Westat®

Improving lives through research™

Development and Validation of a Novice Teacher and Supervisor Survey

Authors

Matthew Finster

August 2017

Prepared by:
Westat
An Employee-Owned Research Corporation®
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Summary

This brief presents the Westat and AACTE Novice Teacher and Novice Teacher Supervisor Surveys and presents initial evidence about their reliability and validity. The Novice Teacher and Novice Teacher Supervisor Surveys assess how well prepared novice teachers are to meet the job requirements of teaching. The surveys are designed to provide educator preparation programs (EPPs) with information that can be used to enable program improvement, provide information for consumers, and/or ensure program accountability (Feuer, Floden, Chudowsky, & Ahn, 2013). The survey items are adapted from InTASC's Model Core Teaching Standards and are grouped into four domain areas: the Learner and Learning, Content, Instructional Practice, and Professional Responsibility.

Online surveys were administered to novice teachers (recent graduates) statewide in Iowa and from multiple EPPs across several states in the spring of 2017. The response rate for the novice teacher survey was 19.7 percent (440 of 2,226), and for the novice teacher supervisor/principal survey the response rate was 22.4 percent (499 of 2,226). The analytic sample for the novice teacher survey is n=433 and for the novice teacher supervisor survey n=480. To assess the structural (or factorial) validity of the survey, we conducted confirmatory factor analyses (CFAs). The results of the CFAs provide evidence that the 17 survey items cover the four intended domains and that each item is strongly related to the underlying domain. The model fit indices indicate that the hypothesized structure of model fits well with the data and that factor loadings for the subscales load significantly and moderately to strongly on each of their respective domains. Furthermore, the four factors, namely, The Learner and Learning, Content, Instructional Practice and Professional Responsibility, load significantly onto one second-order factor, referred to as Teacher Practice. Moreover, a multigroup analysis demonstrates that the instrument operates equivalently across the two groups.

This brief presents evidence of the construct validity of the survey instruments, but validating a survey is an ongoing process (Messick, 1995). Hence, we will continue to work to collect and develop a body of evidence that supports the use of these surveys.

Background

The trend for establishing policies that hold educators accountable for their performance has increased; likewise, calls have shifted to focus on holding educator preparation programs (EPPs)[1] accountable for their graduates' performance. This effort has been spurred on by reports such as the controversial report by the National Council for Teacher Quality (NCTQ), which called teacher education an "industry of mediocrity" (Greenberg, Mckee, & Walsh, 2013, p. 1). While acknowledging some of problems that persist for EPPs, more positive views comment on the difficulty and complexity of preparing educators; for example, Monk (2009) commented that EPPs are "focused on some of the most intellectually complex and interesting phenomena imaginable" (p. 256). Following calls for increased accountability, accreditation agencies (e.g., the Council for Accreditation of Educator Preparation [CAEP]) have bolstered their requirements, demanding EPPs demonstrate more and more evidence of performance, and in the fall of 2016, the federal government put forth a set of regulations proposing unprecedented levels of accountability for EPPs. While these regulations were later repealed, they underscore the general trend for holding EPPs accountable for their graduates' performance.

Graduate and employer surveys offer one method to assess novice teachers' perceptions of training and performance in job-related demands, such as classroom management, planning and delivering instruction, assessment of student learning, and professional responsibilities (Feuer et al., 2013). Surveys offer an important perspective of those closest to the application of the skills that preservice teachers are intended to develop: the teachers themselves and their supervisors.

While the link between teachers' responses on these surveys and student learning has not been thoroughly examined, Boyd et al. (2009) did find a link between EPP graduates' survey responses and impact on student learning. Nonetheless, a common and apt critique is that the surveys alone provide insight only about a graduate's *perceptions* of preparation and program quality, not *actual* preparation and program quality (Darling-Hammond, 2006). This point underscores the importance of combining survey data with other measures of EPP quality. That is, survey data should be triangulated with other data to validate findings.

[1] Also commonly referred to more narrowly as teacher preparation programs (TPPs).

Data and Methods

Survey Development and Administration

For the purposes above, the American Association of Colleges for Teacher Education (AACTE), Westat, and representatives from multiple state education agencies (SEAs) and EPPs met to develop and administer a novice teacher and novice teacher supervisor survey based on the InTASC standards (see Council of Chief State School Officers [CCSSO], 2011).

The structure of the survey instruments closely mirrors the structure of the InTASC standards. The InTASC Core Teaching Standards are grouped into four general categories—The Learner and Learning, Content, Instructional Practice, and Professional Responsibility—with each general category having multiple corresponding standards. The survey instruments have the same four domains and have items under each domain that are intended to reflect each standard, and/or are more broadly intended to cover the teachers' requisite knowledge, dispositions, and performance in each respective domain. For example, for The Learner and Learning (domain 1), there are three InTASC standards: (1) learner development, (2) learning differences, and (3) learning environments (InTASC standards #1-3). This structure is reflected in the survey, which also has three items that are meant to cover domain 1. The three items in domain 1 inquire about a teacher's ability to design and implement learning experiences, ensure an inclusive learning environment, and develop and maintain a positive learning environment. For Content (domain 2), InTASC has two standards: (1) content knowledge and (2) application of content (InTASC standards #4-5). The survey instruments have three items that are intended to cover Content. The three items inquire about a teacher's ability to demonstrate understanding of content area, make a discipline accessible, and integrate cross-disciplinary skills. For Instructional Practice (domain 3), InTASC has three standards: (1) assessment, (2) planning for instruction, and (3) instructional strategies (InTASC standards #6-8). The survey instruments have three corresponding items that are reflective of assessment, planning, and instructional strategies, as well an additional two items that inquire about differentiating instruction and the use of technology. For Professional Practice (domain 4) there are two standards: (1) professional learning and ethical practice and (2) leadership and collaboration (InTASC standards #9-10). The survey has four items that are meant to cover the professional practice domain by inquiring about engaging in professional learning, evaluating outcomes of teaching, reflecting on practice, and working collaboratively with colleagues. (See CCSSO, 2011, for a discussion of the InTASC model core teaching standards. And for a review of the literature pertaining to the content validity of the InTASC standards, see Youngs, 2011.)

The Novice Teacher and Novice Teacher Supervisor surveys are each a 17-item instrument structured on a 4-point Likert-type scale that ranges from 1 (*not at all prepared*) to 4 (*very well prepared*). As discussed above, they are each composed of four subscales, each measuring one facet of teacher practice: (1) The Learner and Learning, (2) Content, (3) Instructional Practice, and (4) Professional Responsibility. The Learner and Learning domain consists of three items (Q1-Q3). The Content domain consists of three items (Q4-6). The Instructional Practice domain consists of seven items (Q7-Q13), and the Professional Practice domain consists of four items (Q14-Q17).[2] Please refer to Exhibits 1 and 2 for copies of the survey.

An online survey was administered to recent graduates (also referred to as novice or beginning teachers) and their supervisors; of participating EPPs in New York, Kansas, Maine, and Wyoming; and was administered statewide in Iowa.[3] Respondent contact information was provided by staff at the respective EPPs and the Iowa Department of Education. The online survey was open for several months (beginning of April to end of June). The response rate for the novice teacher survey was 19.7 percent (440 of 2,226), and for the novice teacher supervisor/principal survey the response rate was 22.4 percent (499 of 2,226). The analytic sample for the novice teacher survey was n=433 and for the novice teacher supervisor survey, it was n=480. (The results of the novice teacher supervisor survey are presented in Appendix A. Results of a multigroup analysis, which demonstrates the factorial equivalence of the surveys across the two groups and provides further evidence of the construct validity of the surveys, are presented in appendix B.)

A variety of statistical techniques were used to examine the reliability and validity of the survey.[4] The analyses focused on the factorial validity of the survey instruments, which is also indicative of construct validity. To validate a survey instrument, one should establish a body of evidence that includes content, construct, convergent, and predictive validity.[5] Content validity is based on the relevance of content of the survey and representativeness to which it covers the domain (Messick, 1989). Construct validity refers to the extent that the scales measure the underlying construct that it is intended to measure. The relationship between the survey scale items and the subscales can be examined via factor analysis, which investigates whether the factor structure of the scale is consistent

[2] The survey also includes four open-ended questions that were excluded from the factor analysis.

[3] Staff from the individual institutes in New York, Kansas, Maine, and Wyoming opted to participate in the survey and sent the online survey to their respective recent graduates and their supervisors; whereas in Iowa, staff from the Iowa Department of Education sent the survey to all novice teachers in Iowa state and their supervisors.

[4] All analyses presented were conducted in Mplus 7 using Maximum Likelihood (ML) unless otherwise noted.

[5] Convergent and predictive validity are not addressed in this study. Convergent validity is demonstrated by scores on a measure being related to scores on another measure that it is intended to measure a similar construct. Similarly, predictive validity is demonstrated by a measure predicting future scores on a measure of a similar construct.

with the theorized structure—in this case, the structure of the survey instrument hypotheses that there are four distinct domains and that the respective survey items load distinctly onto each domain. The CFAs test the extent to which this hypothesized model fits empirically with the data. Combined with evidence of content validity, the results of the CFAs can provide further evidence of construct validity of an instrument by demonstrating that the survey items are in fact empirically related to the respective domains. Content validity demonstrates that the items and domains are theoretically related, and factorial validity demonstrates that the items and domains are empirically related to each other. This information combined is indicative of construct validity.

Descriptive and Correlational Analysis

The descriptive statistics for the novice teacher survey are presented in Table 1. The means for the items range from 2.25 (Q12) to 3.37 (Q17). Since CFA and SEM require multivariate normal data, it is important to check measures of univariate nonnormality. Kurtosis provides one measure of nonnormality of the data. While there is no consensus regarding problematic kurtosis values (Kline, 2011, as cited in Byrne, 2012), values ranging from ±2.0 to ±7.0 have been proposed as initial points of nonnormality (Muthén & Kaplan, 1985; West et al., 1995; as cited in Byrne, 2012). Almost all the values for kurtosis, excluding Q1, are ±2.0, so kurtosis values do not appear problematic for CFA.[6]

Bivariate correlations were used to examine the relationships between the survey items (Q1-Q17). A strong majority of the correlations (94%) were moderate in size, r=0.3-0.6. All correlations were significant. The correlations are presented below in Table 2.

Table 1. Descriptive statistics for novice teacher survey (n=433)

Variable/sample	Mean	Std. Deviation	Variance	Skewness	Kurtosis
Q1	2.98	0.65	0.42	-0.65	2.33
Q2	3.03	0.71	0.50	-0.70	1.86
Q3	3.20	0.72	0.52	-0.83	1.60
Q4	3.11	0.69	0.47	-0.49	0.85
Q5	2.98	0.71	0.50	-0.40	0.60

[6] To assess whether data might be multivariate nonnormally distributed, the chi-square values of models using ML and MLM—maximum likelihood parameter estimates with standard errors and a mean-adjusted chi-square test statistic that are robust to non-normality—estimators were compared based on Byrne's (2012) recommendation. Although the chi-square value did decrease using MLM estimation, nonetheless ML is fairly robust to minor nonnormality (Byrne, 2012), so the analysis proceeded using ML estimation. *Note.* MLM estimation requires listwise deletion in Mplus so for analysis using MLM n=431.

Table 1. Descriptive statistics for novice teacher survey (*n*=433) (continued)

Variable/sample	Mean	Std. Deviation	Variance	Skewness	Kurtosis
Q6	2.93	0.71	0.50	-0.34	0.51
Q7	2.88	0.74	0.54	-0.16	-0.05
Q8	3.07	0.74	0.54	-0.53	0.55
Q9	3.11	0.70	0.48	-0.62	1.27
Q10	2.79	0.86	0.74	-1.02	1.96
Q11	2.68	0.79	0.63	-0.03	-0.10
Q12	2.25	0.87	0.76	0.14	-0.18
Q13	3.08	0.77	0.60	-0.45	-0.07
Q14	3.15	0.69	0.48	-0.51	0.43
Q15	2.90	0.76	0.58	-0.23	-0.26
Q16	3.36	0.67	0.45	-0.95	1.52
Q17	3.37	0.71	0.51	-1.11	1.77

NOTE.: Sample sizes vary by item from 427 to 433 due to missing data patterns.

Table 2. Correlations of novice teacher survey item responses (*n*=433)

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
Q1	1.00																
Q2	0.63	1.00															
Q3	0.58	0.60	1.00														
Q4	0.47	0.33	0.34	1.00													
Q5	0.56	0.50	0.55	0.56	1.00												
Q6	0.53	0.44	0.44	0.50	0.59	1.00											
Q7	0.41	0.33	0.39	0.39	0.39	0.41	1.00										
Q8	0.47	0.31	0.38	0.43	0.38	0.41	0.43	1.00									
Q9	0.54	0.50	0.52	0.46	0.58	0.50	0.48	0.49	1.00								
Q10	0.49	0.54	0.45	0.30	0.48	0.44	0.36	0.31	0.49	1.00							
Q11	0.50	0.52	0.49	0.24	0.49	0.44	0.35	0.31	0.50	0.63	1.00						
Q12	0.39	0.37	0.32	0.27	0.32	0.38	0.33	0.33	0.37	0.40	0.54	1.00					
Q13	0.37	0.29	0.38	0.44	0.42	0.39	0.35	0.46	0.42	0.25	0.21	0.30	1.00				
Q14	0.51	0.47	0.47	0.44	0.47	0.41	0.41	0.42	0.53	0.42	0.35	0.31	0.33	1.00			
Q15	0.46	0.42	0.43	0.43	0.45	0.42	0.44	0.39	0.50	0.39	0.35	0.31	0.35	0.63	1.00		
Q16	0.41	0.38	0.46	0.34	0.38	0.37	0.35	0.33	0.45	0.25	0.26	0.22	0.30	0.56	0.49	1.00	
Q17	0.43	0.33	0.40	0.38	0.36	0.38	0.37	0.33	0.45	0.32	0.31	0.30	0.38	0.54	0.49	0.56	1.00

NOTE.: The default option in Mplus 7 uses all available information to calculate correlations, but does not provide significant tests. All correlations are significant at $p<0.01$ in SPSS using both pairwise and listwise deletion.

Confirmatory Factor Analysis

To test the factorial validity, a first- and second-order confirmatory factor analysis (CFA) was conducted based on the conceptual framework of the survey. The framework postulates a four-factor structure model consisting of (1) The Learner and Learning, (2) Content, (3) Instructional Practice, and (4) Professional Responsibility. This hypothesis is based on the intended relationship between the items that were constructed to provide indicators of multiple related behaviors or skills in each of these four domains The first-order CFA tests for the validity of the factorial structure of the survey by assessing the extent that the items designed to measure the respective factors (e.g., latent constructs) actually do so. The second-order CFA adds an overarching factor (a second-order factor) and tests the structural relationship between the four factors (The Learner and Learning, Content, Instructional Practice, and Professional Responsibility) and an overarching domain (Teacher Practice). In other words, the model hypothesizes that responses can be explained by four first-order factors and one second-order factor in which covariation among the first-order factors is explained by regression on the second-order factor (i.e., Teacher Practice).

The findings indicate that the first- and second-order factor CFAs, based on the hypothesized models, appropriately represent the data and provide evidence to support the structure of the survey instrument and its subscales. Based on the substantive theory and model fit indices, the second-order model is preferable to the first-order model estimates; hence, while the model fit indices are presented for both models, only the item estimates, along with a schematic diagram, are presented for the second-order factor analysis. The model fit indices and model estimates are discussed below.[7]

The goodness-of-fit indices of the first- and second-order CFAs indicate a reasonably good fit between the hypothesized model and observed data, based on model fit recommendations (e.g., Brown & Cudeck, 1993; Hu & Bentler, 1999). The comparative (also referred to as incremental) indices—the CFI and TLI—for both models are in the acceptable range (Bentler, 1990). Finally, the absolute fit indices—SRMR and RMSEA—are in the adequate to well-fitting range (Brown & Cudeck, 1993; Hu & Bentler, 1999). Given the results of the goodness-of-fit indices, no further modifications to the model were considered.

While both CFAs demonstrate reasonable fit, the second-order CFA is preferable based on the substantive theory of all the first-order factors being explained by a single second-order factor of teacher practice and has comparable, if not slightly improved, comparative (e.g., CFI and TLI),

[7] Model results are similar for the novice teacher supervisor/principal survey CFAs, which are presented in the appendix.

absolute (e.g., SRMR, RMSEA), and predicative (e.g., AIC, BIC) fit statistics. Note, whereas comparative fit indices increase as goodness-of-fit improves, absolute fit indices decrease as goodness-of-fit improves (Browne et al., 2002).

Table 3. Model fit indices for novice teacher survey CFAs (*n*=433)

Tests of Model Fit	Obtained values	
	4-Factor CFA	2nd-Order CFA
Chi-square test of model fit		
Value	382.39	382.81
Degrees of freedom	113.00	115.00
p-value	<.001	<.001
Chi-square test of model fit for the baseline model		
Value	3372.04	3372.04
Degrees of freedom	136.00	136.00
p-value	<.001	<.001
Bentler Comparative Fit Index (CFI)	0.92	0.92
Tucker-Lewis Index (TLI)	0.90	0.90
Loglikelihood		
H_0	-6569.03	-6569.24
H_1	-6377.83	-6377.83
Information criteria		
Number of free parameters	57.00	55.00
Akaike (AIC)	13252.06	13248.48
Bayesian (BIC)	13484.09	13472.37
Sample size adjusted BIC	13303.20	13297.83
Root mean square error of approximation (RMSEA)		
Estimate	0.07	0.07
90 percent C.I.	0.07-0.08	0.07-0.08
Probability RMSEA <=.05	<.001	<.001
Standardized root mean square residual (SRMR)	0.048	0.048

Table 4 presents the standardized survey item loadings on the factors for the second-order CFA.[8] All of the item loadings (parameter estimates) are all statistically significant and load on the factors at 0.53 or above. The item loadings ranges for each of the factors are: (1) The Learner and Learning, 0.75 to 0.81, (2) Content, 0.67 to 0.81, (3) Instructional Practice, 0.53 to 0.77, and (4) Professional Responsibility, 0.69 to 0.81. The standardized factor loadings (F1-F4) on the second-order factor of Teacher Practice are all significant and range from 0.82 to 0.96 (Table 5). The second-order CFA is depicted below in Figure 1. All estimates are standardized.

[8] The estimates are STDYX standardized, which is based on background and outcome variables. All variables are rescaled to have a variance of 1.00 (Byrne, 2012).

Table 4. Novice teacher survey item loadings by factor (*n*=433)

Factor	Factor loading	Standard error
The Learner and Learning (F1)		
Q1	0.81	0.02
Q2	0.76	0.03
Q3	0.75	0.03
Content (F2)		
Q4	0.67	0.03
Q5	0.81	0.02
Q6	0.74	0.03
Instructional Practice (F3)		
Q7	0.60	0.04
Q8	0.60	0.04
Q9	0.77	0.02
Q10	0.67	0.03
Q11	0.68	0.03
Q12	0.55	0.04
Q13	0.53	0.04
Professional Responsibility (F4)		
Q14	0.81	0.02
Q15	0.75	0.03
Q16	0.69	0.03
Q17	0.69	0.03

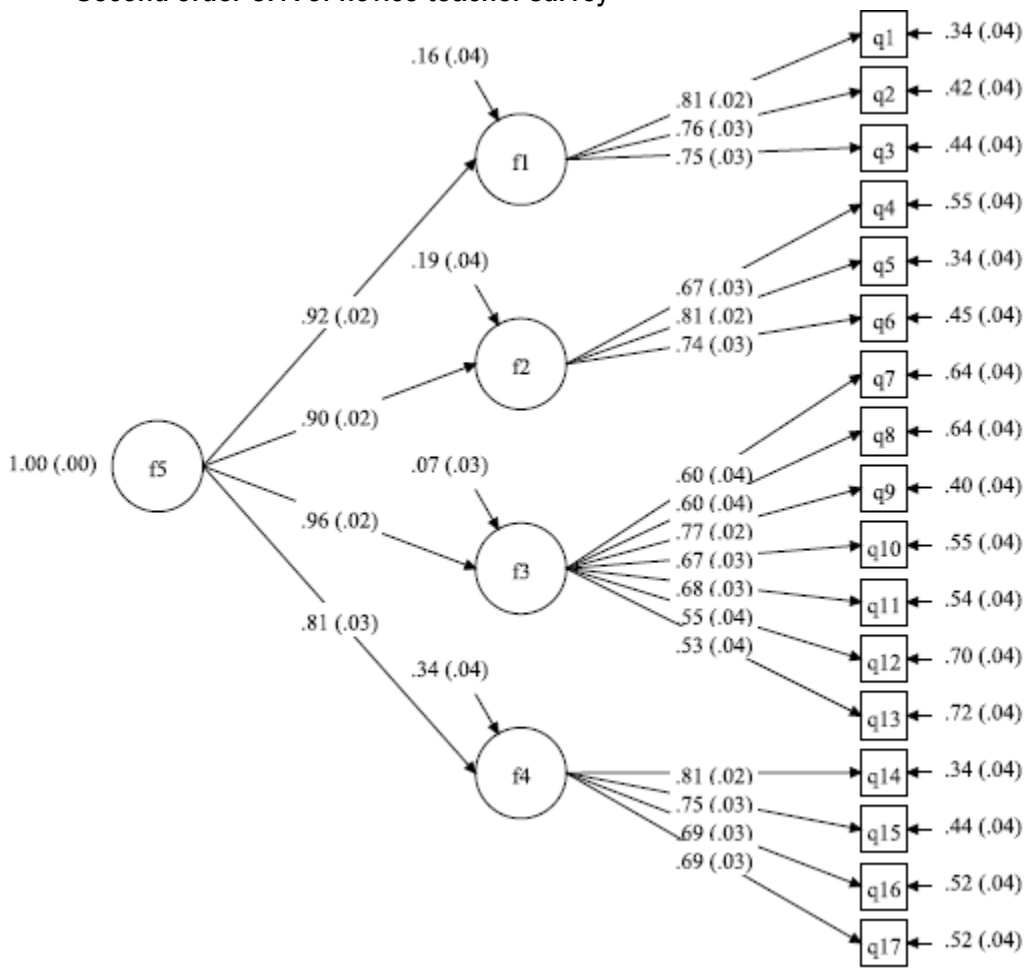
NOTE.: All factors loadings are significant at p<.001. Estimates are STDYX standardized, based on background and outcome variables.

Table 5. Factor loadings on teaching practice (2nd-Order factor) (*n*=433)

Teaching Practice	Factor loading	Standard error
The Learner and Learning (F1)	0.92	0.02
Content (F2)	0.90	0.02
Instructional Practice (F3)	0.96	0.02
Professional Responsibility (F4)	0.82	0.03

NOTE.: All factors loadings are significant at p<.001. Estimates are STDYX standardized, based on background and outcome variables.

Figure 1. Second-order CFA of novice teacher survey



NOTE.: All factors loadings are significant at p<.001. Estimates are STDYX standardized, based on background and outcome variables. The residual variances (error variances) are indicated in parentheses.

Conclusions

Novice teacher and novice teacher supervisor/principal surveys offer another measure of EPP quality and have the potential to enable program improvement by providing information back to EPPs regarding their graduates' performance in different aspects of teaching and to increase program accountability by holding EPPs accountable for their graduates' aggregate performance on such measures. This analysis presents some initial evidence of the validity of the AACTE and Westat novice teacher and novice teacher supervisor/principal surveys. The analysis sought to determine the extent that the survey items measured the factors of the survey, namely, The Learner and Learning, Content, Instructional Practice, and Professional Responsibility. The results of CFAs provide evidence of the factorial structure of the measuring instruments. Combined with the substantive and underlying theory of the InTASC standards (e.g., Youngs, 2011), this empirical analysis provides some initial evidence of the construct validity of the AACTE and Westat novice teacher and novice teacher supervisor/principal surveys.

Limitations of the Analysis

Establishing ture measure validity is an impossible task (Kane, 2006); instead, evidence needs to be compiled that supports the use of the survey. Thus, "validating" a survey is an ongoing process. While this brief cites evidence of the surveys' content validity and presents evidence of the surveys' structural/factorial validity, further evidence that would substantiate the "validity" of the surveys should include information pertaining to convergent and predicative validity—that is, evidence that the scores on this survey are similar to other instruments that intend to measure teacher practices (e.g., educator observation tools, student surveys) and that the scores on this survey are predictive of future scores on measures of related constructs. In addition to examining other types of validity, it would be also be beneficial to examine the invariance of the surveys across specific groups of teachers (e.g., elementary vs. secondary teachers), to show evidence that the relationship between survey items and domain functions similarly across group, which contributes further to the reliability and validity of the surveys.

Exhibit 1. Novice teacher survey

Novice Teacher Survey					
	Survey item	Response scale			
Domain	In your FIRST year of teaching, how well prepared were you to:	Not at all prepared	Somewhat prepared	Well prepared	Very well prepared
Domain 1: The Learner and Learning	1.1 Design and implement developmentally appropriate learning experiences for all learners. (Q1) 1.2 Ensure an inclusive learning environment for all learners. (Q2) 1.3 Develop and maintain a positive learning environment that engages all learners. (Q3)				
Domain 2: Content	2.1 Demonstrate understanding of content area by using central concepts, tools of inquiry, and structures of your discipline. (Q4) 2.2 Make your discipline accessible and meaningful for learners. (Q5) 2.3 Integrate cross-disciplinary skills (e.g., critical thinking, problem solving, creativity, communication) to help learners use content. (Q6)				
Domain 3: Instructional Practice	3.1 Develop and use multiple methods of assessment. (Q7) 3.2 Plan for instruction aligned to content standards. (Q8) 3.3 Use a variety of instructional strategies appropriately. (Q9)				

Exhibit 1. Novice teacher survey (continued)

Novice Teacher Survey					
	Survey item	Response scale			
Domain	In your FIRST year of teaching, how well prepared were you to:	Not at all prepared	Somewhat prepared	Well prepared	Very well prepared
Domain 3: Instructional Practice	3.4 Differentiate instruction for all learners. (Q10) For students with disabilities. (Q11) For English language learners. (Q12) 3.5 Use technology in the classroom appropriately to support instruction. (Q13)				
Domain 4: Professional Responsibility	4.1 Engage in ongoing professional learning to provide all learners with engaging learning experiences. (Q14) 4.2 Evaluate outcomes of teaching using a variety of data (e.g., systematic observation, information about learners, research) to adapt planning and practice. (Q15) 4.3 Reflect on teaching practice to improve instruction. (Q16) 4.4 Work collaboratively with colleagues to meet the needs of all learners. (Q17)				
Comments and recommendations	What were the strengths of your preparation program? In what areas, if any, do you wish your program had prepared you more effectively? What recommendations do you have for your preparation program? Additional comments?	Open ended			

Exhibit 2. Novice teacher supervisor survey

Novice Teacher Supervisor Survey					
	Survey item	Response scale			
Domain	How well does the novice teacher perform each of the following:	Not very well	Somewhat well	Well	Very well
Domain 1: The Learner and Learning	**1.1** Design and implement developmentally appropriate learning experiences for all learners. (Q1) **1.2** Ensure an inclusive learning environment for all learners. (Q2) **1.3** Develop and maintain a positive learning environment that engages all learners. (Q3)				
Domain 2: Content	**2.1** Demonstrate understanding of content area by using central concepts, tools of inquiry, and structures of your discipline. (Q4) **2.2** Make your discipline accessible and meaningful for learners. (Q5) **2.3** Integrate cross-disciplinary skills (e.g., critical thinking, problem solving, creativity, communication) to help learners use content. (Q6)				
Domain 3: Instructional Practice	**3.1** Develop and use multiple methods of assessment. (Q7) **3.2** Plan for instruction aligned to content standards. (Q8) **3.3** Use a variety of instructional strategies appropriately. (Q9) **3.4** Differentiate instruction for all learners. (Q10) For students with disabilities. (Q11) For English language learners. (Q12) **3.5** Use technology in the classroom appropriately to support instruction. (Q13)				

Exhibit 2. Novice teacher supervisor survey (continued)

Novice Teacher Supervisor Survey					
	Survey item	Response scale			
Domain	How well does the novice teacher perform each of the following:	Not very well	Somewhat well	Well	Very well
Domain 4: Professional Responsibility	**4.1** Engage in ongoing professional learning to provide all learners with engaging learning experiences. (Q14) **4.2** Evaluate outcomes of teaching using a variety of data (e.g., systematic observation, information about learners, research) to adapt planning and practice. (Q15) **4.3** Reflect on teaching practice to improve instruction. (Q16) **4.4** Work collaboratively with colleagues to meet the needs of all learners. (Q17)				
Comments and recommendations	In what areas, if any, do you think the preparation program should have prepared the novice teacher more effectively? What recommendations do you have for the novice teacher's preparation program? Additional comments?	Open ended			

References

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wycoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation Policy Analysis*, *31*(4), 416–440.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Brown, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Browne, M. W., MacCallum, R. C., Kim, C.-T., Anderson, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*, 403–412.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* New York, NY: Routledge, Taylor & Francis Group.

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

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for evaluating program outcomes. *Journal of Teacher Education*, *57*(1), 1–19.

Feuer, M. J., Floden, R. E., Chudowsky, N., & Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options.* Washington, DC: National Academy of Education.

Greenberg, J., Mckee, A., & Walsh, K. (2013). Teacher prep review: A review of the nation's teacher preparation programs. Washington, DC: National Council on Teacher Quality.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement equivalence in aging research. *Experimental Aging Research*, 18*,* 117–144.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York, NY: American Council on Education/Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement Issues and Practice. 14*, 5–8.

Monk, D. H. (2009). Reaction from an education school dean. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 251–258). Washington, DC: The Urban Institute Press.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–234). Washington, DC: American Psychological Association.

Youngs, P. (2011, April). *InTASC Model Core Teaching Standards: Research synthesis*. Retrieved from http://www.ccsso.org/Resources/Digital_Resources/InTASC_Research_Synthesis.html

Appendix A
Results of Novice Teacher Supervisor/Principal Survey

The results of the second-order CFA using the novice teacher supervisor/principal survey data are presented below. As mentioned above, the online survey was administered to recent graduates and their supervisors from participating EPPs in New York, Kansas, Maine, and Wyoming, and was administered statewide throughout Iowa. The analytic sample for the novice teacher supervisor survey was *n*=480. The results of the CFA using the supervisor/principal survey data are very similar to the results of the CFA using the teacher survey data, which are discussed above. The tests of model fit indicate that the model adequately fits the data. All survey items significantly load onto their respective factors and all factors significantly and substantially load onto the second-order factor (i.e., teacher practice).

Additional analyses and model modifications were also conducted with the supervisor survey data to test for robustness. One way to test for multivariate nonnormal data is to compare chi-square values using ML and MLM estimators (Byrne, 2012), with substantial differences indicating MLM should be used. Similar to the novice teacher model, the chi-square value did decrease. Nonetheless, ML estimation is considered robust to minor data nonnormality (Byrne, 2012), so the model results presented are based on ML estimation. Further modifications to the model were also tested. For example, based on theory, the modification indices (MIs) and expected parameter changes (EPCs) cross-loadings across the observable items were added to the model (e.g., Q1 with Q3, Q3 with Q2, Q10 with Q11, and Q11 with Q12). While this modification improved model fit (e.g., CFI value >.95), the parameter estimates of the associated items onto the respective factors decreased slightly. Thus, the original hypothesized model is preferable and more parsimonious. The results of the hypothesized second-order CFA are presented below, including the descriptive statistics, correlations, model fit indices, factor loadings estimates (Tables A-1 through A-4), and the schematic model (Figure A-1).

Table A-1. Descriptive statistics for novice teacher supervisor survey (*n*=480)

Variable	Mean	Std. Deviation	Variance	Skewness	Kurtosis
Q1	3.21	0.72	0.51	-0.64	0.21
Q2	3.29	0.75	0.56	-0.84	0.21
Q3	3.32	0.76	0.58	-0.89	0.20
Q4	3.21	0.75	0.55	-0.60	-0.20
Q5	3.18	0.74	0.55	-0.60	-0.05
Q6	3.01	0.79	0.63	-0.37	-0.50

Table A-1. Descriptive statistics for novice teacher supervisor survey (*n*=480) (continued)

Variable	Mean	Std. Deviation	Variance	Skewness	Kurtosis
Q7	2.99	0.76	0.58	-0.27	-0.55
Q8	3.24	0.75	0.56	-0.93	1.19
Q9	3.09	0.83	0.70	-0.94	1.37
Q10	2.97	0.94	0.88	-1.07	1.53
Q11	3.03	0.87	0.76	-0.82	0.80
Q12	2.73	1.07	1.14	-1.04	0.81
Q13	3.16	0.77	0.60	-0.66	0.16
Q14	3.37	0.71	0.50	-1.19	2.23
Q15	3.02	0.77	0.59	-0.49	0.07
Q16	3.24	0.76	0.57	-0.85	0.48
Q17	3.47	0.74	0.54	-1.34	1.51

NOTE.: Sample sizes vary by item from 479 to 480 due to missing data patterns.

Table A-2. Correlations of novice teacher supervisor survey item responses (n=480)

Items	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
Q1	1.00																
Q2	0.69	1.00															
Q3	0.70	0.76	1.00														
Q4	0.70	0.61	0.58	1.00													
Q5	0.73	0.67	0.70	0.71	1.00												
Q6	0.67	0.63	0.60	0.68	0.69	1.00											
Q7	0.65	0.57	0.62	0.61	0.60	0.66	1.00										
Q8	0.59	0.52	0.50	0.63	0.59	0.58	0.65	1.00									
Q9	0.62	0.58	0.59	0.61	0.61	0.64	0.62	0.58	1.00								
Q10	0.55	0.56	0.54	0.51	0.58	0.52	0.61	0.45	0.53	1.00							
Q11	0.54	0.60	0.55	0.52	0.56	0.59	0.61	0.51	0.56	0.65	1.00						
Q12	0.43	0.45	0.42	0.40	0.40	0.46	0.49	0.42	0.46	0.50	0.62	1.00					
Q13	0.49	0.38	0.40	0.50	0.46	0.54	0.51	0.51	0.53	0.42	0.41	0.39	1.00				
Q14	0.56	0.48	0.52	0.56	0.52	0.52	0.54	0.52	0.52	0.41	0.40	0.29	0.39	1.00			
Q15	0.60	0.56	0.59	0.60	0.57	0.62	0.66	0.56	0.57	0.60	0.53	0.42	0.43	0.62	1.00		
Q16	0.56	0.56	0.57	0.57	0.57	0.59	0.56	0.55	0.51	0.48	0.48	0.34	0.45	0.63	0.68	1.00	
Q17	0.53	0.58	0.61	0.53	0.51	0.48	0.47	0.49	0.46	0.45	0.47	0.33	0.37	0.57	0.60	0.65	1.00

NOTE.: The default option in Mplus 7 uses all available information to calculate correlations, but does not provide significant tests. All correlations significant at p<.01 in SPSS using both pairwise and listwise deletion.

Table A-3. Model fit indices of second-order CFA of novice teacher supervisor survey (*n*=480)

Indices	Values
Chi-square test of model fit	
Value	448.99
Degrees of freedom	115.00
p-value	<0.001
Chi-square test of model fit for the baseline model	
Value	5949.65
Degrees of freedom	136.00
p-value	<0.001
Bentler Comparative Fit Index (CFI)[1]	0.943
Tucker-Lewis Index (TLI)[2]	0.932
Loglikelihood	
H_0	-6870.86
H_1	-6646.36
Information criteria	
Number of free parameters	55.00
Akaike (AIC)	13851.72
Bayesian (BIC)	14081.28
Sample size adjusted BIC	13906.71
Root mean square error of approximation (RMSEA)[3]	
Estimate	0.078
90 percent C.I.	0.070-0.085
Probability RMSEA <=.05	<0.001
Standardized root mean square residual (SRMR)[4]	0.037

NOTE.: [1] Values 0.90 to 0.95 indicative of acceptable fit (Bentler, 1990). [2] Values 0.90 to 0.95 indicative of acceptable fit (Bentler, 1990). [3] Values <0.08 indicative of adequate fit (Brown & Cudeck, 1993) and values 0.80 to 0.10 indicative of mediocre fit (MacCallum et al., 1996). [4] Values <0.05 indicative of well-fit (Byrne, 2012) and values <0.08 indicative of acceptable fit (Hu & Bentler, 1999).

Table A-4. Supervisor survey item loadings by factor (*n*=480)

Factor	Factor loading	Standard error
The Learner and Learning (F1)		
Q1	0.86	0.02
Q2	0.84	0.02
Q3	0.84	0.02
Content (F2)		
Q4	0.82	0.02
Q5	0.85	0.02
Q6	0.83	0.02
Instructional Practice (F3)		
Q7	0.82	0.02
Q8	0.74	0.02
Q9	0.77	0.02
Q10	0.73	0.02
Q11	0.76	0.02
Q12	0.61	0.03
Q13	0.62	0.03
Professional Responsibility (F4)		
Q14	0.76	0.02
Q15	0.83	0.02
Q16	0.83	0.02
Q17	0.75	0.02

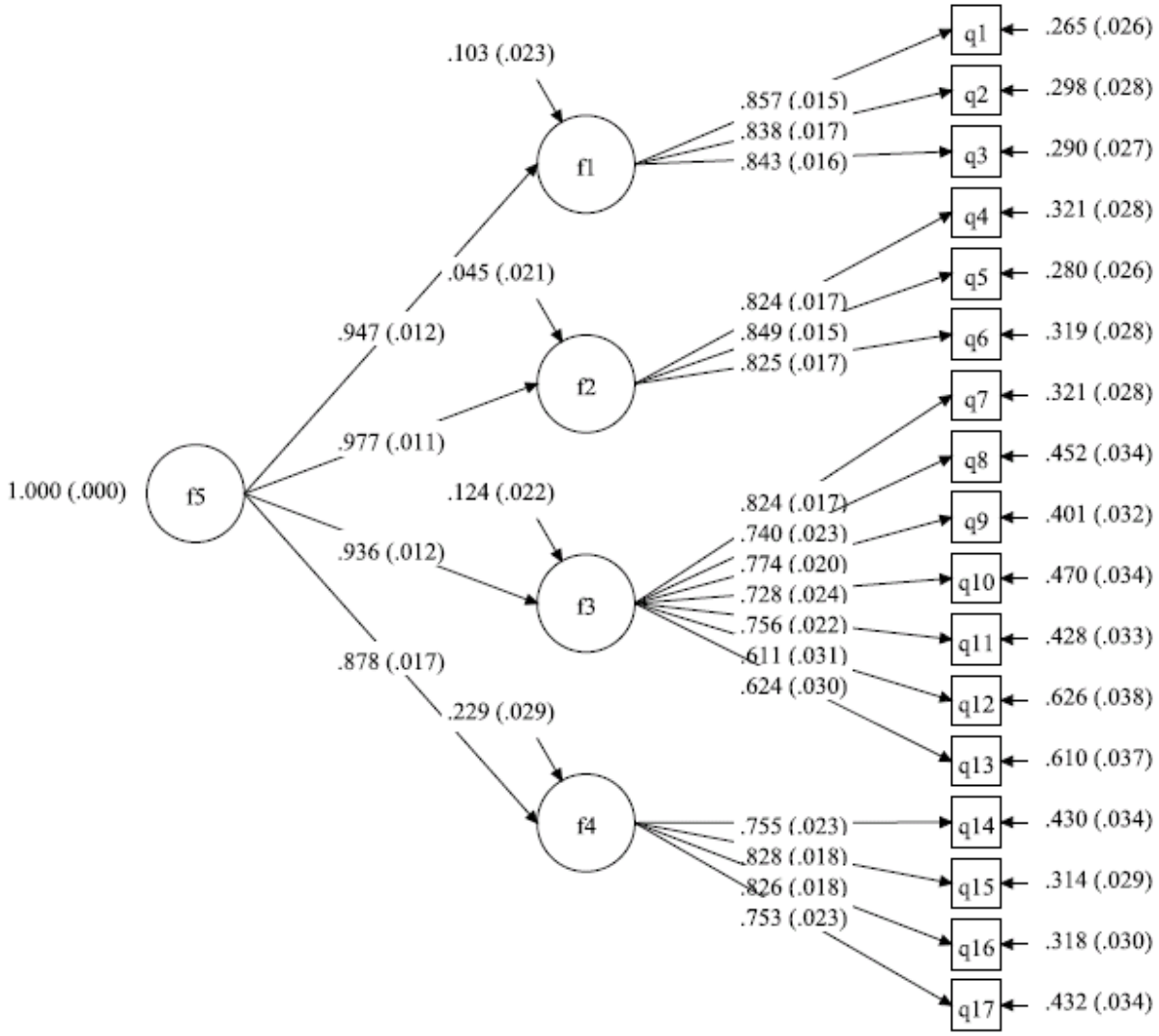
NOTE.: All factors loadings are significant at p<.001. Estimates are STDYX standardized, based on background and outcome variables.

Table A-5. Supervisor survey factor loadings on teaching practice (Second-order factor) (*n*=480)

Teaching Practice	Factor loading	Standard error
The Learner and Learning (F1)	0.95	0.01
Content (F2)	0.98	0.01
Instructional Practice (F3)	0.94	0.01
Professional Responsibility (F4)	0.88	0.02

NOTE.: All factors loadings are significant at p<.001. Estimates are STDYX standardized, based on background and outcome variables.

Figure A-1. Second-order CFA model of novice teacher supervisor/principal survey



NOTE.: All factors loadings are significant at p<.001. Estimates are STDYX standardized, based on background and outcome variables. The residual variances (error variances) are indicated in parentheses.

Appendix B
Multigroup Invariance: Testing Construct Validity Across Teachers and Supervisors

A multigroup structural equation modeling approach was used to compare equivalence of the factorial structure of the survey instruments based on responses from novice teachers and novice teacher supervisors/principals. The purpose of the multigroup analysis is to test whether components of the measurement and structural model are equivalent (also referred to as invariant) across the two groups. This analysis provides additional evidence of construct validity by demonstrating that survey items operate similarly in relation to the intended domains across the two groups, which further demonstrates that the items are related to the intended constructs.

Testing for invariance across the groups can provide further evidence of construct validity by demonstrating that across two different groups, the survey items and their relationships to the domains are similar to each other.[9] In testing for invariance across groups, sets of parameters are gradually tested in an increasingly restrictive fashion. The first level of measurement invariance is configural invariance, in which only the number of factors and loading patterns must be constant. The next level of invariance, factorial invariance, requires the factor loadings to be equivalent. Once factorial invariance is established, the invariance of the structural model can be tested. As equivalence is demonstrated across each of parameters, a higher level of invariance is met (Meredith, 1993; Widaman & Reise, 1997). [10]

The testing of invariance across the two groups proceeded as follows: (1) baseline CFA models were established, (2) a configural model was tested, (3) the first invariant model was tested by constraining the factors loadings equivalent, and (4) the second invariant model was tested by constraining the factor loadings equivalent. Baseline models were established for each group based on the models

[9] Generally, multigroup analysis is used in circumstance when there are different group responses on the same instrument. In this instance, although there are two versions of the survey, the structure of the survey instruments are identical, so a multigroup analysis is appropriate.

[10] Widaman & Reise (1997) distinguish between weak, stong and strict factorial invariance, the last of which requires the factor loadings, indicator intercepts, and indicator residual variances to be the same. Given the purpose of this analysis—that is, to examine construct validity of the surveys across two groups—not all eligible parameters that are eligible for tests of invariance were examined (e.g., item intercepts, factor means). For demonstrating construct validity, group differences in intercepts and the latent means are of no particular interest (Byrne, 2012) and hence are not tested in this analysis.

previously discussed in the brief.[11] Then, the configural invariance was tested (Horn & McArdle, 1992). In the configural model, the number of factors and loading patterns are constant, but no equality constraints were imposed on any of the parameters. That is, factor loadings and structural paths were freely estimated across groups.[12] Next, to test for factorial invariance, equivalence across the measurement model was tested by constraining the item loadings to be equal (i.e., invariant) across groups (Invariant Model_v1). The results of Invariant Model_v1 were compared to results of the configural model to assess invariance across the factor loadings. Finally, structural invariance was tested by constraining the structural pathways to be equal (Invariant Model_v2). The results of Invariant Model_v2 were compared to the results of Invariant Model_v1 to assess invariance across the structural paths. The results of the models are discussed below.

The configural model with the number of loadings and factors constant, but all parameters freely estimated in the two groups fit the data well (CFI =0.955, SRMR =0.036) according to fit criteria suggested by Hu and Bentler (1999). This model provides evidence of configural invariance, which indicates there are same number of factors and factor loading patterns in the models across the two groups. To test for the next level of invariance, factorial invariance, the next model (i.e., Invariant Model_v1) constrained the factor loadings equal across the groups. The chi-square from the model with all parameters allowed to be unequal across groups (i.e., configural model) was compared to the chi-square from the model with only the loadings constrained to be equal across groups (Invariant Model_v1). Differences between the chi-square for each model that are not statistically significant indicate that the additional parameters constrained to be equal are in fact equivalent (invariance) across the groups. The invariant model with loadings constrained to be equal across groups (i.e., Invariant Model_v1) also fit the data well (CFI =0.955., SRMR = 0.043.), and compared to the configural model, the difference in chi-square was not statistically significant, $\Delta\chi2(12) = 14.303$, $p > .05$. The comparison of the two models indicates that survey items operate equivalently across the two groups.[13] To test the equivalence (invariance) of the structural model across the two groups, the chi-square from the model with only the loadings constrained to be equal across groups (i.e., Invariant Model_v1) was compared to the chi-square from the model with the structural paths also constrained to be equal across groups (i.e., Invariant Model_v2). The invariant model with structural paths constrained equal (i.e., Invariant Model_v2) also fit the data well (CFI =0.955., SRMR

[11] The models were slightly modified from the previous CFAs with the addition of cross-loadings between several of the items (noted below in Table B-1).

[12] Latent means were constrained to zero.

[13] Based on the modification indices and fit indices, one modification was made to the model. Item 16 (i.e., reflect on teaching practice to improve instruction) on factor 4 (Professional responsibility) was freely estimated (not constrained equal across groups). This modification indicates that within this dataset the relationship between item 14 and factor 4 is slightly different for teachers vs. supervisors.

=0.047.), and compared to the model that constrained the factor loadings equal (i.e., Invariant Model_v1), was not statistically significant, $\Delta\chi2(3) = 5.144$, $p > .05$. These results indicate that, in addition to the items operating equivalently across the two groups, the structural paths are also equivalent across the groups.[14]

This multigroup analysis provides additional evidence of the construct validity of the survey instrument by demonstrating multigroup configural, factorial, and structural equivalence (invariance), further suggesting that items are related to the intended constructs.

Table B-1. Fit statistics of multigroup models (Configural and invariant models)

Tests of model fit	Configural model	Invariant model_v1.	Invariant model_v2.
Notes	Addition of cross-loadings: Q10 with Q11; Q10 with Q12; Q11 with Q12. No parameters constrained equal.	Factor loadings constrained equal. One freely estimated parameter: F4 by Q16.	Factor loadings (excluding one parameter) and structural paths constrained equal
Number of free parameters	116	104	101
Chi-square test of model fit			
Value	628.513	642.816	647.96
Degrees of freedom	224	236	239
p-value	<0.001	<0.001	<0.001
Δχ2	na	14.303	5.144
Δ Degrees of freedom	na	12	3
Contribution from each group:			
Supervisor	354.18	358.868	360.481
Teacher	274.334	283.948	287.479
Bentler Comparative Fit Index (CFI)[1]	0.955	0.955	0.955
Root mean square error of approximation (RMSEA)[2]	0.063	0.061	0.061
Standardized root mean square residual (SRMR)[3]	0.036	0.043	0.047

NOTES: * $p<.05$. [1] Values 0.90 to 0.95 are indicative of acceptable fit (Bentler, 1990), values >0.95 are indicative of a well-fit model (Hu & Bentler, 1999).[2] Values <0.08 are indicative of adequate fit (Brown & Cudeck, 1993) and values 0.80 to 0.10 are indicative of mediocre fit (MacCallum et al., 1996).[3] Values <0.05 are indicative of a well-fit model (Byrne, 2012; Hu & Bentler, 1999).

[14] Results of comparisons of the nested models also held based on MLR estimation—maximum likelihood estimation with robust standard errors—in addition to ML estimation.