2017 Educator Preparation Surveys: Executive Technical Report

May 2018

Massachusetts Department of Elementary and Secondary Education
75 Pleasant Street, Malden, MA 02148-4906
Phone 781-338-3000 TTY: N.E.T. Relay 800-439-2370
www.doe.mass.edu

This document was prepared by the
Massachusetts Department of Elementary and Secondary Education
Jeffrey Riley
Commissioner

The Massachusetts Department of Elementary and Secondary Education, an affirmative action employer, is committed to ensuring that all of its programs and facilities are accessible to all members of the public.
We do not discriminate on the basis of age, color, disability, national origin, race, religion, sex, gender identity, or sexual orientation.
Inquiries regarding the Department's compliance with Title IX and other civil rights laws may be directed to the
Human Resources Director, 75 Pleasant St., Malden, MA 02148-4906. Phone: 781-338-6105.

This document printed on recycled paper

Massachusetts Department of Elementary and Secondary Education
75 Pleasant Street, Malden, MA 02148-4906
Phone 781-338-3000 TTY: N.E.T. Relay 800-439-2370
www.doe.mass.edu

Content

1. Purpose of this Report

The purpose of this technical report is to provide reliability and validity evidence to support the use of ESE's educator preparation 2017 stakeholder surveys in the evaluation of educator preparation providers' programs. This technical report delivers an executive summary of the 2017 validity data. For a full understanding of the program approval and readiness constructs measured and the Rasch methodology used to assess the validity of the constructs, readers should review the technical report created for the 2016 surveys.

2. Survey Specifications

There are two types of items in the surveys:

1. Items that align to observable outcomes relative to the Program Approval Standards and Review Criteria (henceforth called program approval construct items), which set forth expectation for providers.
2. Items that align to observable practices within the Professional Standards of Teaching (PSTs), which define the pedagogy, knowledge, and skills required of all teachers (readiness construct items).

The survey specification for each survey in 2017 is provided in Table 1. The total number of program approval items increased from 26 in 2016 to 28 items in 2017.The total number of items measuring the PSTs decreased from 42 in 2016 to 32 in 2017 for the teacher candidate (TCD), teacher completer (TCP) and supervising practitioner (SP), respectively. The new teacher hiring principal (NTHP) and teacher-of-record hiring principal (TRHP) surveys retained 6 items.

Table 1: Test specification for scalable measures

Item Type	Topic area	TCD[1]	TCP[1]	SP[1]	NTHP[1]	TRHP[1]
Program Approval Criteria	Program Experience	11	11	NA	NA	NA
	Course Work	3	3			
	Field-base Experience	5	5			
	Supervision[2]	5	5			
	Assessment	4	4			
	Item total	28	28			
Readiness construct: Professional Standards of Practice (PSTs)	Standard I[2]	8	8	8	6	6
	Standard II[2]	14	14	14		
	Standard III[2]	5	5	5		
	Standard IV[2]	5	5	5		
	Item total	32	32	32	6	6

[1] TCD (Teacher Candidate); TCP (Teacher Completer); SP (Supervising Practitioner); NTHP (New Teacher Hiring Principal; TRHP (Teacher-of-Record Hiring Principal: [2]Standard I (Curriculum, Planning and Assessment); Standard II (Teaching All Students); Standard III (Family and Community Engagement); Standard IV (Professional Culture).

3. Survey Response Rates

ESE administered four educator preparation stakeholder surveys in the spring of 2017 to evaluate the perceptions of their preparation programs and teacher readiness in the Commonwealth. ESE surveyed the following stakeholder groups:

1. **Teacher Candidate (TCD) survey,** issued to candidates at the point of program completion
2. **Teacher Completer (TCP) survey,** issued to educators employed in a Massachusetts public school one year after completing a preparation program
3. **Supervising Practitioner (SP) survey,** issued to educators who served as a supervisor to a candidate during the practicum experience
4. **Hiring Principal (NTHP or TRHP) surveys,** issued to principals one year after hiring a preparation program completer

Table 2 provides response rates for each of the 2017 surveys included in this validity study.

Table 2: 2017 Survey Response Rates

Survey	Stakeholder Groups	Survey Topic	Total number of respondents	Number included in validity study	Percent included in study
TCD	Teacher Candidate	Program Criteria	813	813	100%
		PST	813	363	42%
TCP	Teacher Completer	Program Criteria	433	433	100%
		PST	433	218	50%
SP	Supervising Practitioner	PST	931	472	51%
NTHP	New Teacher Hiring Principal	PST	602	600	99%
TRHP	Teacher-of-Record Hiring Principal	PST	459	459*	100%

* There were only 239 hiring principals with productive measures (52%)

In contrast to 2016, TCD, TCP and SP respondents were given the option to complete the PSTs related items. As a result, response rates for PST items were halved in 2017 for these stakeholder groups.

4. Program Approval Criteria survey data

The program approval validity data is summarized in Appendix B for the teacher candidate (TCD) and teacher completer (TCP) surveys. The psychometric properties of the TCD and TCP program approval items are sound.

Content. In both surveys, the clear majority of items fit the Rasch model well. These data support the content validity aspect of construct validity.

Substantive. The rating scale for both surveys is monotonic. The step thresholds are narrow due to respondents' infrequent use of scoring category 2 (Neither agree nor disagree). The item hierarchies replicate the results from 2016 with items measuring TCDs' and TCPs' overall experience harder to endorse than items related to course work or supervision. These data support the substantive validity aspect of construct.

Generalizability. TCD items were invariant across years: the correlation of item deltas between the two years was 0.92. Due to an administration error in 2016, there is no comparable analysis of the TCP survey. Differential item functioning (DIF) analyses indicated there were two items with mild DIF (>0.5 logits but <0.64) in the TCD survey; there was no item DIF related to the TCP survey. Items were largely invariant across years and subgroups. The non-extreme real person separation reliability (henceforth PSR) for the TCD and TCP surveys is 0.80 and 0.84, respectively. Extreme persons make up 10% of respondents. The targeting of the survey is off-center with the person means in both surveys over 2 logits above the item mean (set to 0.00 logits). However, these data combined provide evidence to support the generalizability of the program approval construct.

Structural. The variance explained by the TCD and TCP surveys is 38% and 39%, respectively. Supervisory items break out to form the 1st contrast which explains over 5% of the variance in both surveys. The supervisory items, however, fit the Rasch model well (outfit statistics are less than 1.5 logits) and the point-to-measure (PTM) correlations are all above 0.3. There is sufficient evidence to support the structural validity of the program approval items.

External. The responsiveness and concurrent validity of the instruments were assessed. Both surveys, on average, can differentiate respondents into approximately three statistically distinct scoring groups (high, medium, and low). The correlation between TCD and TCP program approval scores and associated readiness construct scores are 0.68 and 0.67, respectively. The limited validity evidence supports the external validity of the program approval scores.

5. Performance Standards of Teaching (PSTs) survey data

Respondents in each group (TCD, TCP, and SP) responded to the same 32 items. The readiness construct validity data are summarized in Appendix C. The psychometric properties of the TCD, TCP, and SP items are sound.

Content. Except for two items on the TCP survey, the technical quality of the items is good with all items fitting the Rasch model well. These data support the content validity of the readiness construct.

Substantive. The rating scale for the three surveys is used as intended by survey developers. The step thresholds on the TCP survey are narrow due to respondents' infrequent use of scoring category 2 (Neither agree nor disagree). The item hierarchies on each survey replicated the results from 2016 with items measuring "family engagement", on average, harder to endorse than items measuring ‘teaching to all students” which, in turn, are harder to endorse than “professional culture” items. These data support the substantive validity aspect of the readiness construct.

Generalizability. Items on each survey were invariant across years: the correlation of item deltas between the two years was equal to or above 0.90. Item invariance was also evident in the DIF analyses. When using teacher-of-record as the indicator, there was no DIF present for the TCP and SP surveys. However, there were three items with severe DIF (>0.8 logits) in the TCD survey; these items should be considered for removal. With this exception, items were largely invariant across years and subgroups for each survey. The PSR for the TCD, TCP, and SP surveys is 0.88. 0.89, and 0.93, respectively; extreme persons make up 18%, 15%, and 9% of respondents in the TCD, TCP, and SP surveys, respectively. The difference between person and item means is over 2 logits in each survey. However, these data combined provide the evidence to support the generalizability of the readiness construct.

Structural. The variance explained by the readiness construct items is 53%, 48%, and 57% for the TCD, TCP and SP surveys, respectively. Similar to 2016, family engagement items break out to form the 1st contrast. The family engagement items, however, fit the Rasch model well (outfit statistics are less than 1.5 logits) and the PTM correlations are all above 0.5. The sub-scale correlations between measures of the four standards are 0.5 or greater for each survey; the readiness construct explains the positive relationship between the standards. The structural validity evidence for the readiness construct is strong across the three surveys.

External. All three surveys, on average, can minimally differentiate respondents into the three statistically distinct scoring groups (high, medium, and low), with the SP survey able to differentiate 5 separate groups. The correlation between TCD and TCP readiness scores and their program approval construct scores are 0.68 and 0.67, respectively. The limited validity evidence supports the external validity of the readiness construct scores.

6. Hiring Principal survey data

The hiring principal validity data is summarized in Appendix D for the new teacher candidate (NTHP) and teacher-of-record (TRHP) surveys. Each survey will be discussed separately.

NTHP survey

Content. The technical quality of the six items is good; there are no misfitting items in the NTHP survey and PTM correlations are above 0.8. These data support the content validity aspect of construct validity.

Substantive. The rating scale is monotonic and threshold steps clearly differentiate the construct continuum. The item hierarchies replicate the results from 2016 with items measuring Standard I harder to endorse than Standard II items, which are, in turn, harder to endorse than the Standard IV items. These data support the substantive validity aspect of readiness construct.

Generalizability. The six items were invariant across years: the correlation of item deltas between the 2016 and 2017 was 1.00. Items were also invariant across subgroups. The PSR for the NTHP survey was 0.93; extreme persons make up 8% of respondents (4% low; 4% high). The targeting of the survey is optimal with the person mean 0.27 below the item mean (set to 0.00 logits). These data support the generalizability of the readiness construct.

Structural. The variance explained by the NCHP survey is 80.3%. The six items form a unidimensional scale. These data support the structural validity of the readiness construct.

External. The NTHP survey is responsive; it can, on average, differentiate respondents into the 5 to 6 statistically distinct scoring groups. However, the external validity evidence is limited and more data is needed to examine the relationship between other criterion measures and NTHP scores.

TRHP survey

The psychometric properties of the TRHP survey items are problematic for some aspects of construct validity. There was insufficient sample size (N = 239) for some of the validity analyses; due to the number of extreme scores, only half of respondents had productive measures.

Content. There are no misfitting items in the TRHP survey. The technical quality of the six items are good with point-to-measure correlations above 0.8. These data support the content validity aspect of construct validity.

Substantive. The rating scale is monotonic and threshold steps clearly differentiate the construct continuum. However, hiring principals do not use score category 0 (decline) or score category 1 (none). The item hierarchies do not replicate the results from 2016. These data partially support the substantive validity aspect of the readiness construct.

Generalizability. The six items were not invariant across years: the correlation of item deltas between the 2016 and 2017 was 0.65. The compression of the variance in item difficulties and the small sample size likely contributed to item invariance. There was no item DIF related to the INTACT indicator; one item exhibited severe DIF using the READY indicator. The PSR for the TRHP survey was 0.86; *extreme persons make up 47% of respondents*. The targeting of the items is detrimental to precisely measuring respondent views; the person mean was 6.84 logits above the item mean (set to 0.00 logits). These data bring into doubt the generalizability of the readiness construct for this respondent group.

Structural. The variance explained by the TRHP surveys is 63%. The six items appear to form a unidimensional scale, however the data are problematic. The variance explained by the six items is only 1.7% of the total variance making it hard to assess the constructs unidimensionality. The structural validity of the readiness construct for this group of respondents is inconclusive.

External. The TRHP survey is responsive; it can, on average, differentiate respondents into 3 to 4 statistically distinct scoring groups. However, the external validity evidence is limited and more data is needed to examine the relationship between other criterion measures and TRHP scores.

7. Conclusion and Recommendations

Program Approval construct items. Overall, the construct validity data is sufficient to support score use in assessing stakeholder (TCD and TCP) views of their preparation programs in general and, more specifically, their views on their coursework, field-based experiences, supervision and assessment. The 28-item program approval items largely meet the assumptions of the Rasch model; they are well-fitting, reliable and unidimensional. The program approval scores can be used for their intended purpose.

Readiness (PST) construct items. Overall, the construct validity data is sufficient to support score use in assessing stakeholder (TCD, TCP and SP) views of readiness to teach in general and, more specifically, their views on the four professional standards of teaching. The 32-item readiness items largely meet the assumptions of the Rasch model; they are well-fitting, reliable and unidimensional. The readiness scores can be used for their intended purpose.

Hiring Principal surveys. The construct validity of the 6 items measuring hiring principal views of new teacher hires (NTHP) meets the assumption of the Rasch model and the readiness scores can be used for their intended purpose. The data needed to support the construct validity of the 6 items measuring hiring principal views of teacher-of-record (TRHP) hires are inconclusive. Like 2016, ceiling effects impact the validity of the scale; this survey should be revised before the scores are used for their intended purpose.

Appendices

Appendix A: Validity criteria used to assess construct validity

Validity Aspect	Statistic/Data	Cutoff Criteria or Typical Standard
Content	Point-to-measure correlation	• Positive and >0.3
Content & Structural	Infit/Outfit Mean-Square error fit statistics (MNSQ)	• MNSQ error fit statistics of between 0.5 – 1.5 • Disruption of pattern in magnitude of misfit
Substantive	Rating scale functioning	• Minimum of 10 responses per category. • Observed score averages and Andrich item threshold parameters increase monotonically • Outfit Mean Square error of less than 1.5 for each threshold • Threshold steps are greater than 0.8 logits
	Item difficulty hierarchy	• Ordering of item deltas corresponds to theoretical expectations (item/person variable maps)
Generaliz-ability	Item difficulty (Delta) invariance and Differential Item Functioning (DIF)	• Within standard error, items retain same item difficulty (deltas) across administrations and survey forms • Correlation of mean item difficulties between survey forms or administrations is greater than or equal to 0.9 • For DIF, recommended criteria vary: delta difference of <0.3 – 0.64 logits (0.5 used in study)
	Person Separation Reliability (PSR) and targeting.	• Medium stakes ~ 0.8 • High stakes > 0.9 • Item and person distributions overlap on scale metric axis (means less than one logit apart)
Structural	Sub-scale correlations	• Correlations are positive and substantial (> 0.5 but < 0.9)
	Principal component analyses of the residuals to assess unidimensionality	• Total variance explained: >40% very good; >50% excellent • 1st contrast: < 5% of total variance; • 1st contrast: Eigen value < 3 • Item variance 4x variance of 1st contrast • Cluster correlations ♦ > 0.82 likely only one latent trait ♦ > 0.71 more dependency than independence
External	Responsiveness	• Typical ~ 3-person strata (low, medium, high) • Person strata = (4 person separation index +1)/3
	Concurrent validity	• Correlational relationship between readiness construct scores or program approval scores and criterion measures

Appendix B: Program criteria validity evidence

Validity Aspect	Statistic/Data	Teacher Candidate (TCD) N = 813	Teacher Completer (TCP) N = 433
Content	Point-to-measure correlation	• 0.30– 0.66	• 0.30– 0.68
Content & Structural	Infit/Outfit Mean-Square error fit statistics (MNSQ)	• Infit: 0.65– 1.51 • Outfit: 0.60 – 1.81 • 2 items under fit model	• Infit: 0.56– 1.60 • Outfit: 0.51 – 1.79 • 2 items under fit model
Substantive	Rating scale functioning	• Monotonic observed averages; disordinal Andrich thresholds • Threshold MNSQ for Cat0 is above 1.5 • Narrow step thresholds	• Monotonic • Thresholds' MNSQ good • Narrow step thresholds
	Item difficulty hierarchy	• As expected	• As expected
Generalizability	Item difficulty (Delta) invariance and Differential Item Functioning (DIF)[1]	• Delta correlation 2016/2017: 0.92 • Teacher-of-Record: Two items mild DIF • YrComp: No DIF	• Delta correlation 2016/2017: Not available • trecord: No DIF • YrComp: No DIF
	Person Separation Reliability (PSR) and targeting.	• Non-extreme: Real (0.80); Model (0.83) • With extremes: Real (0.74); Model (0.75) • Extreme person mean is 2.21 logits above the item mean; 10% extreme	• Non-extreme: Real (0.84); Model (0.86) • With extremes: Real (0.78); Model (0.79) • Extreme person mean is 2.10 logits above the item mean; 10% extreme
Structural	Sub-scale correlations	• Not Applicable	• Not Applicable
	Principal component analyses	• Variance explained: 38.4% • Supervision items associate with 1st contrast (mild multidimensionality)	• Variance explained: 39.4% • Supervision items associate with 1st contrast (mild multidimensionality)
External	Responsiveness	• Person strata: 2 to 3	• Person strata: 3

[1] DIF variables: - trecord: candidates who were employed as a teacher-of-record before or during the program were compared to new teacher candidates; YrComp: candidates that took 1-2 years to complete the program were compared to candidates who took 3+ years to complete it.

Appendix C: Performance Standards of Teaching (PSTs) validity evidence

Validity Aspect	Statistic/Data	Teacher Candidate TCD, N = 363	Teacher Completer TCP, N = 218	Superv. Practitioner[1] SP N = 472
Content	Point-to-measure correlation	• 0.46 – 0.74	• 0.46 – 0.71	• 0.56 – 0.74
Content & Structural	Infit/Outfit Mean-Square error fit statistics (MNSQ)	• Infit: 0.60 – 1.52 • Outfit: 0.62 – 1.57 • No misfit	• Infit: 0.63 – 1.54 • Outfit: 0.49 – 2.00 • 2 items misfit	• Infit: 0.72 – 1.48 • Outfit: 0.74 – 1.33 • No misfit
Substantive	Rating scale functioning	• Monotonic • Thresholds' MNSQ good • Threshold steps good	• Monotonic • Cat0 threshold MNSQ: 1.81 • Threshold steps low differentiation	• Monotonic • Thresholds' MNSQ good • Threshold steps good
	Item difficulty hierarchy	• As expected	• As expected	• As expected
Generaliz-ability	Item invariance and Differential Item Functioning (DIF)[1]	• Correlation 2016/2017: 0.95 • trecord: Three items severe to moderate DIF	• Correlations 2016/2017: ≥ 0.90 • No DIF but insufficient sample size	• Correlation 2016/2017: 0.98 • No DIF
	Person Separation Reliability (PSR) and targeting.	• Non-extreme: Real (0.88); Model (0.89) • With extremes: Real (0.80); Model (0.80) • Extreme person mean is 3.24 logits above the item mean; 18% extreme	• Non-extreme: Real (0.89); Model (0.91) • With extremes: Real (0.83); Model (0.84) • Extreme person mean is 2.78 logits above the item mean; 15% extreme	• Non-extreme: Real (0.93); Model (0.94) • With extremes: Real (0.89); Model (0.90) • Extreme person mean is 2.57 logits above the item mean; 9% extreme
Structural	Sub-scale correlations	• Std.III/Std.IV: 0.50 - • Std.I/Std.IV: 0.75	• Std.I/Std.IV: 0.51 - • Std.I/Std.II: 0.76	• Std.III/Std.IV: 0.54 - • Std.I/Std.II: 0.86
	Principal component analyses	• Variance explained: 52.9% • Unidimensional	• Variance explained: 47.7% • Unidimensional	• Variance explained: 57.1% • Unidimensional
External	Responsiveness	• Person strata: 3 to 4	• Person strata: 3 to 4	• Person strata: 4 to 6

[1] Suprvising Practitioner; [2] DIF variables: - TCD and TCP: TRECORD: candidates who were employed as a teacher-of-record before or during the program were compared to new teacher candidates; YrComp: candidates that took 1-2 years to complete the program were compared to candidates who took 3+ years to complete it. For supervising practitioner survey: announced and unannounced number of observations were used as DIF indicators (0-2 versus 3+); teaching experience (0-15 versus 16+), and number of candidates supervised (1-2 versus 3+).

Appendix D: Hiring Principal survey validity evidence

Validity Aspect	Statistic/Data	NTHP (N = 600)	TRHP (N = 239) *After removing 2 misfitting persons*
Content	Point-to-measure correlation	• 0.89 – 0.93	• 0.85 – 0.89
Content & Structural	Infit/Outfit Mean-Square error fit statistics (MNSQ)	• Infit: 0.83 – 1.46 • Outfit: 0.76 – 1.55 • No misfit	• Infit: 0.84 – 1.15 • Outfit: 0.68 – 1.00 • No misfit
Substantive	Rating scale functioning	• Monotonic • Thresholds' MNSQ good • Threshold steps good	• Observed averages monotonic • Andrich thresholds monotonic – low use Cat0 and Cat1 • Thresholds' MNSQ good • Threshold steps good
	Item difficulty hierarchy	• As expected	• Item hierarchy is compressed with little to no variation in mean deltas
Generaliz-ability	Item invariance and Differential Item Functioning (DIF)[1]	• Correlation 2016/2017: 1.00 • No DIF	• Correlation 2016/2017: 0.65 • DIF: one item READY • No DIF: INTACT • Insufficient sample size
	Person Separation Reliability (PSR) and targeting.	• Non-extreme: Real (0.93); Model (0.95) • With extremes: Real (0.94); Model (0.96) • Extreme person mean is -0.27 logits below the item mean; 8% extreme	• Non-extreme: Real (0.86); Model (0.87) • With extremes: Real (0.78); Model (0.79) • Extreme person mean is 6.84 logits above the item mean; 47% extreme
Structural	Sub-scale correlations	• Not Applicable	• Not Applicable
	Principal component analyses	• Variance explained: 80.3% • Unidimensional • Item variance is only 4.5% of total variance	• Variance explained: 62.6% • Unidimensional: undetermined • **Item variance is only 1.7% of total variance**
External	Responsiveness	• Person strata: 5 to 6	• Person strata: 3 to 4

[1] DIF variables: - INTACT: depth of interaction (minimal to moderate versus substantial to very extensive) and READY: ready to meet needs of students (fully ready versus not to mostly ready)