable and representative of the teacher readiness construct.

7.3. Hiring Principal Instruments (6-item).

The validity evidence for the 6-item hiring principal surveys is mixed. The item technical quality is good across the two instruments (NTHP and TRHP). The rating scale structures appear to perform well but the distance between some thresholds is above 5 logits for both surveys suggesting that more response options are required to help differentiate respondent views. The DIF results comparing candidates or completers from public and private were excellent for both instruments, thereby supporting the invariance of items across sub-groups. However, there is only a moderate correlation (<0.5) between the average item parameter estimates of the new teacher and teacher-of-record surveys. This suggests that despite the same items being used, the different stem and response options provided to the hiring principals have impacted the parameter estimates of the two surveys. The NTHP instrument is responsive (five person

strata) supporting the external validity aspect of construct validity. In contrast, the TRHP instrument is less responsive but still adequate for low stakes decisions (3 person strata).

Overall, there is sufficient validity evidence to support the continued use of the hiring principal surveys, although improvement to the TRHP, in particular, is warranted.

7.4. Limitations and Recommendations

7.4.1. Study Limitations

The major limitation of the technical data within this report is related to the small sample size for some of the surveys. ESE could not perform some of the analyses due this limitation. For example, ESE did not perform the CFA for the teacher-of-record surveys. Similarly, the DIF analyses that compared item deltas across the size of the preparation provider institutions were not performed. These analyses would have provided more comprehensive validity data to support the use of the survey scores. Although each survey was administered to a census of stakeholders, the respondents were not required to participate. As a result, the data shown in this report are not likely representative of the institutions and, as a result, any data received by the preparation providers should be treated with caution. The average number of respondents per institution was also too low to perform any external validity analyses at this level.

7.4.2. Survey and Item-Level Recommendations

a. Given the strong correlation between item deltas of the new teacher and teacher-of-record items and the replicability of the results overall, the program office should consider merging the two sets of surveys and survey results. A study should investigate the feasibility of this. Merging the two surveys into one will improve the number of respondents by institution and allow for more reliable institution-level comparisons and analyses.
b. The hiring principal survey should be increased to an eight-item survey and include an item from Standard III. This should improve the variance of the scale, particularly the variance of the items for the teacher-of-record survey where ceiling effects were evident. A six-response option Likert scale should also be considered due to the large distance between category thresholds.
c. At this time, it is not recommended reducing the number of response options on the 42-item surveys to four. This could impact the variance and responsiveness of the scales. In addition, the limited sample sizes for some of the surveys could have impacted the estimation of threshold parameters.
d. The PSTs items should be reviewed to determine if more difficult but relevant teacher practices could be added to better capture and assess teacher readiness for respondents' at the upper end of the scoring distribution. This process was undertaken between the pilots and this administration of the surveys; it is doubtful whether new items can be developed to assess respondents with very favorable views of their teacher readiness. One avenue to explore is to develop more items related to social and emotional learning.
e. The number of items for some topics (elements) could be reduced. For example, in Standard I, there are six items related to assessment, and in Standard II, there are five items related to classroom management. Although the items from both elements form hierarchies in terms of item difficulty, the representation of these two elements could be reduced in order to lower the overall number of items on the surveys. The program office should determine if they want to report separate subscales for these topics; if they do, the number of items should remain the same. If not, the number of items in each topic could be reduced to four thereby reducing the total number of items measuring Standard II by three items.
f. Items that are redundant could be removed. The Rasch analyses point to some items that may have redundant content. For example, items 2.7, 2.13, and 2.18 are all related to providing students with a cognitively

demanding learning environment and their item difficulties are similarly located on the scale metric axis. One or two of these items could be removed.

References

Andrich, D. (1978a). Application of psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2* (4), 581-594.

Andrich, D. (1978b). Rating formulation for ordered response categories. *Psychometrika*, *43* (4), 561-573.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. (2nd Ed). New Jersey: Lawrence Erlbaum Associates.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*, 253-269.

Boone, W. J., Staver, J. R., and Yale, M. S. (2014). *Rasch analysis in the human sciences*, New York: Springer.

Byrne, B. M. (2012). *Structural Equation Modeling with MPLUS: Basic concepts, applications and programming.* New York, New York: Routledge Taylor & Francis Group.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral and health sciences.* Routledge Taylor & Francis Group, New York, New York.

Ferguson, R. (2010). *Student perceptions of teaching effectiveness: Discussion brief.* National Center for Teacher Effectiveness and the Achievement Gap Initiative, Cambridge, Massachusetts: Harvard University.

Gable, R.K., Ludlow, L.H. & Wolf, M.B. (1990). The Use of Classical and Rasch Latent Trait Models to Enhance the Validity of Affective Measures. *Educational and Psychological Measurement*, *50* (4), 869-878.

Hambleton, R. K. & Jones, R. W. (1993). An NCME instructional model on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, Fall, 38-47.

Linacre, J. M. (2014a). WINSTEPS [Computer program, v3.81.0]. Chicago: MESA Press.

Linacre, J.M. (2014b). *A user's guide to Winsteps Rasch-model computer programs.* Program manual 3.81.0. Chicago, US: MESA Press.

Ludlow, L. H. & Haley, S. M. (1995). Rasch model logits: Interpretation, use and transformation. *Educational and Psychological Measurement*, *55* (6), 967-975.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012-1027.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50* (9), 741-749.

Muthén, L. K. & Muthén, B. O. (2010). *MPLUS, (Computer program 6.11)*. Los Angeles, US: Muthén & Muthén.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press).

Schumacker, R. E. & Lomax, R. G. (2004). *A beginners guide to structural equation modeling*. (2nd Ed). New Jersey: Lawrence Erlbaum Associates.

Schumacker, R. E. & Smith, E. V. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement*, *67* (3), 394-409.

Sinnema, C. E. L. & Ludlow, L. H. (2013). A Rasch approach to the measurement of responsive curriculum practice in the context of curricula reform. *The International Journal of Educational and Psychological Assessment, 12* (2), 33-55.

Smith, E. V. Jr. (2000). Metric Development and Score Reporting in Rasch Measurement. *Journal of Applied Measurement*, *1*(3), 303-326.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205-231.

Wolfe, E. W., & Smith, E. V. Jr. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I – Instrument development tools. *Journal of Applied Measurement*, *8*(1), 97-123.

Wolfe, E. W. & Smith Jr., E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. *Journal of Applied Measurement*, *8* (2), 204-234.

APPENDICES

A-1: Respondent profile by preparation provider characteristics 28
B-1: Rasch-based instrument validity framework summary 29
C-1: Confirmatory factor analyses 30
C-2: Guide for evaluating Rasch model and CFA validity data 31
C-3: Rasch Rating Scale Model 32
D-1: New Teacher Item Technical Quality (Item Fit Statistics) 33
D-2: Teacher-of-Record Item Technical Quality (Item Fit Statistics) 34
D-3: Hiring Principal Item Technical Quality (Item Fit Statistics) 35
E-1: New Teacher Candidate (NTCD) Rating Scale Structure 36
E-2: Teacher-of-Record Candidate (TRCD) Rating Scale Function 37
E-3: New Teacher Completer (NTCP) Rating Scale Function 38
E-4: Teacher-of-Record Completer (TRCP) Rating Scale Function 39
E-5: Supervising Practitioner Teacher Candidate (SPCD) Rating Scale Function 40
E-6: New Teacher Hiring Principal (NTHP) Rating Scale Function 41
E-7: Teacher-of Record-Hiring Principal (TRHP) Rating Scale Function 42
E-8: New Teacher Candidate (NTCD) Item Variable Map 43
E-9: New Teacher Completer (NTCP) Item Variable Map 44
E-10: Supervising Practitioner Candidate (SPCD) Item Variable Map 45
E-11: Assessment Item Hierarchy (Standard I) 46
E-12: Classroom Management Item Hierarchy (Standard II) 46
F1: Descriptive Data and Real Person Separation Reliabilities (PSR) for Performance Standard I 47
F2: Descriptive Data and Real Person Separation Reliabilities (PSR) for Performance Standard II 47
F3: Descriptive Data and Real Person Separation Reliabilities (PSR) for Performance Standard III 48
F4: Descriptive Data and Real Person Separation Reliabilities (PSR) for Performance Standard IV 48
F-5: New Teacher Candidate (NTCD) Differential Item Functioning 49
F-6: Teacher-of-Record Candidate (TRCD) Differential Item Functioning 50
F-7: New Teacher Completer (NTCP) Differential Item Functioning 51
F-8: Teacher-of-Record Completer (TRCP) Differential Item Functioning 52
F-9: Supervising Practitioner Candidate (SPCD) Differential Item Functioning 53
F-10: New Teacher Hiring Principal Candidate (NTHP) Differential Item Functioning 54
F-11: Teacher-of-Record Hiring Principal (NTHP) Differential Item Functioning 55
G-1: Performance Standard I Item-Level Invariance across 42-Item Surveys 56
G-2: Performance Standard II Item-Level Invariance across 42-Item Surveys 57
G-3: Performance Standard III Item-Level Invariance across 42-Item Surveys 58
G-4: Performance Standard IV Item-Level Invariance across 42-Item Surveys 59
G-5: Item-Level Invariance across 6-Item Hiring Principal Surveys 60
G-6: Item Invariance Correlations across Hiring Principal Surveys 61
H-1: Sub-Scale Correlations for the Four Performance Standards of Teaching (PST) 62
J-1: Confirmatory Factor Analyses Fit Statistics 63
K-1: New Teacher Candidate (NTCD) Non-Extreme Person Measure Distribution 64
K-2: Preparation Provider Aggregate NTCD Person Measure Distribution 64
K-3: New Teacher Completer (NTCP) Non-Extreme Person Measure Distribution 65
K-4: Preparation Provider Aggregate NTCP Person Measure Distribution 65

K-5: Supervising Practitioner Candidate (SPCD) Non-Extreme Person Measure Distribution 66
K-6: Preparation Provider Aggregate SPCD Person Measure Distribution 66
K-7: New Teacher Hiring Principal (NTHP) Non-Extreme Person Measure Distribution 67
K-8: Preparation Provider Aggregate NTHP Person Measure Distribution 67

Appendix A-1: Respondent Profile by Preparation Provider Type and Size

Percent of Respondents		Type of Institution			Size of Institution		
Survey	Number of Institutions	Attending a Public School	Attending a Nationally Accredited Provider	Attending a Traditional School[1]	Small (0 – 29 students)	Medium (30 – 99 students)	Large (100+ students)
NTCD	52	35.0	51.2	95.8	12.0	16.4	71.6
TRCD	41	29.1	43.6	85.5	7.3	21.8	70.7
SP	51	42.7	53.5	95.1	14.0	16.3	69.6
NTCP	48	33.2	45.8	91.6	12.4	19.9	67.7
TRCP	43	29.6	40.3	81.7	9.7	18.8	71.5
NTHP	57	40.0	52.7	93.6	9.6	13.5	76.9
TRHP	45	35.0	45.7	87.3	5.8	15.4	78.8

[1] A Traditional preparation provider is classified as a degree-granting Institution of Higher Education (HE); an Alternative preparation provider is an organization, other than an HE institution, that prepares individuals for educator licensure (e.g., a school district, educational collaborative etc.).

Appendix B-1:

Rasch-Based Instrument Validity Framework and Evidence Collected for Performance Standards Scales[1]

	Validity Aspect		
	Content	Substantive	Generalizability
Evidence	Instrument Purpose	Rating Scale Functioning	Differential Item Functioning
	Test Specification	Item Difficulty Hierarchy	
	Expert Reviews[2]		Person Separation Reliability
	Item Technical Quality		Item Invariance

	Validity Aspect		
	Structural[3]	External	Consequential[5]
Evidence	Rasch Dimensionality Analyses	Responsiveness	*Standard Setting Score Use*
		Sub-scale correlations	
		Relationship between Performance standard scaled-scores with scores from similar/dissimilar constructs[4]	

[1] Based on: Messick (1995) and Wolfe and Smith (2007b) conceptualization and representation.

[2] Experts reviewed items for: item representativeness (did they measure the standard it was designed to measure?), accessibility (would respondents understand it?), actionability (would preparation providers be able to use the information?), and would the items measured form a continuum of item difficulty.

[3] Confirmatory factor analyses (CFA) were performed on performance standards' items related to three surveys to provide additional structural validity evidence. CFA is founded on classical theory methodology

[4] This aspect of external validity is beyond the scope of this study.

[5] Consequential validity is beyond the scope of this study.

Appendix C-1: Confirmatory Factor Analyses

In confirmatory factor analyses (CFA), items that share variance and covariances are hypothesized to form a dimension. These analyses are based on ordinal measures (Likert responses); as a result, the CFA use the correlation matrix structure to assess the relationship between the variables of the model. These polychoric correlations are used to determine if the relationships between the observed variables are explained by an underlying continuous latent variable (Byrne, 2012). Because the data are ordinal in nature, a Weighted Least Square Mean Variance (WLSMV) estimator is the default estimator for producing parameter estimates. This estimator corrects for the categorical nature of the data and the likely violation of the normality assumption required of CFA (Byrne, 2012). As a result, the chi-square statistic is scaled to take these factors into account and this scaled value is used in the CFA DIFFTEST by MPLUS to determine if the model fit between two nested models is significantly different.

Appendix C-2: Guide for evaluating Rasch Model validity data

Validity Aspect	Statistic/Data	Cutoff Criteria or Typical Standard	Comment
Content	Point-to-measure Correlation	Positive and >0.3.	Analog to CTT item-total correlation.
Content & Structural	Infit and Outfit Mean-square Fit Statistics (MNSQ)	• 0.7 – 1.3 • Disruption of pattern in magnitude of misfit.	Mean square errors should have a mean of one i.e. (observed = expected).
Substantive	Rating Scale Functioning	• Minimum of 10 responses per category. • Categories are unimodal. • Observed score averages and item threshold parameters increase monotonically. • Un-weighted MNSQ < 2.0 for ea. category.	Scale is being used according to the intent of instrument developers – supports score use and inferences.
	Item Difficulty Hierarchy	• Ordering of item deltas correspond to theoretical expectations. • Item/person variable maps.	Qualitative assessment of items in the construct and/or standards.
Generaliz-ability	Item Invariance and Differential Item Functioning (DIF)	• Within standard error, items should retain same item difficulty (deltas) across administrations and survey forms (correlation of greater than 0.9). • For DIF, recommended criteria vary: delta difference of 0.3 – 0.64 Logits (0.5 used in study)	DIF flags items that need further review. Items may need revision to eliminate bias or removal when estimating scores if bias is significant.
	Person Separation Reliability (PSR)	• Typical ~ 0.8; High Stakes > 0.9 • 0.8 Standards 0.9 Construct	PSR is similar to Cronbach α and ranges from 0 to 1.
Structural	Sub-scale Correlations	• Positive and substantive (> 0.5 but < 0.9)	The items that form a 2nd dimension should be reviewed qualitatively to determine their commonality and if their co-variation is meaningful.
	Standardized Residuals	• No correlation between residuals from separate calibrations of two item subsets.	
	Winsteps Software (PCA: Principal Component Analyses of Residuals).	• Total variance explained: >40% very good; >50% excellent • 2nd dimension: < 5% of total variance. • 2nd dimension Eigen < 3 • 1st contrast variance 4x variance of 2nd contrast • Cluster correlations ♦ > 0.82 likely only one latent trait ♦ > 0.71 more dependency than independence	
	Confirmatory Factor Analyses (CFA)	• RMSEA < 0.05 excellent; <0.08 good • CFI and TLI > 0.95 excellent ; • CFI and TLI > 0.90 good • Factor Loadings > 0.5	CFA can be used in conjunction with Rasch analyses. The results are comparable and CFA is used to support the results of the Rasch model.
External	Responsiveness	• Typical ~ 3 person strata (low, medium, high). • Teacher readiness measures need a minimum of 4 person strata. • H = (4G +1)/3 where H is the number of person strata and G is the person separation index.	Instruments that are responsive can better differentiate high and low scorers by reliably separating individuals into a greater number of performance levels, thereby facilitating the measurement of change of respondent views on a construct.

Appendix C-3: Rasch Rating Scale Model

The Rasch model uses an exponential transformation to place ordinal Likert responses on to an equal-interval logit scale (Rasch, 1960). This transformation ensures that stakeholder perceptions are measured appropriately and that the data meet the assumptions of parametric testing (Ludlow and Haley, 1995; Boone, 2014). In addition, the sample independence features of the Rasch model overcome the fundamental drawbacks of classical test theory (CTT) analyses Smith (2000). In CTT, the difficulty of a test is sample dependent, making it problematic to measure change on a variable (Smith, 2000; Boone & Scantlebury, 2006). In contrast, the Rasch property of item invariance implies that the relative endorsements and location of the items do not change (within measurement error), or are independent of the sample responding; in kind, the relative item endorsements should behave as expected across different samples (Smith, 2002, Engelhard, 2013). When items are invariant, the Rasch model is particularly discerning in differentiating between high and low scorers (Gable, Ludlow, and Wolf, 1990; Sinnema & Ludlow, 2013) on a measurement scale as it places persons and items on a common scale metric (Hambleton and Jones, 1993; Engelhard, 2013).

The Rasch rating scale model provides a mathematical model for the probabilistic relationship between a person's ability (β_n) and the difficulty of items (δ_i) on a test or survey. Andrich's (1978a, 1978b) rating scale model (RSM) used in this study is defined in Equation 1.

$$\emptyset_{nij} = \frac{exp(\beta_n - (\delta_i + \tau_j))}{\left(1 + exp(\beta_n - (\delta_i + \tau_j))\right)}, \qquad j = 1, 2, \ldots, m_i. \qquad (1)$$

Where $\emptyset_{nij}$ is the "conditional probability of person, *n* responding in category *j* to item *i*". Tau is the estimate of the location of the *jth* step for each item relative to that item's scale value (δ_i). The number of response categories is equal to m_i +1 where m_i is the number of thresholds. In the RSM, moving from one threshold to the next contiguous threshold is assumed to have the same mean difference across all items of the survey. The unit of measurement resulting from the natural log transformation of person responses results in separate ability and item difficulty estimates called logits (Ludlow & Haley, 1995). The persons and items are placed on a common continuum (the scale metric axis of the variable map) and as such, the persons can be characterized by their location on the continuum by the types and level of items of which they are associated. By taking the natural log of the odds ratio, stable replicable information about the relative strengths of persons and items is derived with equal differences in logits translating into equal differences in the probability of endorsing an item no matter where on the scale metric an item is located; this interval-level unit of measurement is a fundamental assumption of parametric tests (Boone, Townsend, and Staver, 2011). By default, in WINSTEPS, the item mean summed across the thresholds equals zero; the person and item measures are generated and reported on the logit scale. In the context of this study, a respondent with a positive logit value on an educator preparation survey feels relatively more positive about the program than a respondent with a negative logit value.

Appendix D-1: New Teacher Item Technical Quality (Item Fit Statistics)

Standard	New Teacher Candidate (NTCD) MNSQ: N=408			New Teacher Completer (NTCP) MNSQ: N=202			Supervising Practitioner (SPCD) MNSQ: N=649		
Item	Infit[1]	Outfit[2]	PTM[3]	Infit[1]	Outfit[2]	PTM[3]	Infit[1]	Outfit[2]	PTM[3]
1.1	1.19	1.35	0.56	1.35	1.62	0.56	1.25	1.26	0.63
1.2	1.20	1.46	0.55	1.07	0.98	0.61	1.09	1.05	0.65
1.3	1.02	1.03	0.60	0.99	1.14	0.62	0.89	0.81	0.70
1.4	1.07	1.05	0.60	1.25	1.30	0.58	1.06	1.06	0.68
1.5	1.01	0.93	0.56	1.27	1.17	0.56	1.24	1.12	0.62
1.6	0.94	0.93	0.65	1.02	0.96	0.65	0.94	0.89	0.70
1.7	1.12	1.11	0.60	1.13	1.02	0.63	1.10	1.05	0.68
1.8	0.87	0.76	0.63	0.92	0.86	0.65	0.94	0.88	0.71
1.9	1.29	1.39	0.60	1.19	1.36	0.64	1.14	1.20	0.67
1.10	1.05	0.98	0.64	1.07	1.14	0.66	0.95	0.93	0.70
1.11	0.98	0.94	0.64	0.71	0.86	0.68	0.95	0.91	0.69
1.12	0.94	1.02	0.66	0.92	0.98	0.67	1.04	1.05	0.68
1.13	1.29	1.38	0.62	1.24	1.29	0.66	1.50	1.57	0.61
2.1	0.83	0.88	0.62	0.90	0.78	0.65	0.89	0.86	0.66
2.2	1.01	1.00	0.65	1.26	1.25	0.62	1.08	1.06	0.67
2.3	0.87	0.87	0.64	1.03	0.95	0.67	1.07	0.97	0.69
2.4	0.87	0.84	0.63	0.70	0.69	0.69	0.79	0.73	0.72
2.5	0.93	0.96	0.67	0.77	0.73	0.72	0.94	0.93	0.70
2.6	0.96	0.91	0.60	0.83	0.71	0.66	1.01	0.99	0.67
2.7	0.90	0.92	0.64	0.68	0.64	0.73	0.82	0.80	0.73
2.8	1.36	1.40	0.56	1.41	1.18	0.60	1.20	1.20	0.67
2.9	0.83	0.82	0.63	0.84	0.81	0.67	0.84	0.80	0.71
2.10	1.04	1.11	0.64	0.97	0.90	0.68	1.06	1.07	0.67
2.11	0.89	0.81	0.60	0.82	0.73	0.67	0.78	0.74	0.71
2.12	0.95	1.09	0.66	1.30	1.22	0.65	1.14	1.17	0.67
2.13	0.75	0.77	0.66	0.93	1.10	0.66	0.87	0.83	0.71
2.14	0.85	0.90	0.59	1.00	0.91	0.62	0.95	0.96	0.68
2.15	0.89	0.81	0.64	0.84	0.82	0.66	1.03	0.96	0.69
2.16	0.82	0.75	0.66	0.70	0.75	0.70	0.62	0.61	0.76
2.17	0.77	0.88	0.64	0.77	0.67	0.68	0.71	0.67	0.73
2.18	0.66	0.63	0.69	0.73	0.67	0.71	0.68	0.65	0.73
3.1	1.08	1.18	0.67	1.10	1.21	0.67	0.98	1.05	0.69
3.2	1.22	1.31	0.62	0.94	1.02	0.67	1.09	1.17	0.65
3.3	1.13	1.28	0.66	1.05	1.30	0.68	0.98	1.22	0.68
3.4	1.13	1.16	0.64	1.06	1.16	0.66	1.11	1.25	0.63
3.5	1.05	1.10	0.65	0.74	0.80	0.70	1.10	1.21	0.65
3.6	1.07	1.18	0.63	1.09	1.30	0.65	0.97	1.08	0.67
4.1	1.41	1.29	0.49	1.11	1.00	0.53	1.15	1.19	0.59
4.2	1.05	0.91	0.58	1.08	1.01	0.60	1.20	1.20	0.60
4.3	1.09	1.06	0.52	1.15	1.06	0.50	1.19	1.21	0.58
4.4	0.95	0.81	0.60	0.95	0.82	0.64	1.00	0.96	0.64
4.5	0.94	0.87	0.56	0.84	0.78	0.59	1.26	1.24	0.57

[1]Infit mean square is the average of the standardized residuals weighted by the individual item and person variances.
[2]Outfit mean square is the average of the standardized residual variance and is un-weighted.
[3]PTM: Point to Measure Correlation

Appendix D-2: Teacher-of-Record Item Technical Quality (Item Fit Statistics)

Standard	Teacher-of-Record Candidate (TRCD) MNSQ: N=179			Teacher-of-Record Completer (TRCP) MNSQ: N=186		
Item	Infit[1]	Outfit[2]	PTM[3]	Infit[1]	Outfit[2]	PTM[3]
1.1	1.09	1.13	0.69	1.27	1.42	0.62
1.2	1.31	1.22	0.69	0.77	0.99	0.68
1.3	0.95	0.87	0.75	0.88	0.94	0.69
1.4	1.02	1.04	0.71	1.04	1.00	0.66
1.5	0.84	0.71	0.74	0.92	0.82	0.65
1.6	0.97	0.95	0.74	1.29	1.42	0.62
1.7	0.96	0.91	0.75	1.10	1.14	0.66
1.8	0.76	0.72	0.75	0.81	0.76	0.69
1.9	1.55	1.76	0.70	1.41	1.60	0.63
1.10	1.25	1.31	0.69	0.88	0.86	0.71
1.11	0.81	0.80	0.78	0.87	0.92	0.71
1.12	1.26	1.26	0.72	1.04	1.16	0.68
1.13	1.56	1.68	0.70	1.16	1.29	0.71
2.1	0.81	0.81	0.75	0.80	0.73	0.70
2.2	1.07	1.13	0.73	1.21	1.18	0.67
2.3	1.07	1.03	0.73	1.21	1.23	0.67
2.4	0.71	0.69	0.77	0.85	0.81	0.71
2.5	1.06	1.07	0.74	0.98	0.99	0.71
2.6	1.02	0.97	0.71	0.92	0.85	0.68
2.7	0.76	0.81	0.75	0.71	0.75	0.73
2.8	1.06	1.19	0.71	1.41	1.35	0.66
2.9	0.77	0.71	0.78	0.74	0.73	0.73
2.10	1.34	1.40	0.71	1.13	1.19	0.69
2.11	0.72	0.72	0.77	0.83	0.78	0.71
2.12	1.26	1.65	0.70	1.31	1.29	0.68
2.13	1.08	1.17	0.71	0.87	0.89	0.69
2.14	0.76	0.72	0.75	0.99	0.94	0.67
2.15	0.65	0.70	0.77	0.74	0.79	0.71
2.16	0.54	0.52	0.78	0.65	0.67	0.74
2.17	0.57	0.52	0.78	0.77	0.73	0.70
2.18	0.62	0.57	0.79	0.73	0.72	0.73
3.1	1.05	1.16	0.74	0.93	1.07	0.74
3.2	0.99	1.04	0.73	1.04	1.03	0.71
3.3	1.29	1.46	0.72	1.13	1.32	0.70
3.4	1.01	1.04	0.74	1.05	1.05	0.71
3.5	1.09	1.12	0.73	1.07	1.20	0.70
3.6	0.94	0.96	0.75	1.03	1.04	0.70
4.1	1.04	0.93	0.71	1.11	1.04	0.62
4.2	1.04	1.03	0.72	1.18	1.07	0.64
4.3	1.00	1.00	0.70	1.02	1.08	0.63
4.4	0.81	0.97	0.70	0.89	0.86	0.71
4.5	1.01	1.38	0.69	0.86	0.74	0.67

[1]Infit mean square is the average of the standardized residuals weighted by the individual item and person variances.
[2]Outfit mean square is the average of the standardized residual variance and is un-weighted.
[3]PTM: Point to Measure Correlation

Appendix D-3: Hiring Principal Item Technical Quality (Item Fit Statistics)

Standard	New Teacher Hiring Principal MNSQ: N=628			Teacher-of-Record Hiring Principal MNSQ: N=444		
Item	Infit[1]	Outfit[2]	PTM[3]	Infit[1]	Outfit[2]	PTM[3]
1.1	1.12	1.05	0.90	0.95	0.83	0.88
1.2	0.83	0.78	0.92	1.05	0.97	0.87
2.1	0.94	0.86	0.93	0.97	0.95	0.88
2.2	0.96	0.92	0.92	0.94	0.81	0.88
2.3	0.67	0.60	0.94	0.93	0.82	0.89
4.1	1.38	1.47	0.89	1.10	0.98	0.88

[1]Infit mean square is the average of the standardized residuals weighted by the individual item and person variances.
[2]Outfit mean square is the average of the standardized residual variance and is un-weighted.
[3]PTM: Point to Measure Correlation

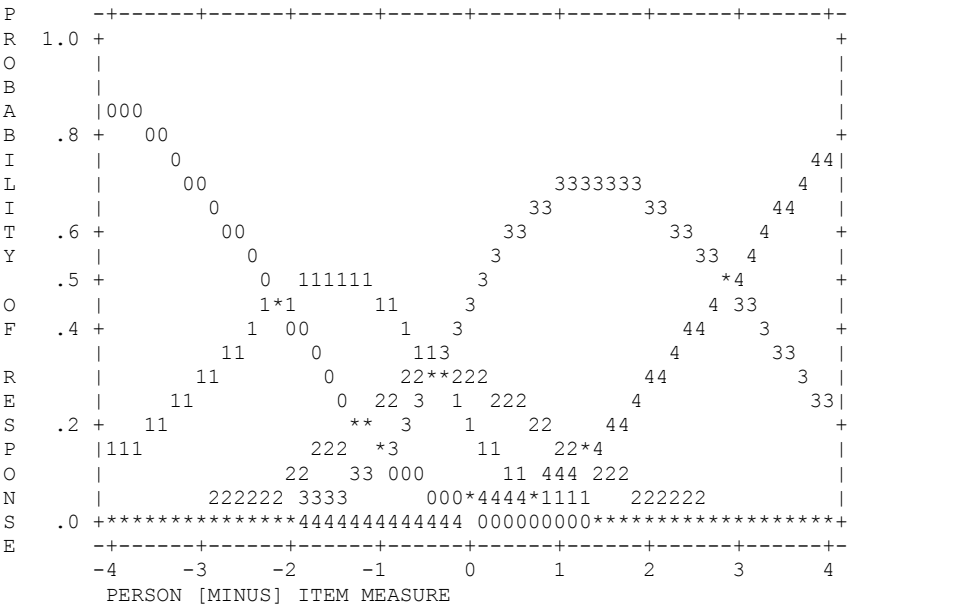
Appendix E: Rating Scale Function

Appendix E-1: New Teacher Candidate (NTCD) Rating Scale Function

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
0	0	102	1	-.51	-.49	.99	1.05	NONE	(-3.34)	0
1	1	681	4	.27	.06	1.19	1.39	-2.12	-1.49	1
2	2	1441	8	.77	.77	1.02	1.08	-.35	-.28	2
3	3	7490	44	1.73	1.79	.97	.89	-.40	1.39	3
4	4	7308	43	3.66	3.61	.94	.95	2.87	(4.00)	4
MISSING		72	0	3.14						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

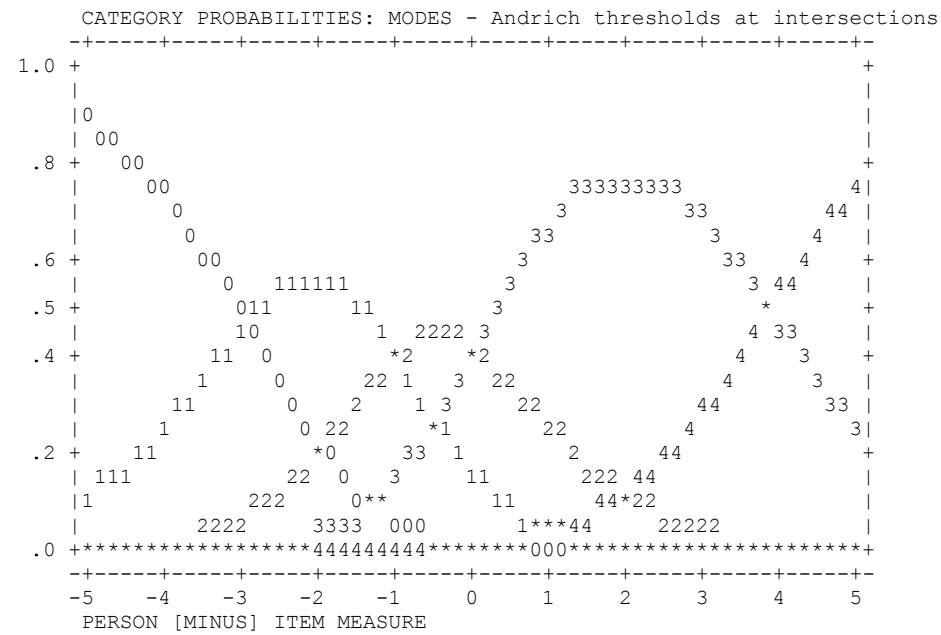


Appendix E-2: Teacher-of-Record Candidate (TRCD) Rating Scale Function

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
0	0	105	1	-3.68	-3.56	1.00	1.19	NONE	(-4.13)	0
1	1	295	4	-.29	-.63	1.23	1.39	-2.93	-2.07	1
2	2	834	11	.72	.72	1.09	1.13	-.98	-.41	2
3	3	3531	47	2.24	2.31	.93	.93	.03	2.00	3
4	4	2734	36	4.60	4.51	.89	.90	3.88	(4.99)	4
MISSING		103	1	1.43						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

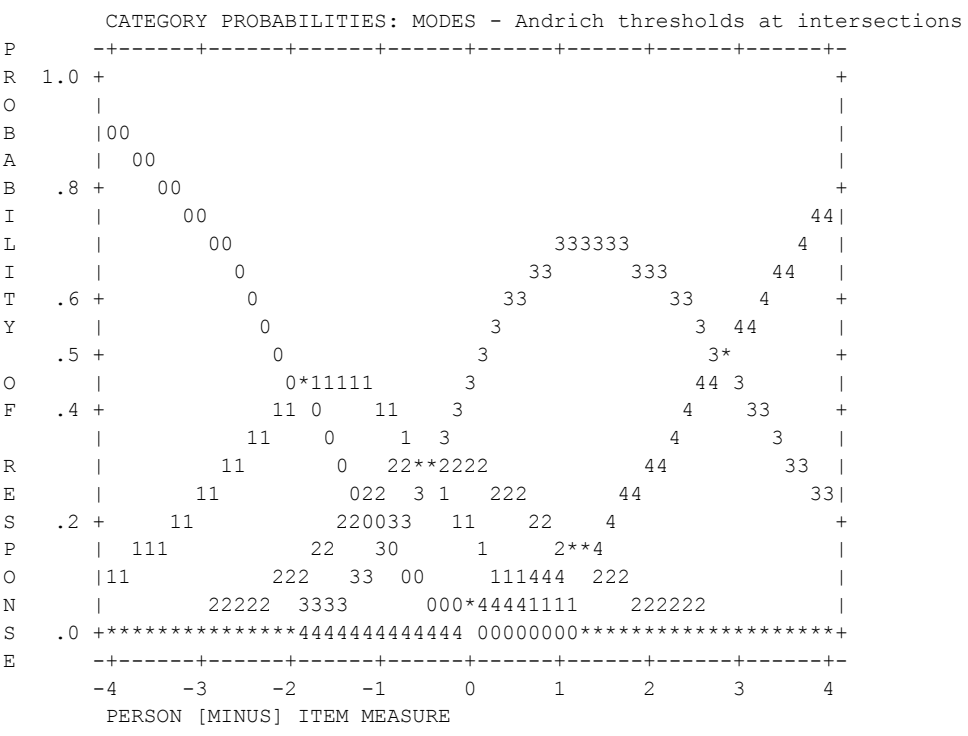


Appendix E-3: New Teacher Completer (NTCP) Rating Scale Function

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
0	0	137	2	-1.18	-1.09	.92	1.10	NONE	(-3.13)	0
1	1	444	5	-.19	-.28	1.05	1.10	-1.86	-1.47	1
2	2	865	10	.58	.57	1.04	1.09	-.53	-.34	2
3	3	3811	45	1.54	1.57	.95	.88	-.44	1.33	3
4	4	3173	38	3.50	3.47	1.00	.97	2.82	(3.95)	4
MISSING		54	1	3.00						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

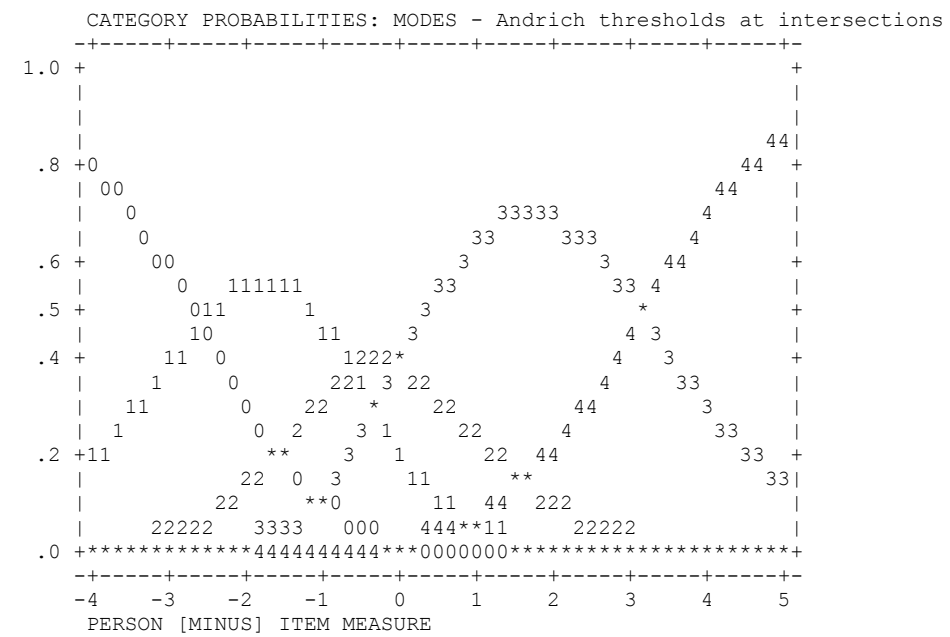


Appendix E-4: Teacher-of-Record Completer (TRCP) Rating Scale Function

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
0	0	109	1	-.91	-1.10	1.17	1.42	NONE	(-3.78)	0
1	1	450	6	-.21	-.25	1.05	1.15	-2.58	-1.73	1
2	2	1001	13	.70	.71	.98	.99	-.58	-.24	2
3	3	3501	45	1.80	1.82	.95	.94	-.01	1.68	3
4	4	2686	35	3.50	3.46	.98	.96	3.17	(4.30)	4
MISSING		65	1	1.90						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

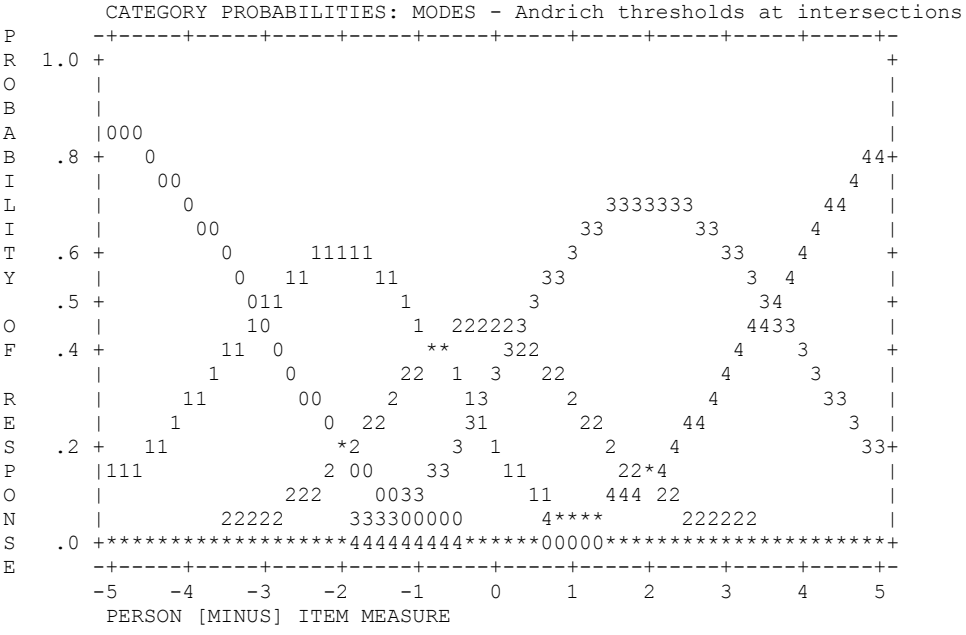


Appendix E-5: Supervising Practitioner Teacher Candidate (SPCD) Rating Scale Function

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
0	0	274	1	-1.86	-2.09	1.24	1.36	NONE	(-4.24)	0
1	1	1632	6	-.20	-.36	1.14	1.19	-3.07	-2.02	1
2	2	4441	16	.72	.78	1.02	1.07	-.78	-.22	2
3	3	13390	49	1.98	2.02	.94	.91	.27	1.99	3
4	4	7521	28	3.92	3.85	.96	.94	3.58	(4.71)	4

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

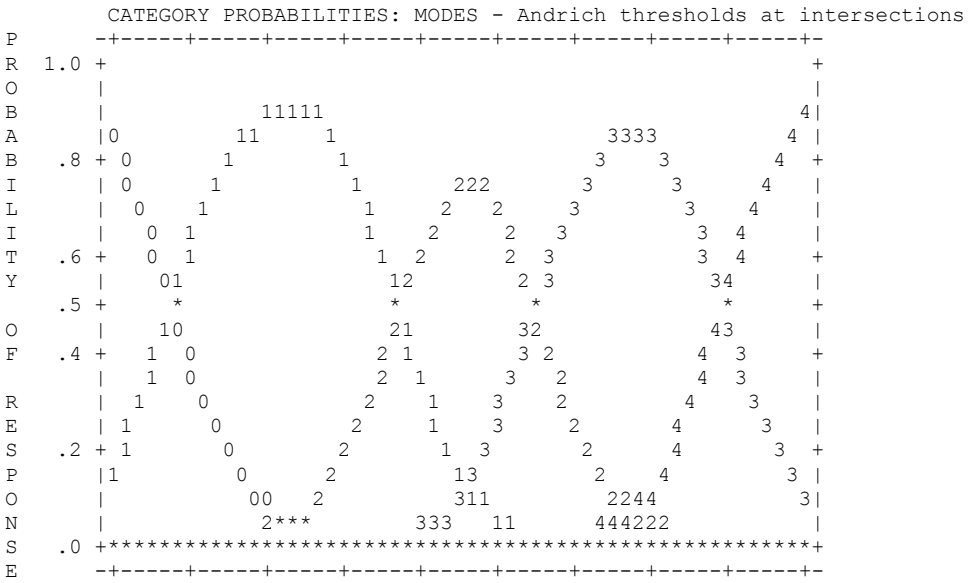


Appendix E-6: New Teacher Hiring Principal (NTHP) Rating Scale Function

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
0	0	410	11	-7.41	-7.46	1.05	.84	NONE	(-8.52)	0
1	1	1091	29	-3.64	-3.61	1.00	1.01	-7.42	-4.46	1
2	2	980	26	.16	.16	.87	.79	-1.50	.25	2
3	3	947	25	3.86	3.82	1.00	1.09	2.00	4.46	3
4	4	314	8	6.54	6.65	1.18	1.09	6.92	(8.02)	4
MISSING		14	0	-.16						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

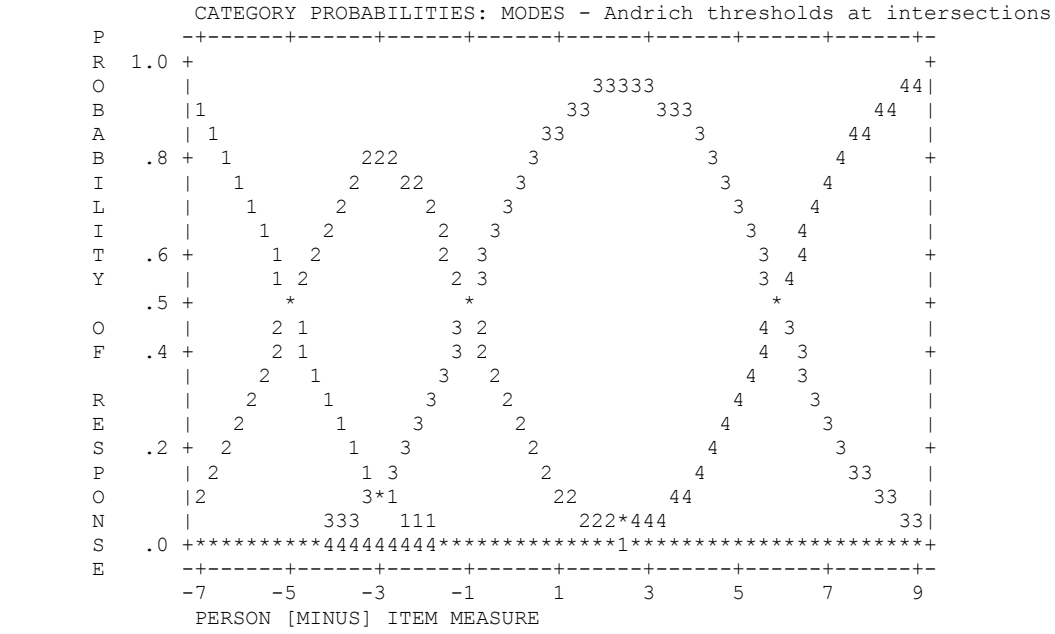


Appendix E-7: Teacher-of Record-Hiring Principal (TRHP) Rating Scale Function

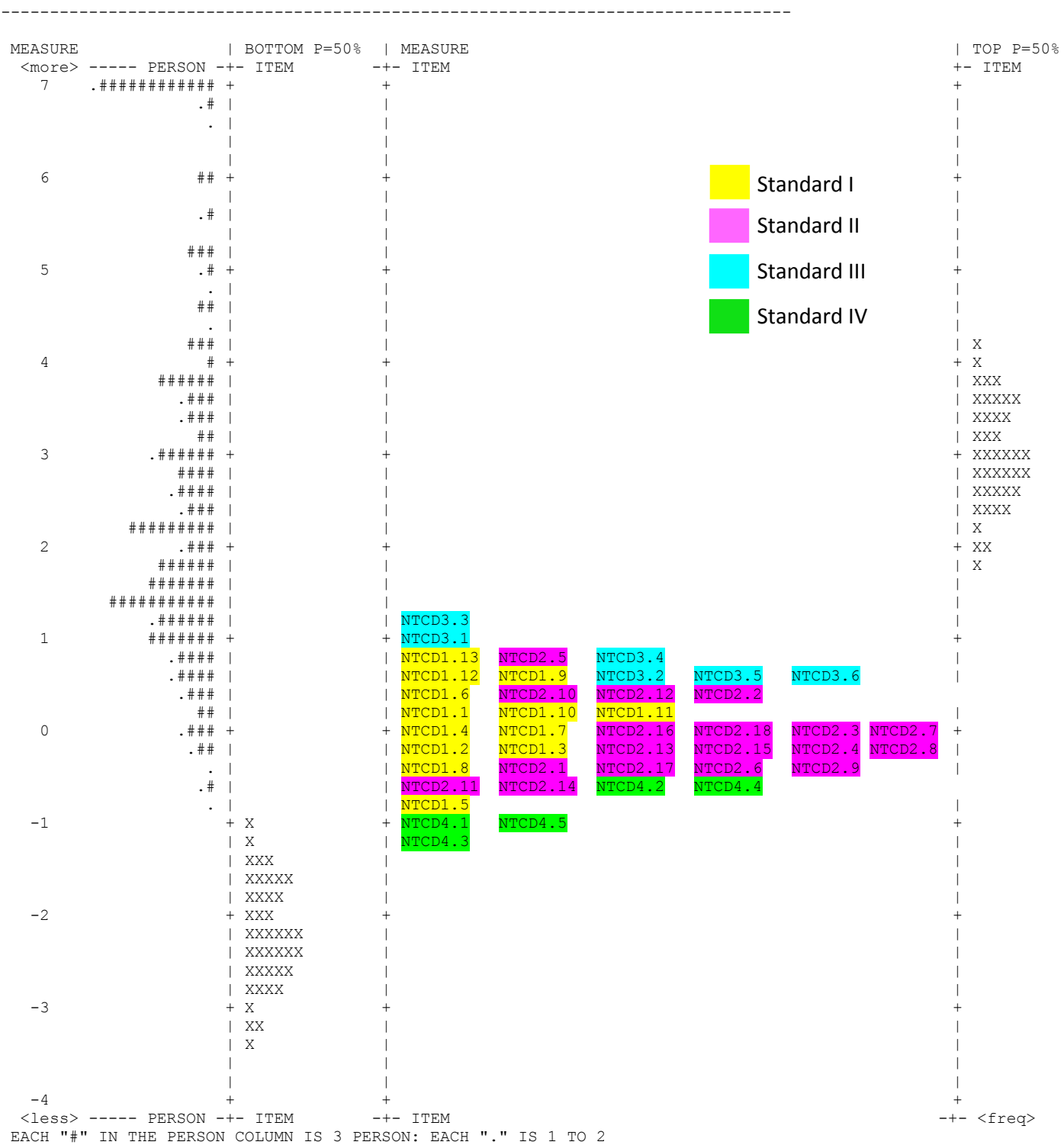
SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
0		18	1					NONE		0
1	1	51	2	-4.34	-4.61	1.18	1.10	NONE	(-6.07)	1
2	2	128	5	-1.54	-1.45	1.02	.80	-4.96	-2.97	2
3	3	973	36	3.51	3.51	.96	.95	-.98	2.48	3
4	4	1504	56	6.51	6.49	.99	.88	5.94	(7.04)	4
MISSING		8	0	5.08						

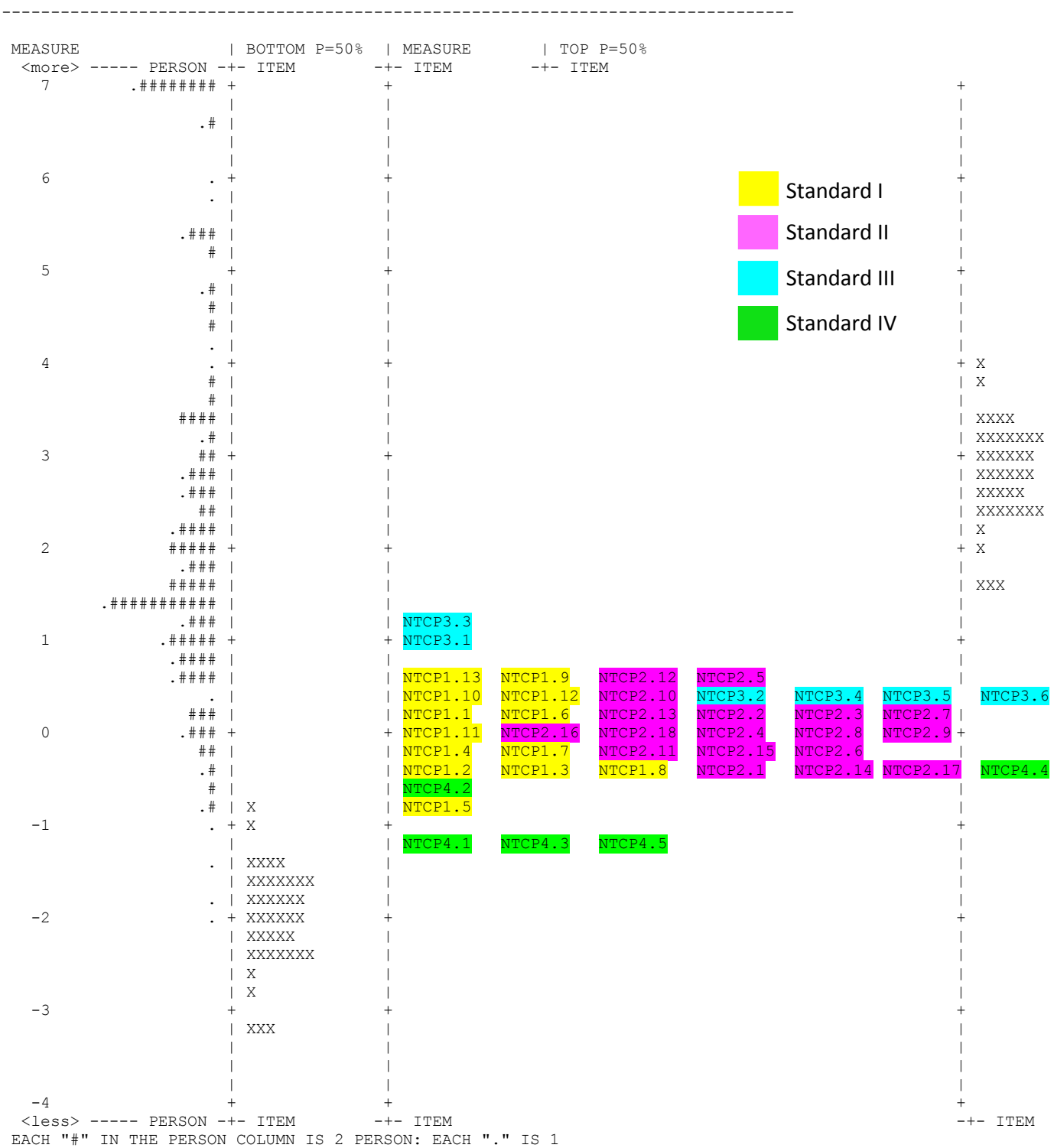
OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.



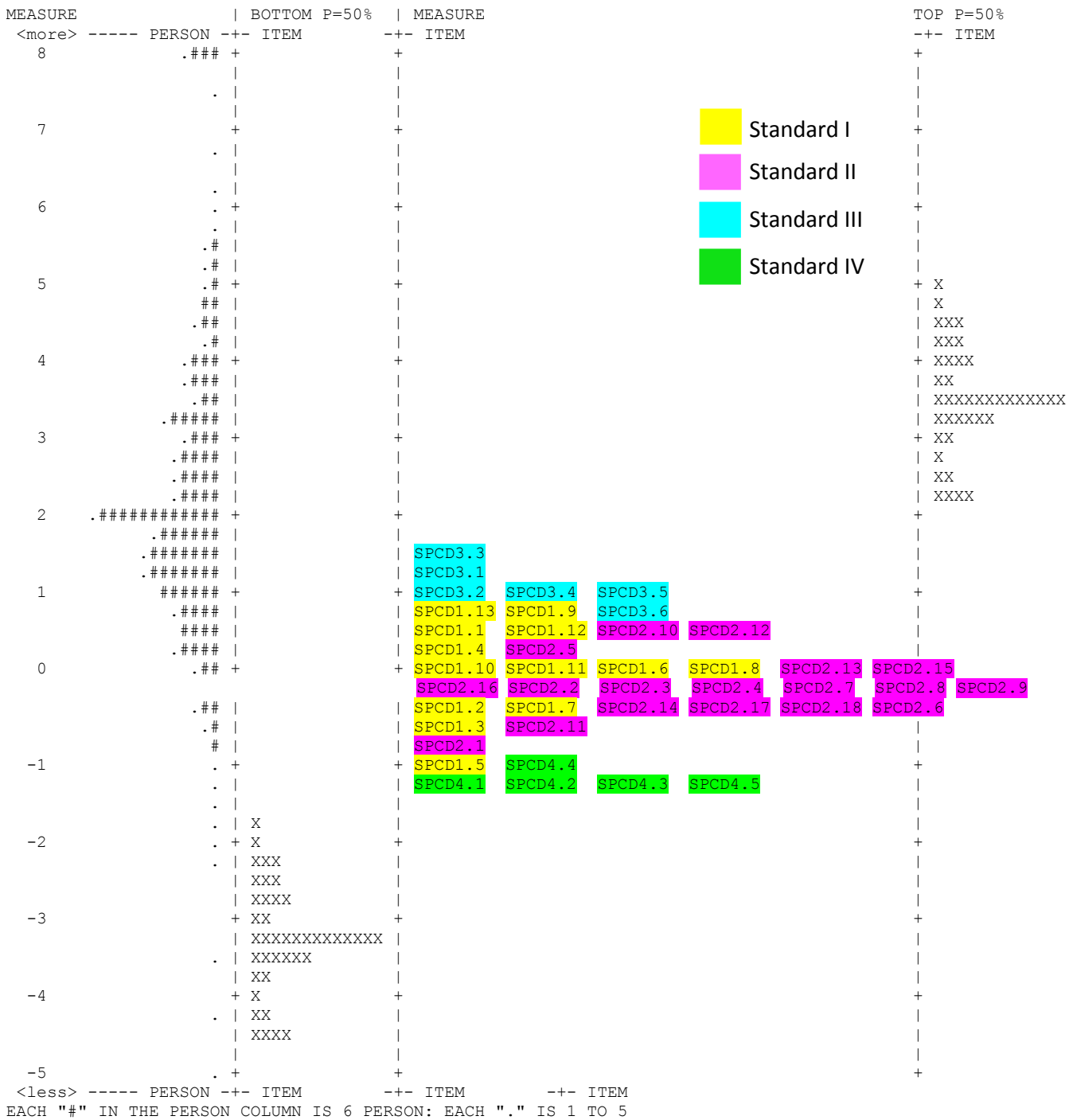
Appendix E-8: New Teacher Candidate (NTCD) Item Variable Map



Appendix E-9: New Teacher Completer (NTCP) Item Variable Map



Appendix E-10: Supervising Practitioner Candidate (SPCD) Item Variable Map



Appendix E-11: Assessment Item Hierarchy (Standard I)

Performance Standard I		Average Item Deltas (Standard Error)				
Item	**New Teacher** Prompt- I am able to: **Teacher-of-Record** Prompt- I have improved in my ability to: **Supervising Practitioner**- The Candidate is able to:	NTCD	TRCD	NTCP	TRCP	SPCD
1.13	use technology to analyze and track student achievement.	0.85 (.08)	1.05 (.13)	0.67 (.11)	0.97 (.11)	0.65 (.06)
1.12	use assessment methods that enable students to rate their own understanding of student learning objectives.	0.53 (.08)	0.40 (.14)	0.34 (.11)	0.44 (.12)	0.52 (.06)
1.6	consistently use assessment data to guide changes to my instructional practice.	0.36 (.08)	0.03 (.14)	0.18 (.11)	-0.08 (.13)	-0.01 (.07)
1.11	analyze student performance data to provide students with timely feedback that enables them to improve their work.	0.29 (.08)	0.24 (.14)	0.07 (.12)	0.33 (.12)	-0.11 (.07)
1.10	use assessment data to differentiate instruction for different groups of students within the classroom (e g , English Language Learners, Special Needs).	0.20 (.08)	-0.03 (.14)	0.37 (.11)	0.12 (.12)	0.08 (.07)
1.2	design formative assessments to check student understanding.	-0.20 (.09)	-0.66 (.15)	-0.39 (.13)	-0.42 (.13)	-0.21 (.07)

Appendix E-12: Classroom Management Item Hierarchy (Standard II)

Performance Standard II		Average Item Deltas (Standard Error)				
Item	**New Teacher** Prompt- I am able to: **Teacher-of-Record** Prompt- I have improved in my ability to: **Supervising Practitioner**- The Candidate is able to:	NTCD	TRCD	NTCP	TRCP	SPCD
2.5	effectively engage students who resist wanting to learn.	0.70 (.08)	0.54 (.14)	0.59 (.11)	0.56 (.12)	0.32 (.07)
2.12	detect and prevent potential behavioral problems from happening in the classroom.	0.47 (.08)	0.30 (.14)	0.58 (.11)	0.66 (.12)	0.49 (.06)
2.2	respond appropriately to misunderstandings between students that arise from difference in backgrounds, languages or identities.	0.46 (.08)	0.25 (.14)	0.22 (.11)	0.38 (.12)	0.07 (.07)
2.3	effectively guide students to refocus their efforts in class when they become distracted.	-0.02 (.09)	0.21 (.14)	0.26 (.11)	0.35 (.12)	-.06 (.07)
2.8	use classroom management techniques that promote students' staying on task (e g , routines, transition and response strategies).	-0.22 (.09)	-0.58 (.15)	-0.01 (.12)	0.17 (.12)	-0.01 (.07)

Appendix F1: Descriptive Data and Real Person Separation Reliabilities (PSR) for Performance Standard I

Non-Extreme Person Report	New Teacher Candidate (NTCD)	Teacher-of-Record Candidate (TRCD)	New Teacher Completer (NTCP)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)
Number of Respondents[1]	352 (408)	157 (179)	176 (202)	163(186)	599 (649)
Maximum Extreme Score	13.5%	13.3%	12.9%	11.8%	7.6%
Number of Items	13	13	13	13	13
Mean	2.12	2.37	1.93	2.15	1.98
Standard Deviation	1.64	2.07	1.56	1.70	1.82
Variance Explained	47.5%	60.0%	50.1%	52.8%	54.4%
Real PSR2	0.87	0.91	0.87	0.88	0.90
Responsiveness (person strata)	3.7	4.5	3.8	3.9	4.3

[1]There were not enough items administered to Hiring Principals to assess standard-based reliabilities.

Appendix F2: Descriptive Data and Real Person Separation Reliabilities (PSR) for Performance Standard II

Non-Extreme Person Report	New Teacher Candidate (NTCD)	Teacher-of-Record Candidate (TRCD)	New Teacher Completer (NTCP)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)
Number of Respondents[1]	340 (408)	155 (179)	171 (202)	161 (186)	597 (649)
Maximum Extreme Score	16.3%	13.4%	15.3%	12.4%	7.7%
Number of Items	18	18	18	18	18
Mean	2.68	2.99	1.79	1.98	2.31
Standard Deviation	1.71	2.46	1.78	1.79	1.98
Variance Explained	47.2%	47.2%	55.5%	54.8%	54.3%
Real PSR[2]	0.90	0.94	0.92	0.92	0.93
Responsiveness (person strata)	4.4	5.6	4.8	4.9	5.2

[1]There were not enough items administered to Hiring Principals to assess standard-based reliabilities.

Appendix F3: Descriptive Data and Real Person Separation Reliabilities (PSR) for Performance Standard III

Non-Extreme Person Report	New Teacher Candidate (NTCD)	Teacher-of-Record Candidate (TRCD)	New Teacher Completer (NTCP)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)
Number of Respondents[1]	336 (408)	150 (179)	165 (202)	155 (186)	602 (649)
Maximum Extreme Score	17.0%	13.5%	16.3%	16.1%	6.6%
Number of Items	6	6	6	6	6
Mean	2.19	2.58	2.44	4.37	2.10
Standard Deviation	2.55	2.97	2.69	3.57	2.94
Variance Explained	66.0%	71.0%	67.0%	69.0%	67.4%
Real PSR[2]	0.86	0.87	0.87	**0.59**	0.87
Responsiveness (person strata)	3.6	3.8	3.8	1.9	3.8

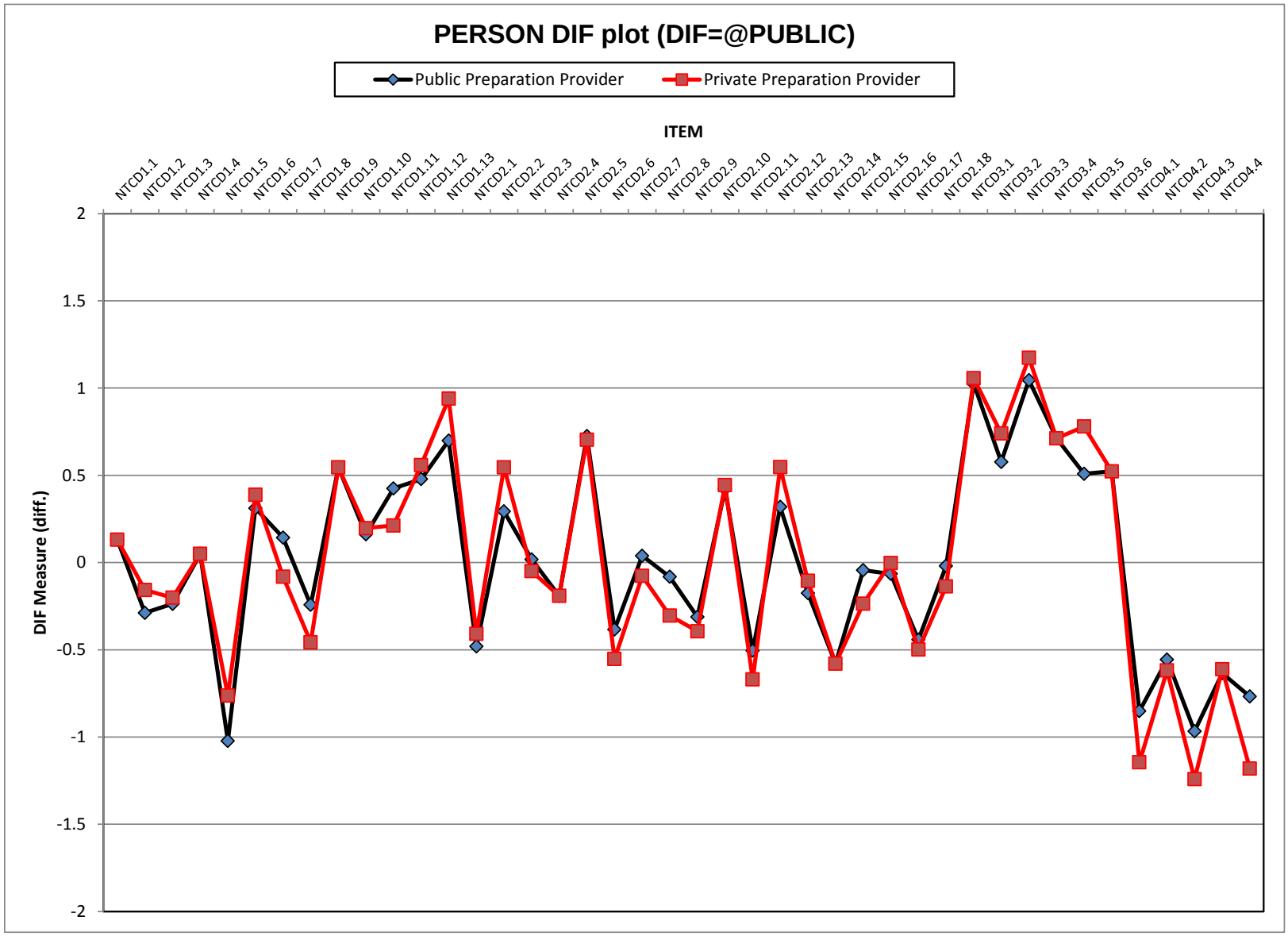
[1]There were not enough items administered to Hiring Principals to assess standard-based reliabilities.

Appendix F4: Descriptive Data and Real Person Separation Reliabilities (PSR) for Performance Standard IV

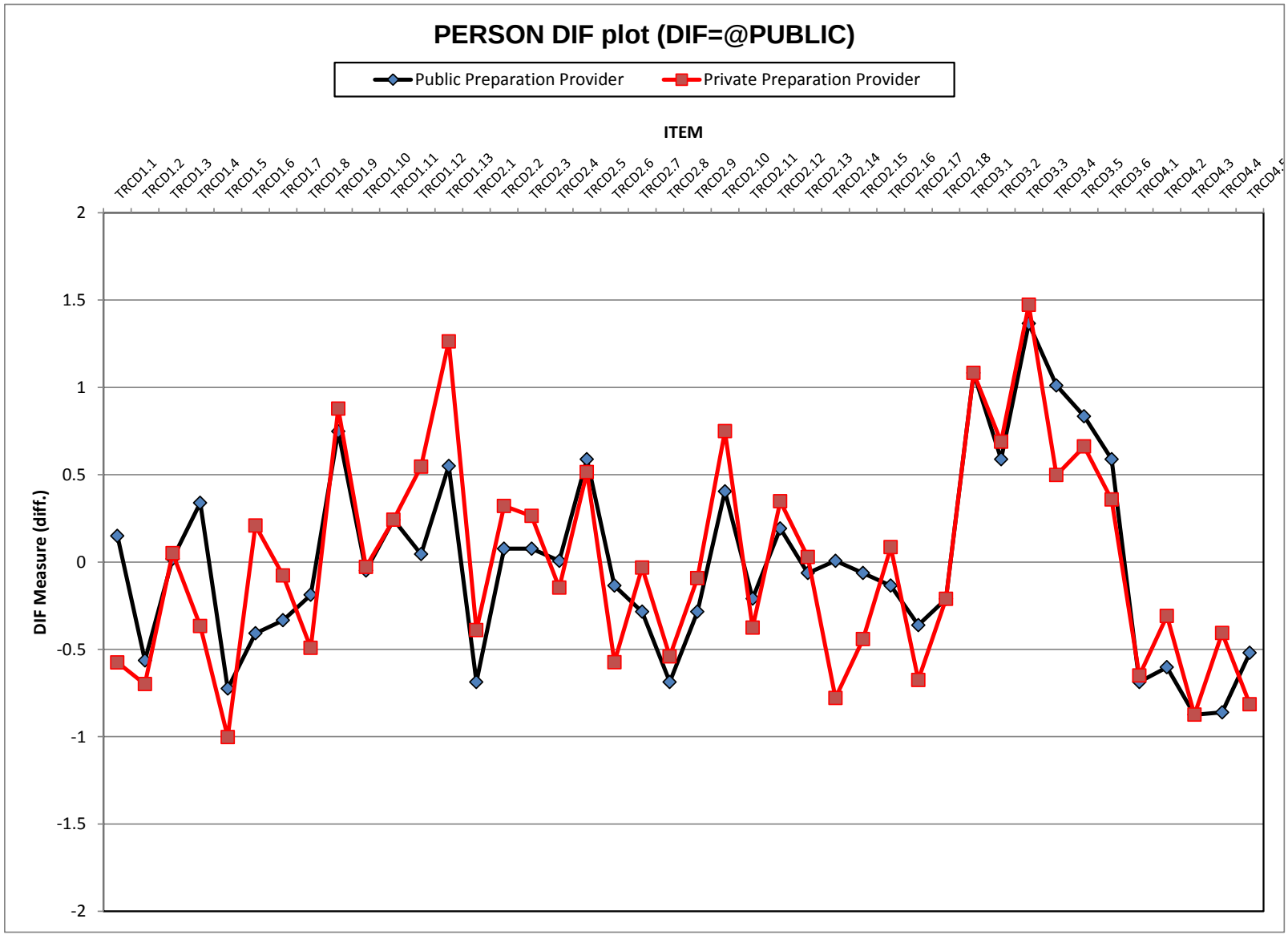
Non-Extreme Person Report	New Teacher Candidate (NTCD)	Teacher-of-Record Candidate (TRCD)	New Teacher Completer (NTCP)	Teacher-of-Record Completer (TRCP)	Supervising Practitioner (SPCD)
Number of Respondents[1]	256 (408)	120 (179)	133 (202)	127 (186)	459 (649)
Maximum Extreme Score	**37.1%**	**32.6%**	**34.2%**	**30.8%**	**29.0%**
Number of Items	5	5	5	5	5
Mean	3.78	4.12	3.95	2.98	3.73
Standard Deviation	2.27	3.57	2.47	2.47	2.55
Variance Explained	57.4%	71.0%	62.6%	60.7%	61.9%
Real PSR[2]	0.69	0.67	0.74	0.77	0.70
Responsiveness (person strata)	2.4	2.2	2.6	1.9	2.4

[1]There were not enough items administered to Hiring Principals to assess standard-based reliabilities.

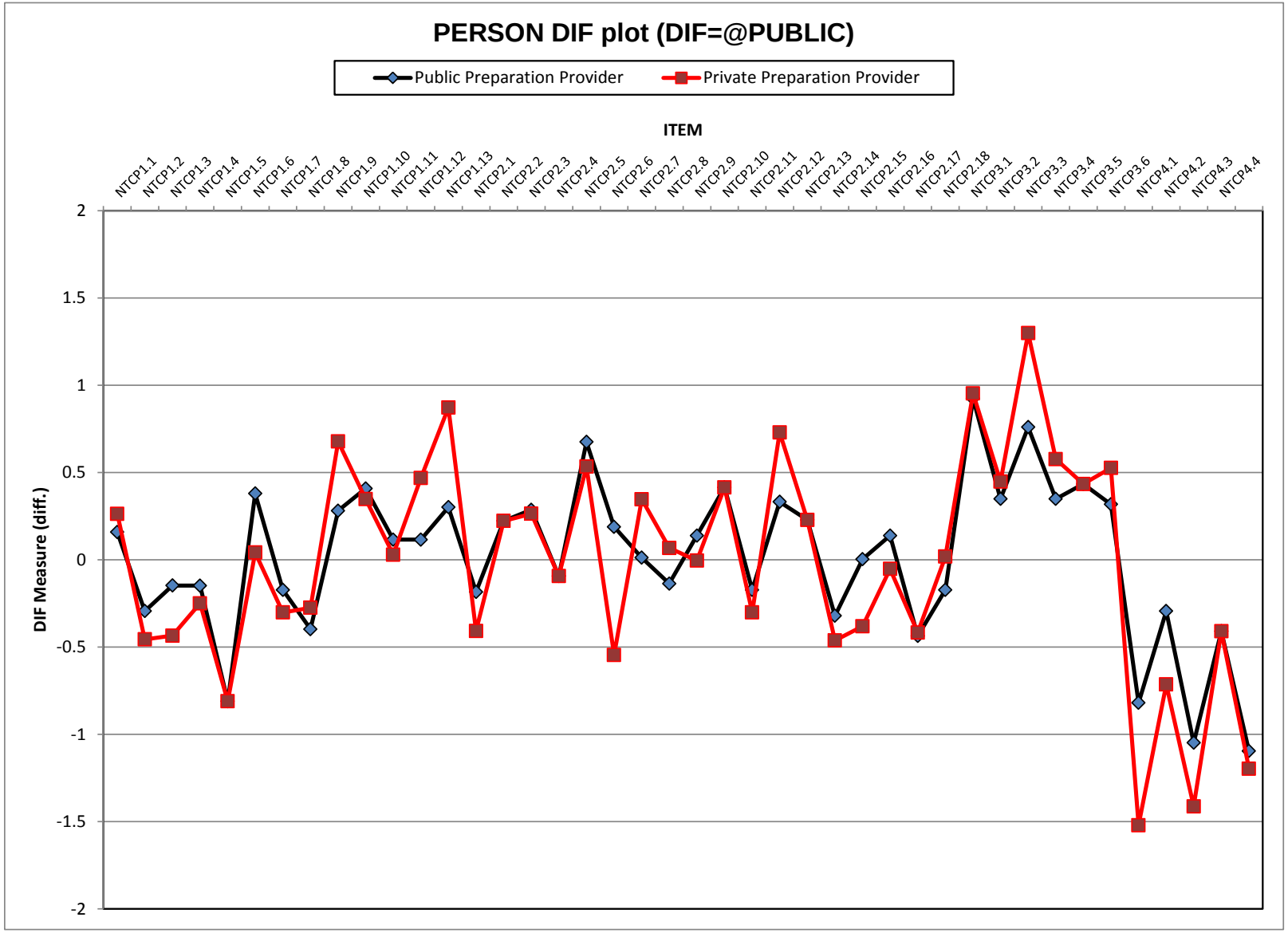
Appendix F-5: New Teacher Candidate (NTCD) Differential Item Functioning



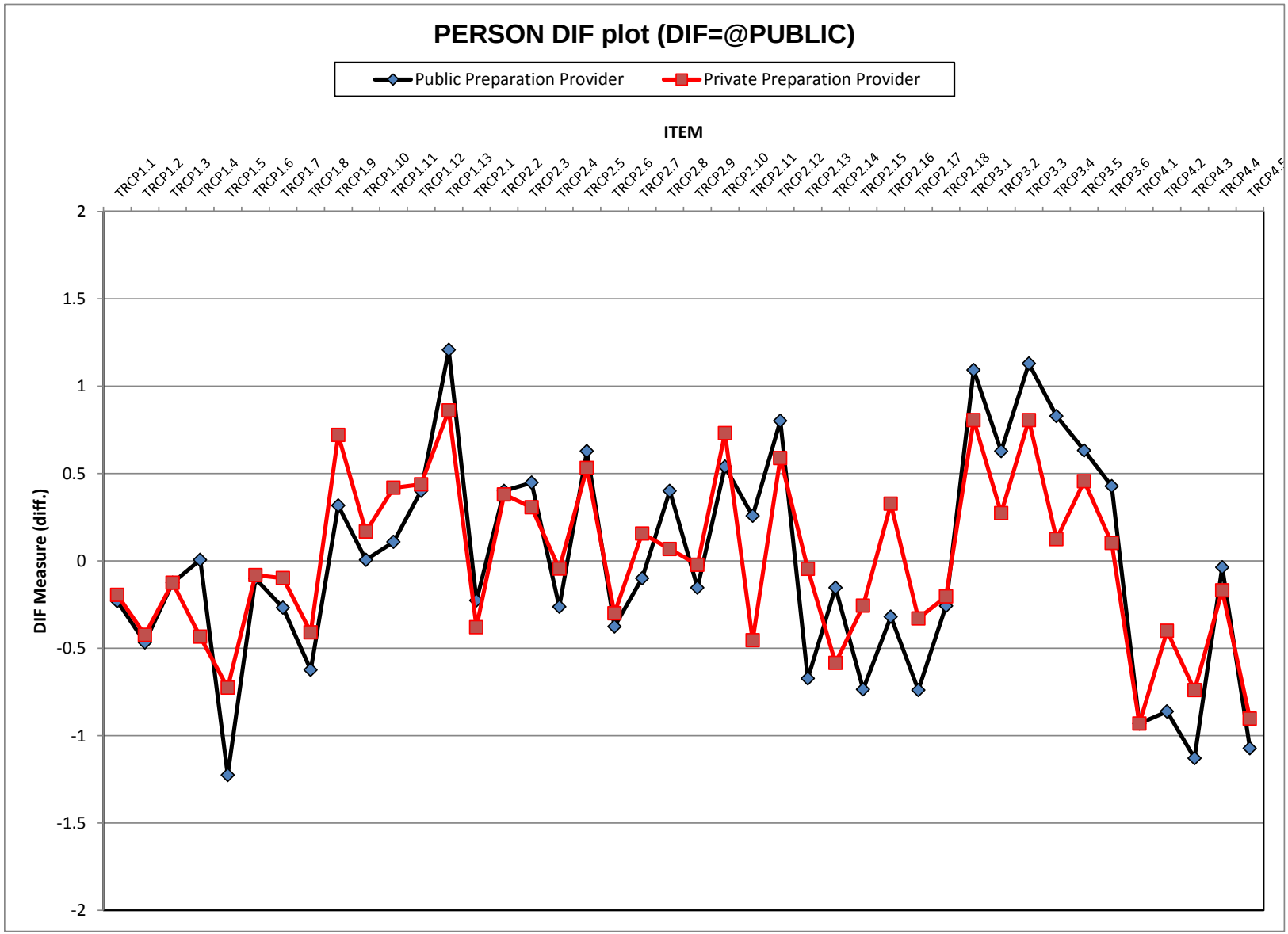
Appendix F-6: Teacher-of-Record Candidate (TRCD) Differential Item Functioning



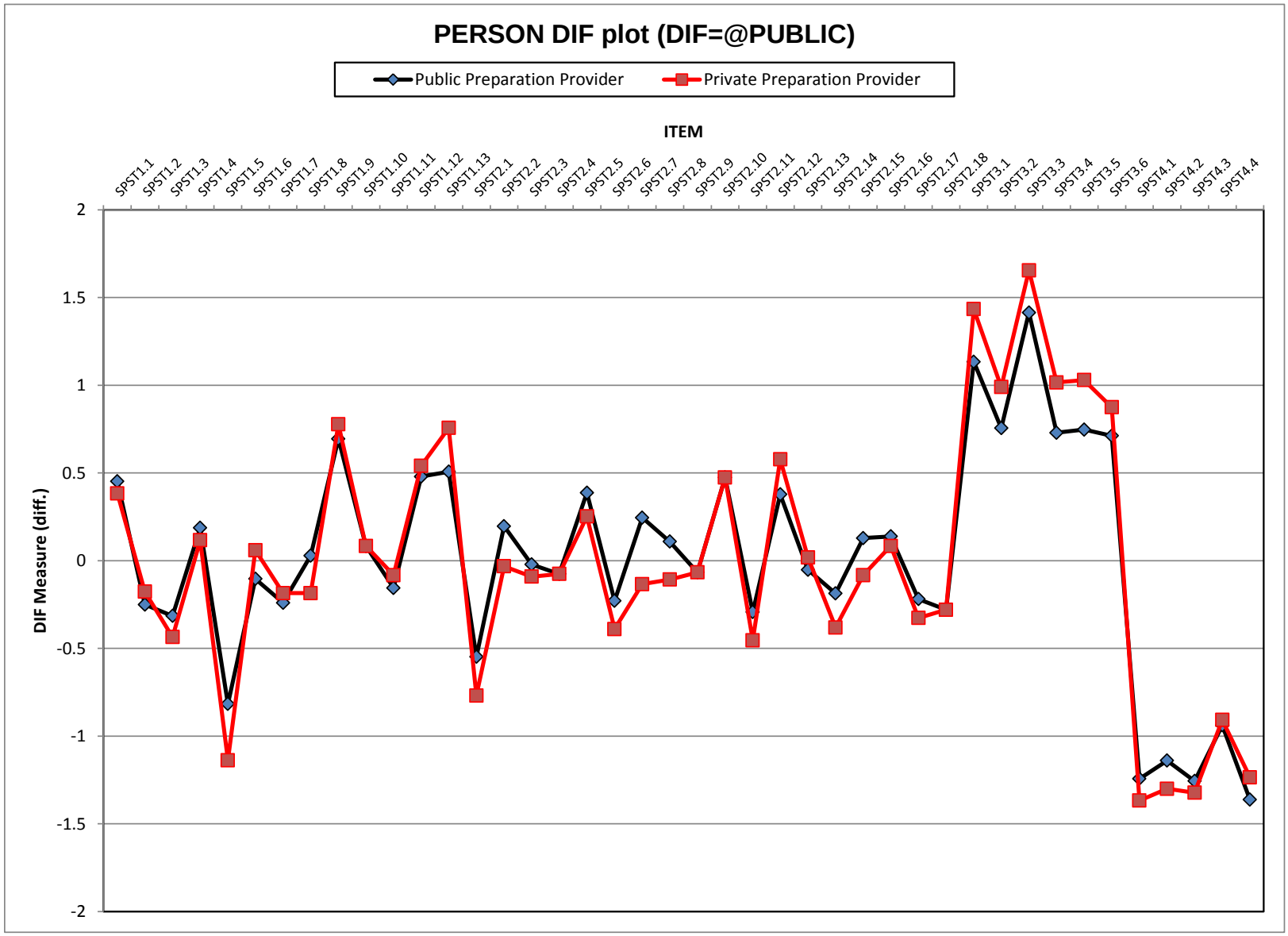
Appendix F-7: New Teacher Completer (NTCP) Differential Item Functioning



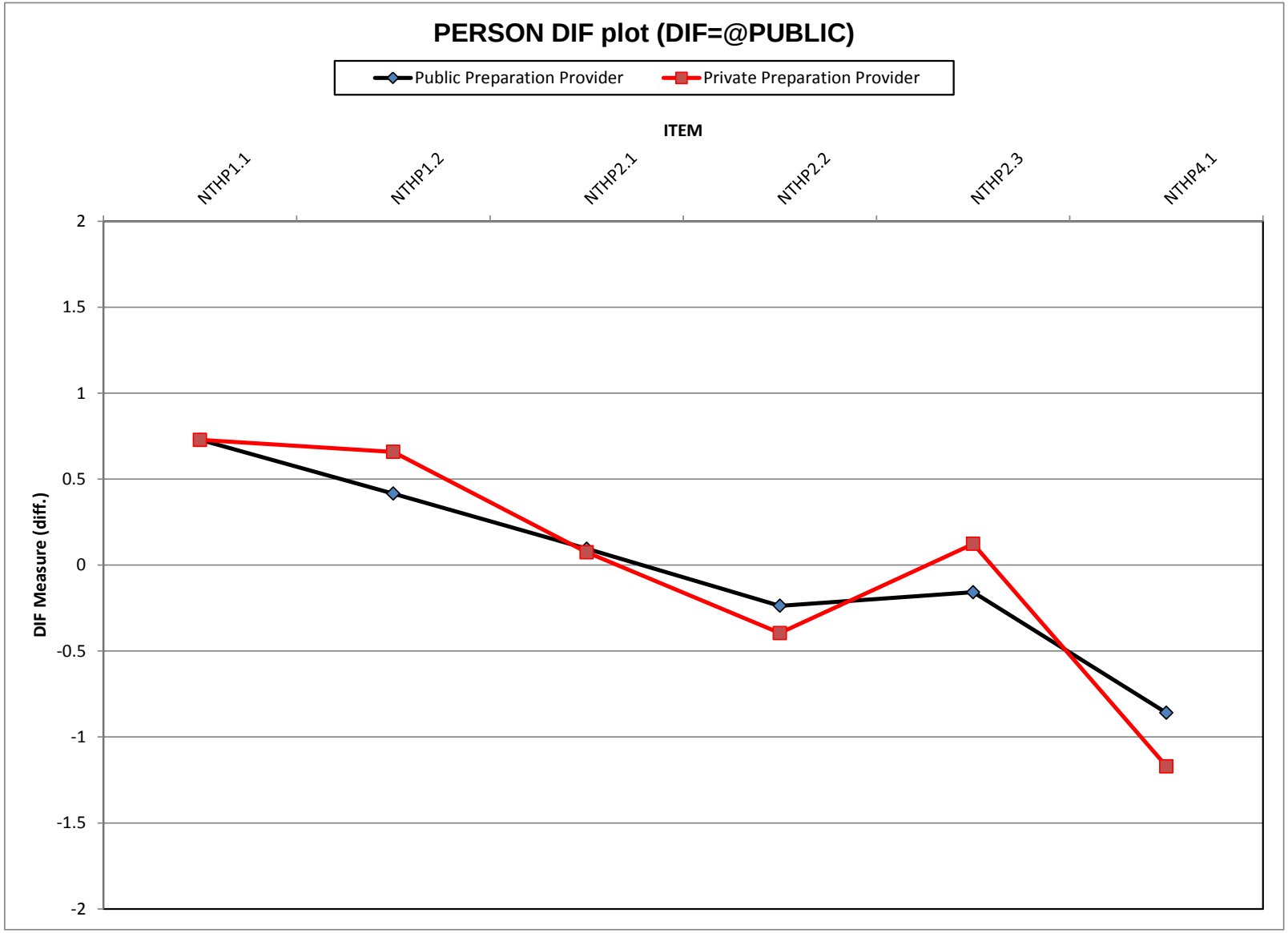
Appendix F-8: Teacher-of-Record Completer (NTCP) Differential Item Functioning



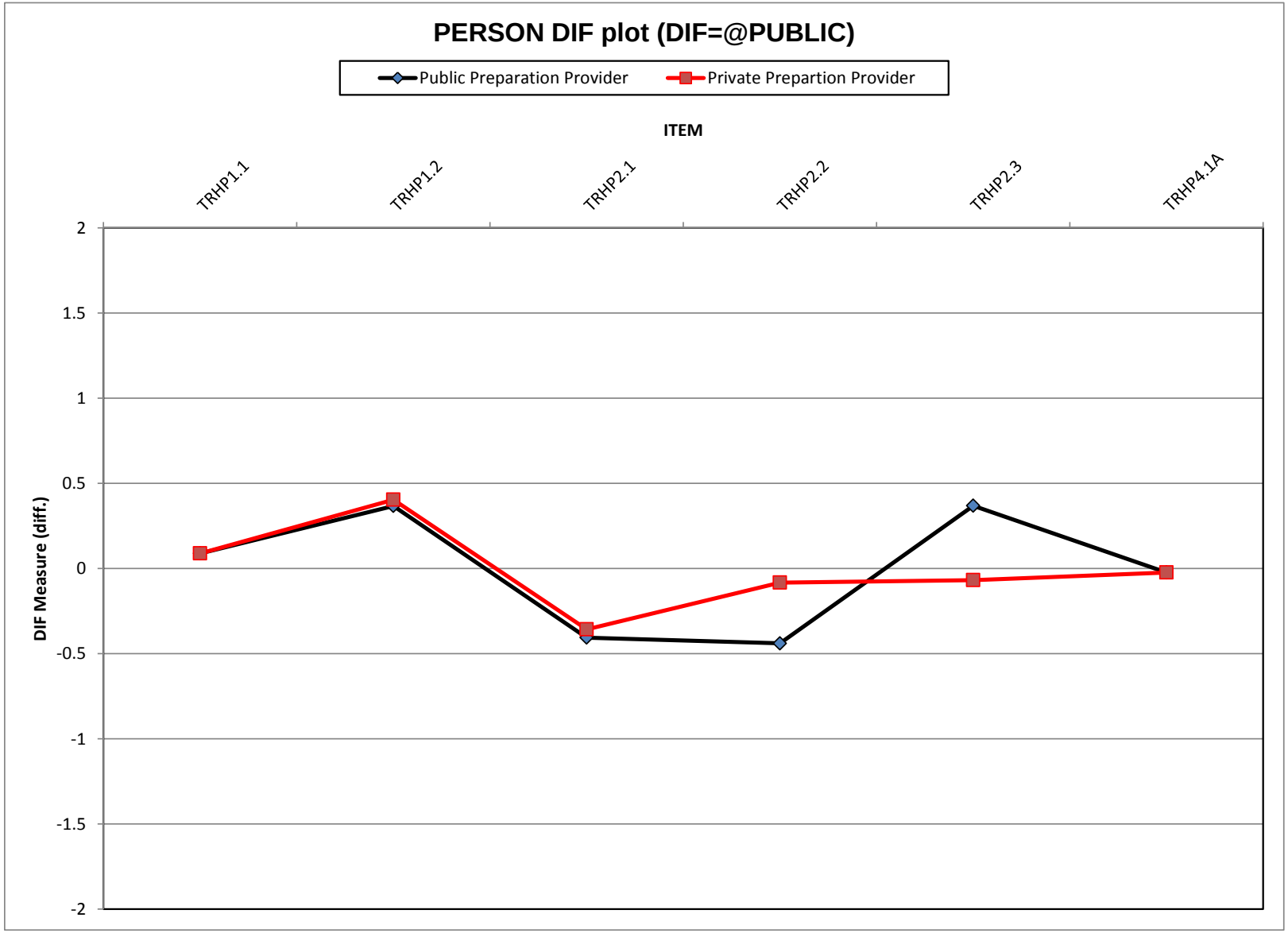
Appendix F-9: Supervising Practitioner Teacher Candidate (SPCD) Differential Item Functioning



Appendix F-10: New Teacher Hiring Principal (NTHP) Differential Item Functioning



Appendix F-11: Teacher-of-Record Principal (TRHP) Differential Item Functioning



Appendix G-1: Performance Standard I Item-Level Invariance across 42 Item Surveys

Performance Standard I		Average Item Deltas (Standard Error)				
Item	**New Teacher** Prompt- I am able to: **Teacher-of-Record** Prompt- I have improved in my ability to: **Supervising Practitioner**- The Candidate is able to:	NTCD	TRCD	NTCP	TRCP	SPCD
1.1	integrate language acquisition into content instruction so English Language Learner students learn as they build language skills.	0.13 (.09)	-0.34 (.15)	0.22 (.11)	-0.19 (.13)	0.42 (.06)
1.2	design formative assessments to check student understanding.	-0.20 (.09)	-0.66 (.15)	-0.39 (.13)	-0.42 (.13)	-0.21 (.07)
1.3	use intellectual engagement strategies that require students to support their answers or reasoning in class.	-0.20 (.09)	0.05 (.14)	-0.31 (.12)	-0.13 (.13)	-0.38 (.07)
1.4	consistently model and use academic language that can be understood by English Language Learner students at all proficiency levels so they can build their content knowledge.	0.05 (.09)	-0.13 (.15)	-0.21 (.12)	-0.30 (.13)	0.15 (.07)
1.5	develop well-structured lessons that incorporated students' interests in the planning of class activities.	-0.85 (.10)	-0.92 (.16)	-0.81 (.13)	-0.87 (.14)	-0.99 (.07)
1.6	consistently use assessment data to guide changes to my instructional practice.	0.36 (.08)	0.03 (.14)	0.18 (.11)	-0.08 (.13)	-0.01 (.07)
1.7	design units of instruction that help students develop many ways to think deeply about an activity or a problem.	0.01 (.09)	-0.15 (.15)	-0.25 (.12)	-0.15 (.13)	-0.21 (.07)
1.8	scaffold and unpack content so all students can understand the material.	-0.37 (.09)	-0.40 (.15)	-0.32 (.12)	-0.47 (.13)	-0.08 (.07)
1.9	develop interdisciplinary curriculum.	0.54 (.08)	0.84 (.13)	0.53 (.11)	0.61 (.12)	0.74 (.06)
1.10	use assessment data to differentiate instruction for different groups of students within the classroom (e g , English Language Learners, Special Needs).	0.20 (.08)	-0.03 (.14)	0.37 (.11)	0.12 (.12)	0.08 (.07)
1.11	analyze student performance data to provide students with timely feedback that enables them to improve their work.	0.29 (.08)	0.24 (.14)	0.07 (.12)	0.33 (.12)	-0.11 (.07)
1.12	use assessment methods that enable students to rate their own understanding of student learning objectives.	0.53 (.08)	0.40 (.14)	0.34 (.11)	0.44 (.12)	0.52 (.06)
1.13	use technology to analyze and track student achievement.	0.85 (.08)	1.05 (.13)	0.67 (.11)	0.97 (.11)	0.65 (.06)

Appendix G-2: Performance Standard II Item-Level Invariance across 42 Item Surveys

Performance Standard II		Average Item Deltas (Standard Error)				
Item	**New Teacher** Prompt- I am able to: **Teacher-of-Record** Prompt- I have improved in my ability to: **Supervising Practitioner**- The Candidate is able to:	NTCD	TRCD	NTCP	TRCP	SPCD
2.1	provide learning experiences that encourage students to be supportive of each other's success.	-0.43 (.09)	-0.47 (.15)	-0.31 (.12)	-0.33 (.13)	-0.66 (.07)
2.2	respond appropriately to misunderstandings between students that arise from difference in backgrounds, languages or identities.	0.46 (.08)	0.25 (.14)	0.22 (.11)	0.38 (.12)	0.07 (.07)
2.3	effectively guide students to refocus their efforts in class when they become distracted.	-0.02 (.09)	0.21 (.14)	0.26 (.11)	0.35 (.12)	-.06 (.07)
2.4	teach so, when asked, students can explain what they are learning and why.	-0.19 (.09)	-0.10 (.15)	-0.09 (.12)	-0.11 (.13)	-0.08 (.07)
2.5	effectively engage students who resist wanting to learn.	0.70 (.08)	0.54 (.14)	0.59 (.11)	0.56 (.12)	0.32 (.07)
2.6	create a learning environment where the teacher has the same high academic expectations for her or his English Language Learner students as she or he does for her native English learners.	-0.49 (.10)	-0.44 (.15)	-0.23 (.12)	-0.32 (.13)	-0.31 (.07)
2.7	use instructional practices that encourage students to challenge each other's thinking in the classroom.	-0.03 (.09)	-0.10 (.15)	0.22 (.11)	0.08 (.12)	0.04 (.07)
2.8	use classroom management techniques that promote students' staying on task (e g , routines, transition and response strategies).	-0.22 (.09)	-0.58 (.15)	-0.01 (.12)	0.17 (.12)	-0.01 (.07)
2.9	use student generated ideas to further student understanding during a lesson.	-0.36 (.09)	-0.15 (.15)	0.06 (.12)	-0.06 (.13)	-0.07 (.07)
2.10	connect students with socio-emotional problems with appropriate support.	0.44 (.08)	0.65 (.14)	0.41 (.11)	0.67 (.12)	0.47 (.06)
2.11	instill in students a growth mindset (perseverance, learn from mistakes, high expectations valued) so all students believe in their ability to learn.	-0.60 (.10)	-0.32 (.15)	-0.25 (.12)	-0.22 (.13)	-0.38 (.07)
2.12	detect and prevent potential behavioral problems from happening in the classroom.	0.47 (.08)	0.30 (.14)	0.58 (.11)	0.66 (.12)	0.49 (.06)
2.13	differentiate instruction so all students are challenged at all times during a lesson.	-0.13 (.09)	0.00 (.15)	0.23 (.11)	-0.22 (.13)	-0.01 (.07)

Appendix G-2: Performance Standard II Item-Level Invariance across 42 Item Surveys continued

Performance Standard II continued		Average Item Deltas (Standard Error)				
Item	**New Teacher** Prompt- I am able to: **Teacher-of-Record** Prompt- I have improved in my ability to: **Supervising Practitioner**- The Candidate is able to:	NTCD	TRCD	NTCP	TRCP	SPCD
2.14	plan effective techniques (e g , use of visuals, model discussion, group work) for making content accessible to English Language Learners such that English Language Learners of mixed proficiency can participate.	-0.58 (.10)	-0.53 (.15)	-0.40 (.13)	-0.45 (.13)	-0.29 (.07)
2.15	comfortably take instructional risks (e g , deviate from planned instruction) to make student learning more accessible.	-0.16 (.09)	-0.32 (.15)	-0.22 (.12)	-0.39 (.13)	0.02 (.07)
2.16	translate knowledge of the diverse experiences that students bring to class to improve the effectiveness of my instruction.	-0.02 (.09)	0.02 (.15)	0.03 (.12)	0.14 (.12)	0.11 (.07)
2.17	choose instructional strategies (e g , tiered instruction, scaffolding, connections) that support student understanding of complex concepts.	-0.48 (.10)	-0.58 (.15)	-0.42 (.13)	-0.45 (.13)	-0.27 (.07)
2.18	create a cooperative but cognitively demanding learning environment where students support each other to strengthen their work.	-0.09 (.09)	-0.21 (.15)	-0.06 (.12)	-0.20 (.13)	-0.28 (.07)

Appendix G-3: Performance Standard III Item-Level Invariance across 42 Item Surveys

Performance Standard III		Average Item Deltas (Standard Error)				
Item	**New Teacher** Prompt- I am able to: **Teacher-of-Record** Prompt- I have improved in my ability to: **Supervising Practitioner**- The Candidate is able to:	NTCD	TRCD	NTCP	TRCP	SPCD
3.1	make strategies available to parents so they can help support and reinforce student learning at home and in school.	1.06 (.07)	1.08 (.13)	0.95 (.11)	0.90 (.11)	1.30 (.06)
3.2	listen to a parent's concerns regarding the progress of their child and use the information to adapt my instruction towards the child.	0.68 (.08)	0.66 (.14)	0.41 (.11)	0.39 (.12)	0.89 (.06)
3.3	connect families to resources outside of school to support student learning in school.	1.13 (.07)	1.44 (.13)	1.10 (.10)	0.91 (.11)	1.55 (.06)
3.4	communicate effectively with families from diverse backgrounds and cultures.	0.71 (.08)	0.66 (.14)	0.49 (.11)	0.35 (.12)	0.89 (.06)
3.5	effectively implement two-way communication strategies (e g , back-to-school nights, office hours) to include parent perspectives in the classroom.	0.68 (.08)	0.72 (.14)	0.43 (.11)	0.51 (.12)	0.91 (.06)
3.6	demonstrate cultural responsiveness when communicating with English Language Learner students' families.	0.52 (.08)	0.43 (.14)	0.45 (.11)	0.20 (.12)	0.80 (.06)

Appendix G-4: Performance Standard IV Item-Level Invariance across 42 Item Surveys

Performance Standard IV		Average Item Deltas (Standard Error)				
Item	**New Teacher** Prompt- I am able to: **Teacher-of-Record** Prompt- I have improved in my ability to: **Supervising Practitioner**- The Candidate is able to:	NTCD	TRCD	NTCP	TRCP	SPCD
4.1	reflect on my practice to develop challenging professional practice goal.	-1.03 (.10)	-0.65 (.16)	-1.21 (.14)	-0.93 (.14)	-1.31 (.08)
4.2	use evaluation data (e g , your feedback) to diagnose my strengths and weaknesses and make adjustment to my practice (e.g., instructional goals, learning objectives).	-0.59 (.10)	-0.39 (.15)	-0.53 (.13)	-0.53 (.13)	-1.22 (.08)
4.3	reflect on my practice in order to identify areas for professional growth (e.g., professional development opportunities).	-1.13 (.10)	-0.87 (.16)	-1.25 (.14)	-0.85 (.14)	-1.29 (.08)
4.4	when established, apply school expectations for student behavior.	-0.61 (.10)	-0.53 (.16)	-0.41 (.13)	-0.13 (.13)	-0.91 (.07)
4.5	effectively act upon colleagues' ideas and/or suggestions to improve my students' learning.	-1.02 (.10)	-0.72 (.16)	-1.15 (.14)	-0.95 (.14)	-1.29 (.08)

Appendix G-5: Item-Level Invariance across Surveys across Hiring Principal Surveys

Hiring Principal		Average Item Deltas (Standard Error)	
Item	**New Teacher –** Relative to all other teachers (both novice and experienced) you've worked with; please indicate the extent to which this teacher's performance is significantly below or above average. **Teacher-of-Record –** Please rate the extent of change in the teacher's performance since completing an educator preparation program.	**NTHP**	**TRHP**
1.1	Implements well-structured lessons.	0.73 (.09)	0.09 (.15)
1.2	Makes adjustments to practice based on assessment data.	0.56 (.09)	0.40 (.15)
2.1	Meets the diverse needs of learners within the classroom.	0.07 (.09)	-0.36 (.15)
2.2	Maintains an academic learning environment where students are unafraid to take academic risks.	-0.33 (.09)	-0.20 (.15)
2.3	Consistently enforces high expectations for all students.	0.01 (.09)	0.09 (.15))
4.1	Uses self-reflection to improve practice.	-1.05 (.09)	-0.02 (.15)

Appendix G-6: Item Invariance Correlations across Hiring Principal 6 Item Surveys

Number of Items = 6	New Teacher Hiring Principal (NTHP)	Teacher-of-Record Hiring Principal (TRHP)
New Teacher Hiring Principal (NTHP)	1	0.48
Teacher-of-Record Hiring Principal (TRHP)	0.42	1

[1]Correlations observed are shown below the diagonal; Disattenuated correlations are shown above the diagonal.

Appendix H-1: Sub-Scale Correlations for the Four Performance Standards of Teaching (PST)

New Teacher Candidate (NTCD)

	PST I	PST II	PST III	PST IV
PST I		0.89	0.76	0.78
PST II	0.90		0.79	0.81
PST III	0.72	0.76		0.67
PST IV	0.80	0.81	0.62	

Sub-scale correlations below the diagonal are from the four-factor confirmatory factor analysis model; Correlations above the diagonal are the person measure correlations from the four standard-based consecutive (separate) Rasch models.

New Teacher Completer (NTCP)

	PST I	PST II	PST III	PST IV
PST I		0.87	0.70	0.63
PST II	0.90		0.76	0.70
PST III	0.77	0.83		0.57
PST IV	0.75	0.81	0.67	

Sub-scale correlations below the diagonal are from the four-factor confirmatory factor analysis model; Correlations above the diagonal are the person measure correlations from the four standard-based consecutive (separate) Rasch models.

Supervising Practitioner (SPCD)

	PST I	PST II	PST III	PST IV
PST I		0.85	0.67	0.62
PST II	0.90		0.67	0.67
PST III	0.72	0.72		0.44
PST IV	0.74	0.78	0.50	

Sub-scale correlations below the diagonal are from the four-factor confirmatory factor analysis model; Correlations above the diagonal are the person measure correlations from the four standard-based consecutive (separate) Rasch models.

Appendix J-1: Confirmatory Factor Analyses Fit Statistics

New Teacher Candidate (NTCD)

Number of Items = 42	One-Factor (All 42 Items)	Four-Factor (Four PSTs)	Three-Factor (PST I items combined with PST II items)
RMSEA	0.094*	0.055*	0.061*
RMSEA (90% CI)	0.091 – 0.97	0.052 – 0.059	0.058 – 0.064
CFI	0.92	0.97	0.97
TLI	0.92	0.97	0.97
Factor Loadings Above 0.5	YES	YES	YES

[1]RMSEA: Root Mean Square Error of Approximation point estimate; [2]RMSEA 90% Confidence Interval; [3]CFI: Comparative Fit Index; [4]TLI: Tucker-Lewis Index; *$p < 0.0005$.

New Teacher Completer (NTCP)

Number of Items = 42	One-Factor (All 42 Items)	Four-Factor (Four PSTs)	Three-Factor (PST I items combined with PST II items)
RMSEA	0.095*	0.066*	0.072*
RMSEA (90% CI)	0.091 – 0.100	0.061 – 0.071	0.067 – 0.077
CFI	0.93	0.97	0.96
TLI	0.93	0.97	0.96
Factor Loadings Above 0.5	YES	YES	YES

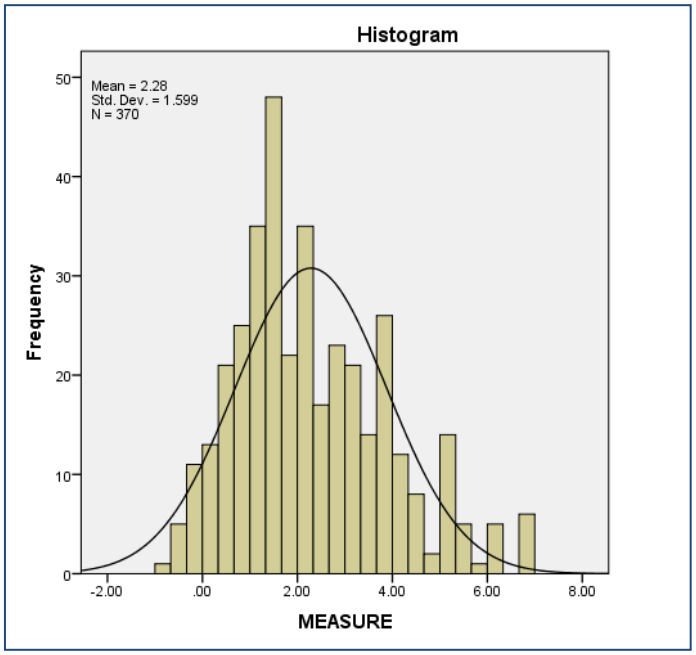
[1]RMSEA: Root Mean Square Error of Approximation point estimate; [2]RMSEA 90% Confidence Interval; [3]CFI: Comparative Fit Index; [4]TLI: Tucker-Lewis Index; *$p < 0.0005$.

Supervising Practitioner (SPCD)

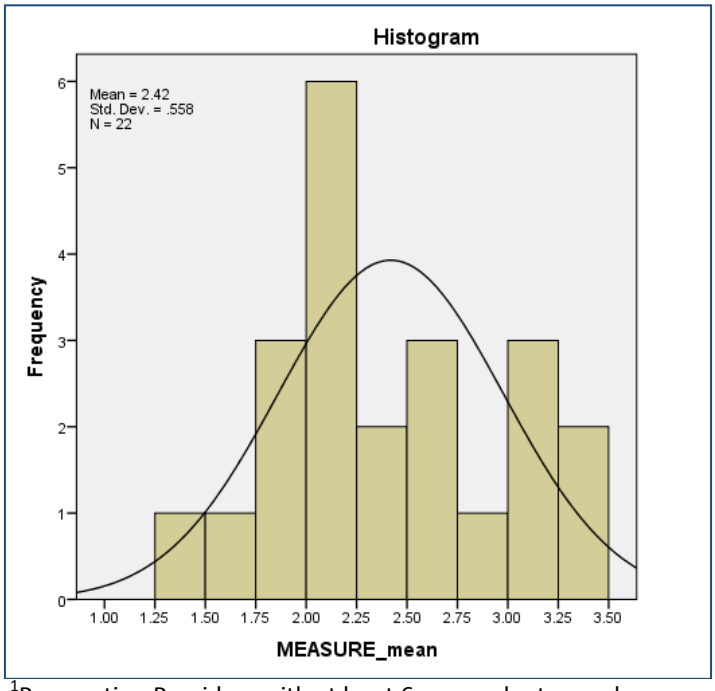
Number of Items = 42	One-Factor (All 42 Items)	Four-Factor (Four PSTs)	Three-Factor (PST I items combined with PST II items)
RMSEA[1]	0.117*	0.067*	0.073*
RMSEA (90% CI)[2]	0.115 – 0.120	0.064 – 0.069	0.071 – 0.076
CFI[3]	0.89	0.96	0.95
TLI[4]	0.87	0.96	0.95
Factor Loadings Above 0.5	YES	YES	YES

[1]RMSEA: Root Mean Square Error of Approximation point estimate; [2]RMSEA 90% Confidence Interval; [3]CFI: Comparative Fit Index; [4]TLI: Tucker-Lewis Index; *$p < 0.0005$.

Appendix K-1: New Teacher Candidate (NTCD) Non-Extreme Person Measure Distribution

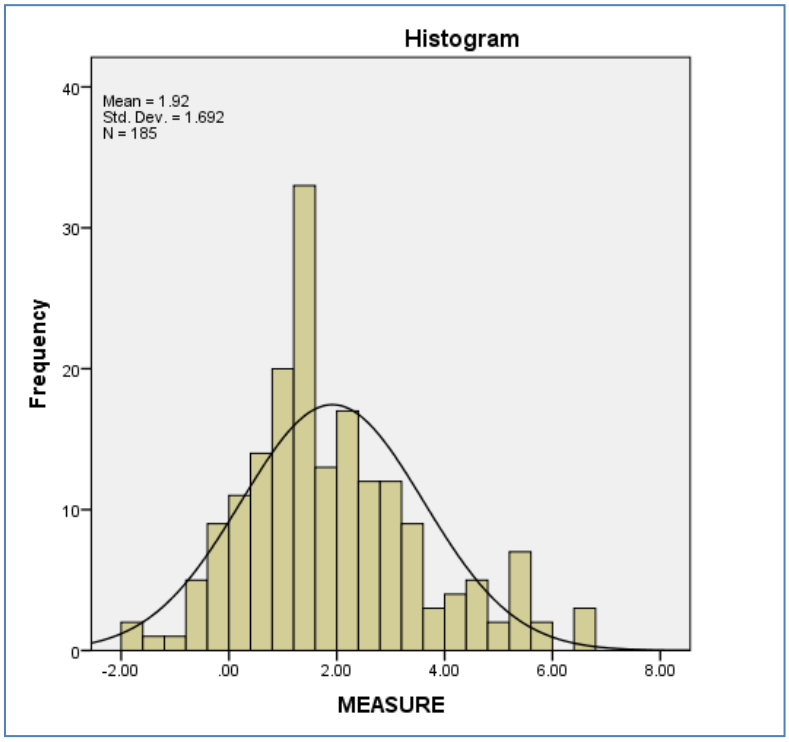


Appendix K-2: Preparation Provider Aggregate NTCD Person Measure Distribution

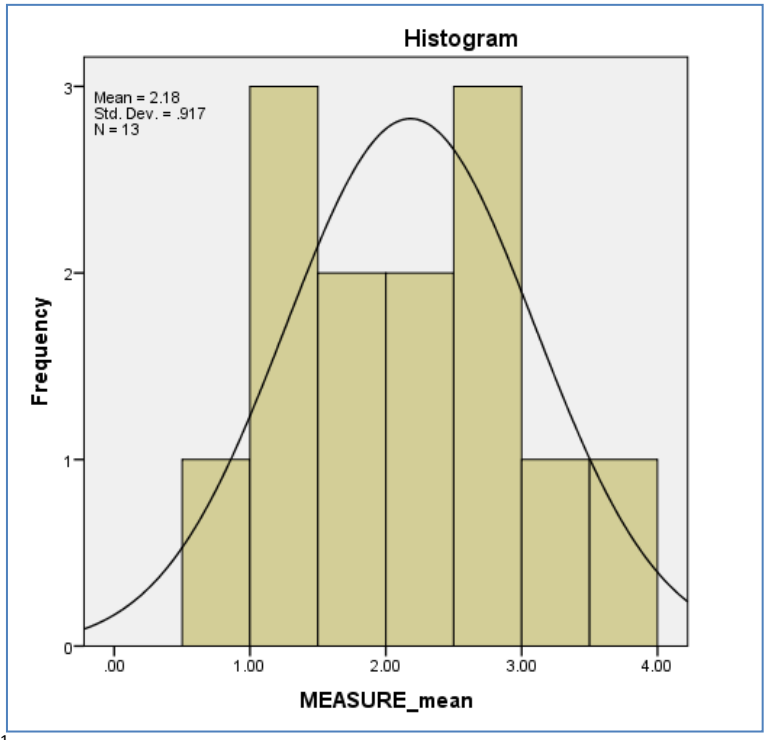


[1]Preparation Providers with at least 6 respondents are shown.

Appendix K-3: New Teacher Completer (NTCP) Non-Extreme Person Measure Distribution

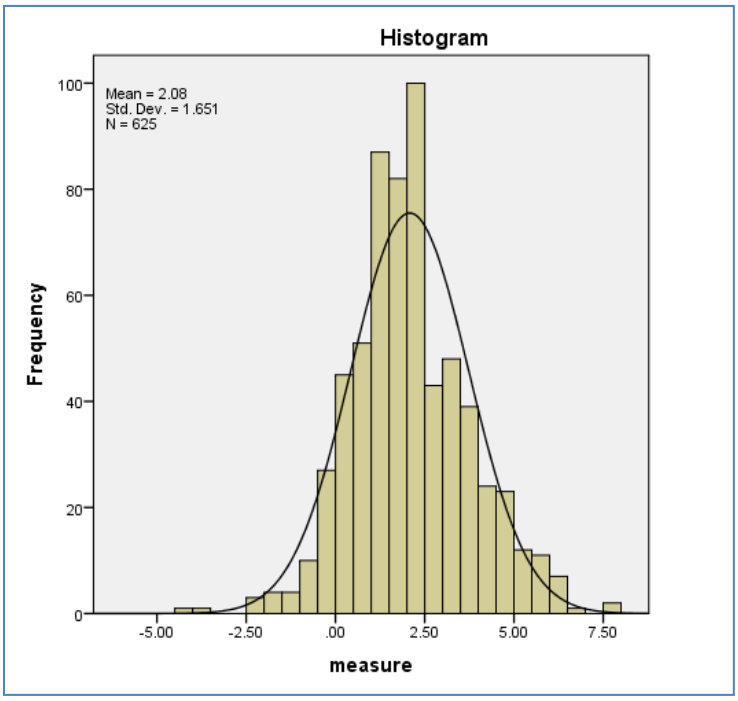


Appendix K-4: Preparation Provider Aggregate NTCP Person Measure Distribution

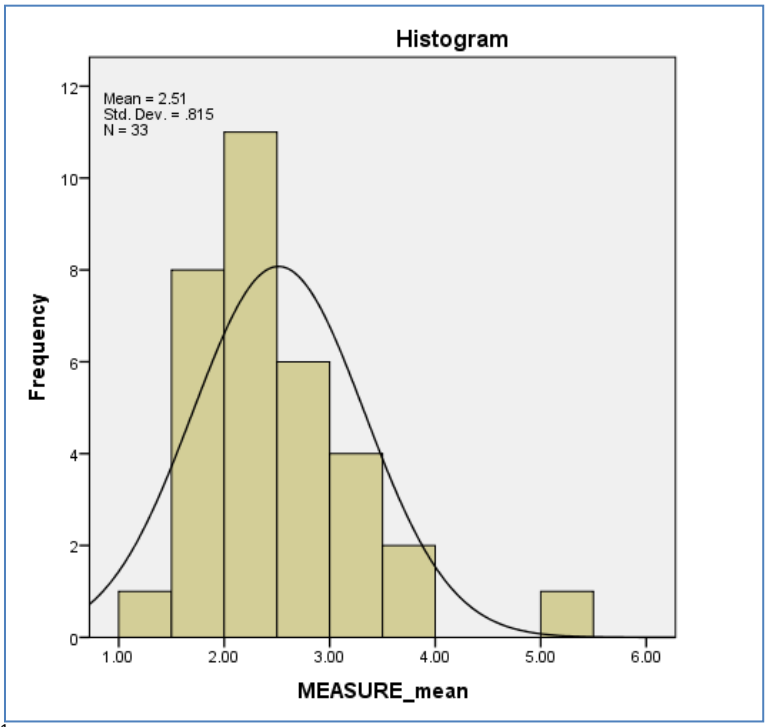


[1]Preparation Providers with at least 6 respondents are shown.

Appendix K-5: Supervising Practitioner Candidate (SPCD) Non-Extreme Person Measure Distribution

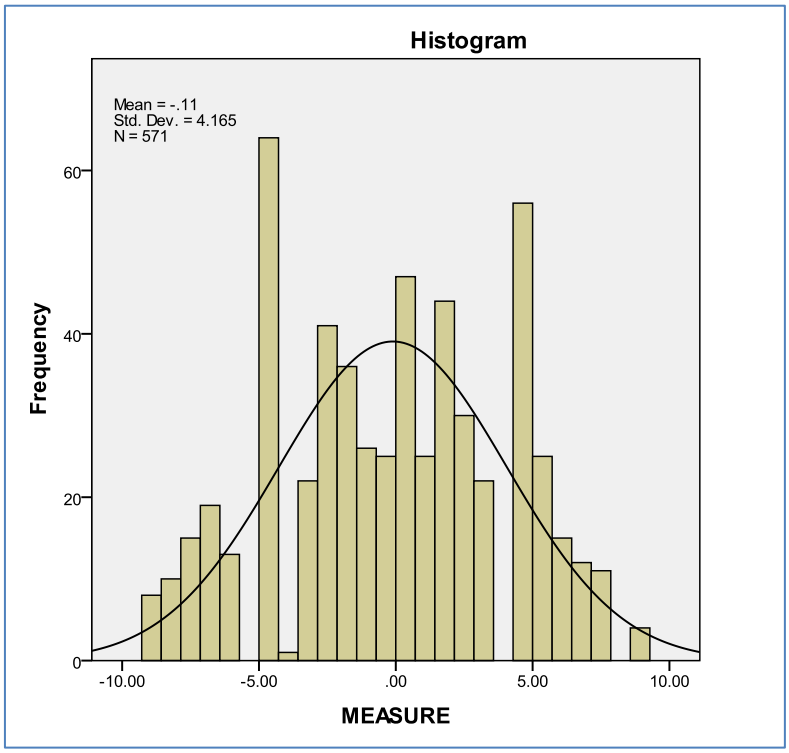


Appendix K-6: Preparation Provider Aggregate SPCD Person Measure Distribution

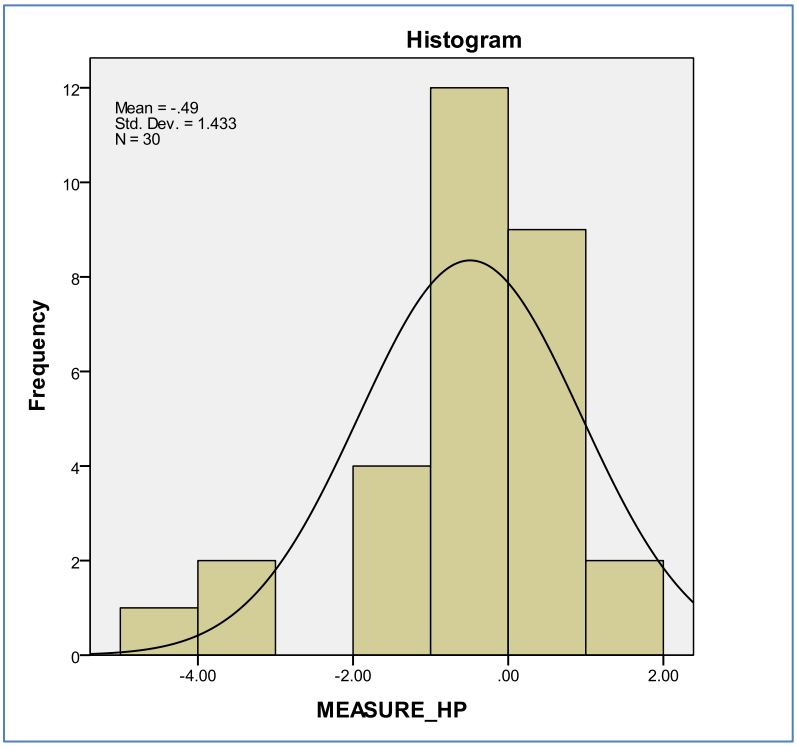


[1]Preparation Providers with at least 6 respondents are shown.

Appendix K-7: New Teacher Hiring Principal (NTHP) Non-Extreme Person Measure Distribution



Appendix K-8: Preparation Provider Aggregate NTHP Person Measure Distribution



[1]Preparation Providers with at least 6 respondents are shown.