Data Analyses on the

Utah Preservice Teacher Observation Protocol (UPTOP)

Lisa McLachlan

Brigham Young University

September 22, 2017

Successful teacher education programs combine clinical experiences with academic curriculum to provide pre-service teachers opportunities to connect theory and practice (Hammerness, Darling-Hammond, & Shulman, 2002; Ball & Bass, 2000, Grossman & Stodolsky, 1995). However, simply providing these opportunities is not enough. Teacher candidates need to be observed and supervised by quality cooperating teachers and evaluated using performance assessments created by research-based practices. Performance assessments based on nationally recognized teaching standards can contribute to program development and provide valuable feedback to teacher candidates (Darling-Hammond, 2012).

This study provides evidence of reliability and validity of a performance assessment adopted by a state consortium of teacher preparation programs. The instrument was developed by experts and stakeholders within the consortium to evaluate teaching effectiveness based on national teaching standards. By aligning performance assessment to national standards, evaluations can help beginning teachers improve their practice in ways that continue after the assessment experience has ended. Performance assessments can also support exemplary practice, pedagogical learning, and inform ongoing program improvement.

Theoretical Framework

Two features that are important in successful teacher preparation programs is: 1) a common, clear vision of good teaching that permeates all coursework and clinical experiences, and 2) well-defined standards of professional practice and performance that are used to guide and evaluate coursework and clinical work (Darling-Hammond, 2012). The consortium of university preparation programs in this study share a common vision of good teaching and are guided by the Interstate Teacher Assessment and Support Consortium (InTASC) Core Teaching standards (Council of Chief State School Officers, 2011). The InTASC teaching standards are grouped into four general categories: the learner and learning, content, instructional practice, and professional responsibility.

The Utah Preservice Teacher Observation Protocol (UPTOP) was created by the Teacher Education Assessment and Accreditation Council for evaluating teacher candidates in college teacher preparation programs in the state of Utah. The UPTOP is based on the state's Effective Teaching Standards, which align with the national InTASC teaching standards (Council of Chief State School Officers, 2011). These standards reflect current research on effective instruction and demonstrate the knowledge and skills necessary to teach the Utah Core Standards. Items on the UPTOP were reviewed by experts in the field and piloted at seven of the ten university preparation programs within the state consortium.

Methods

This study analyzed data collected at three universities: Brigham Young University (BYU), Utah State University (USU), and Utah Valley University (UVU). The UPTOP used in the BYU and UVU samples contains 22 items, measuring three factors or latent constructs: Learner and Learning, Instructional Practice, and Professionalism. The UPTOP used in the USU sample contains 20 items, measuring the same three factors as the BYU and UVU versions. The UPTOP items and factors are listed in Table 5.

During the student teacher or internship experience, the evaluator will score the teacher candidate using the *Not Present* (0), *Beginning* (1), *Emerging* (2), or *Preservice Effective* (3) response categories on the rubric. Using these four levels, the total number of points possible is 66. A student must achieve 80 percent of the total score, which would be 53 points to pass their student teaching or internship experience and have no items scored at the *Not Present* (0) level.

Participants

Participants included 948 early childhood, elementary, secondary, and special education teacher candidates completing a student teaching or internship experience at three university teacher preparation programs during the 2016 – 2017 school year. During their field experience, each teacher candidate was evaluated twice, once by a mentor or cooperating teacher (CT) and once by a university supervisor (US). These evaluations were completed toward the end of the students' teaching experience, but were not necessarily completed on the same day.

Some teacher candidates in the BYU and USU samples were evaluated three or four times, depending on individual situations and/or major requirements. For example, some teachers were evaluated by a CT and a US in both their major and minor subject areas, some were evaluated by more than one CT at the schools they taught, and some were evaluated more than once if they failed or performed poorly on the first observation. In the BYU sample, about 5% of individuals had more than two UPTOP evaluations, for a total of 991 observations. In the USU sample, approximately 31% of individual teachers had more than two UPTOP evaluations, for a total of 953 observations. All individuals in the UVU sample had two and only two UPTOP evaluations, for a total of 480 observations (see Table 1).

Table 1. *Descriptive statistics of sample UPTOP observations.*

	Brigham Young University	Utah State University	Utah Valley University
Total number of observations	991	952	480
Elementary Education	306	479	240
Secondary Education	526	371	240
Special Education	93	102	N/A
Early Childhood Education	66	N/A	N/A

*N/A = not available

Analyses

Data were analyzed using SPSS and Mplus statistics software. A series of confirmatory factor analyses (CFA) were conducted using Mplus statistical software on the factor structure and measurement model of the instrument within each school sample. Because several of the item responses were not normally distributed, the variables were treated as categorical and estimated using WLSMV (Muthen, 2017). McDonald's (1985, 1999) omega (Ω) was used for estimating the reliability of scores instead of Cronbach's alpha (α) to correct for the multidimensionality and nestedness within the measurement models (Brunner, Nagy, & Wilhelm, 2012; Gignac, 2015; Reise, 2012).

Multiple measurement invariance tests were conducted between the two groups of raters (cooperating teacher, CT and university supervisor, US) within each school sample. The 'cluster' command in Mplus was used to account for the multiple observations or clusters of data around individual teachers in each sample.

Findings

A series of paired-sampled t-tests were performed to compare summed scores of the cooperating teachers (CT) to the university supervisors (US) within each school. Results of these analyses are presented in Table 2. Results indicate there was not a significant difference between the summed scores of the CT and the US within the BYU and UVU samples. However, there was a significant difference in summed scores at USU between the CT (M=56.22, SD=5.68) and the US (M=57.31, SD=4.929); *t* (492) = -4.275, *p* < *.001*. These results suggest that the mentor teachers and the university supervisors at USU may be interpreting the items on the UPTOP differently.

Table 2. *Results of paired-sample t-tests of UPTOP summed scores*

	Mean	SD	*t-value*
Brigham Young University (BYU)			
cooperating teacher	59.81	6.81	.16
university supervisor	59.85	5.74	---
Utah State University (USU)			
cooperating teacher	56.22	5.68	4.72*
university supervisor	57.31	4.93	---
Utah Valley University (UVU)			
cooperating teacher	63.89	5.48	-0.038
university supervisor	63.87	4.20	---

*significant at *p < .001*

A CFA conducted on the BYU sample analyzing the three-factor model based on theoretical assumptions produced reasonably good model-fit statistics (χ^2 = 841.067, RMSEA = .056, CFI = .938, TLI = .930). However, the three-factors were highly correlated, with correlations ranging from .767 to .966. Given the degree of correlation between the factors, it was reasonable to determine whether a hierarchical model would account for the estimated correlation between the first-order factors.

A series of CFAs and chi-square difference tests were conducted on the BYU and USU samples to compare nested models: the bifactor model; the model with three correlated, first-order factors; and the model with a single first-order factor. Results of these analyses are present in Table 3 and 4. Unfortunately, a model with a second-order factor and three first-order factors was just identified and could not be compared in these analyses.

Table 3. *Model-fit indices and χ2 difference tests of nested models for Brigham Young University.*

Model	χ2	*df*	Δ χ2	Δ *df*	*p*-value	RMSEA	CFI	TLI
Bifactor	555.454	187	---	---	---	0.045	0.964	0.955
Three, first-order factors	841.066	206	258.823	19	0.000	0.056	0.938	0.930
Single, first-order factor	913.797	209	67.526	3	0.000	0.059	0.931	0.924

Table 4. *Model-fit indices and χ2 difference tests of nested models for Utah State University.*

Model	χ2	*df*	Δ χ2	Δ *df*	*p*-value	RMSEA	CFI	TLI
Bifactor	259.017	150	---	---	---	0.028	0.992	0.990
Three, first-order factors	325.807	167	70.392	17	0.000	0.032	0.988	0.986
Single, first-order factor	324.474	170	17.605	3	0.000	0.033	0.987	0.986

The model-fit indices and the results of the chi-square difference tests on these samples suggested that the bifactor model fits the data better than the other nested models. For the BYU sample, the bifactor model value for Ω was .980, indicating that approximately 98.0% of the variance in the total items was due to the general factor. For the USU sample, the bifactor model value for Ω was .988, indicating that approximately 98.8% of the variance in the total items was due to the general factor. The standardized factor loadings for the bifactor model for the BYU and USU samples are presented in Table 5 and 6, respectively.

Multiple measurement invariance tests were conducted between the two groups of raters (cooperating teacher and university supervisor) within each school sample. Measurement invariance was obtained within the RATER group in the BYU and USU samples. These results suggest the factor structure and scale of the item responses does not differ between the two groups of raters (CT and US) at Brigham Young University and Utah State University. Unfortunately, the small sample size collected from Utah Valley University prevented a reliable analysis of the measurement model and invariance test for the raters at this school.

Discussion

The results of this study provide evidence of reliability and validity of scores from two samples of the UPTOP performance assessment used to evaluate teacher candidates in a consortium of college preparation programs in the state of Utah. Given that the bifactor model fits data better than the other tested models and that the general factor tends to be dominant relative to the specific factors, a single, composite or summed score is appropriate for evaluating each teacher candidate. The score received by an individual student teacher can be interpreted as a measure of his or her teaching effectiveness; the higher the score, the more effective the student teacher is in delivering high-quality instruction.

Measurement invariance tests demonstrate that the observed items measure the same theoretical constructs in both cooperating teacher and university supervisor groups. Obtaining measurement invariance between raters provides evidence of reliability of scores between raters and allows researchers to compare raters' scores (Horn & McArdle, 1992). For example, because measurement invariance was obtained between raters in the Utah State University sample, the difference in means between raters is not attributed to the raters interpreting the items in the scale differently but to observed differences in performance of teachers evaluated by each rater. These results may suggest that more training is needed among cooperating teachers and university supervisors at USU.

While model-fit statistics indicate that the bifactor model fit these data better than the other models, results suggest some estimation problems were encountered with these data. One advantage of the bifactor model is that it can be used to evaluate the importance of domain-specific factors. It is possible that a domain-specific factor will not be relevant to the prediction of the observed measures when the general factor is included in the model. Once the general factor is partialed out, the domain-specific factor(s) do not account for unique variance in the indicators. While the parameter estimates for the general factor in the BYU and USU samples are reasonable, some factor loadings and factor variances are small and statistically nonsignificant (see Tables 5 and 6). Further work is needed to examine the validity of these dimensions and to determine if these subdomains contribute in the expected manner to measuring teacher effectiveness in addition to the general factor.

Using valid and reliable teacher performance assessments to provide feedback to candidates and programs is essential to improving teacher preparation (Darling-Hammond, 2014). The results of this study can be used to improve the student teaching and learning experiences of teacher candidates within the consortium and provide a common reference for educator collaboration among programs. This study also adds to the growing research on measuring teacher effectiveness and how these measures can be used to improve clinical preparation and teacher education.

Table 5. *Standardized factor loadings for the bifactor model of Brigham Young University*

		Domain-specific factors		
Item	General factor	Learner and Learning	Instructional practices	Professionalism
1.1 - Creates developmentally appropriate and challenging learning experiences based on each learner's strengths, interests, and needs.	0.784	-0.095		
1.2 - Collaborates with families, colleagues, and other professionals to promote student growth and development	0.592	0.115		
2.1 - Allows learners multiple ways to demonstrate learning sensitive to diverse experiences, while holding high expectations for all.	0.736	-0.167		
3.1 - Develops learning experiences that engage and support students as self-directed learners who internalize classroom routines, expectations, and procedures.	0.752	0.313		
3.2 - Collaborates with students to establish a positive learning climate of openness, respectful interactions, support, and inquiry.	0.731	0.398		
3.3 - Utilizes positive classroom management strategies, including the resources of time, space, and attention effectively.	0.626	0.475		
4.1 - Bases instruction on accurate content knowledge using multiple representations of concepts.	0.646		0.156	
4.2 - Supports students in learning and using academic language accurately and meaningfully.	0.637		0.270	
5.1 - Uses data sources to assess the effectiveness of instruction and to make adjustments in planning and instruction.	0.638		-0.450	
5.2 - Engages students in understanding and identifying the elements of quality work.	0.723		-0.029	

5.3 - Documents student progress and provides descriptive feedback to student, parent, and other stakeholders in a variety of ways.	0.636		-0.347	
6.1 - Demonstrates knowledge of the Utah Core Standards and references it in short- and long-term planning.	0.580		-0.095	
6.2 - Integrates cross-disciplinary skills into instruction to purposefully engage learners in applying content knowledge.	0.593		0.303	
7.1 - Practices a range of developmentally, culturally, and linguistically appropriate instructional strategies to meet the needs of individuals and groups of learners.	0.703		-0.001	
7.2 - Provides multiple opportunities for students to develop higher-order and meta-cognitive skills.	0.649		0.360	
7.3 - Supports and expands learner's communication skills through reading, writing, listening, and speaking.	0.710		0.160	
7.4 - Uses a variety of available and appropriate technology and resources to support learning.	0.574		0.174	
7.5 - Develops learners' abilities to find and use information to solve real-world problems.	0.650		0.437	
7.6 - Uses a variety of strategies, including questioning, to promote engagement and learning.	0.785		0.178	
8.1 - Adapts and improves practice based on reflection and new learning.	0.762			0.147
9.1 - Participates actively in decision-making processes, while building a shared culture that affects the school and larger educational community.	0.519			1.243
9.2 - Advocates for the learners, the school, the community, and the profession.	0.606			0.312

Table 6. *Standardized factor loadings for the bifactor model of the Utah State University sample*

		Domain-specific factors		
Item	General factor	Learner and Learning	Instructional practices	Professionalism
1.1 - Creates developmentally appropriate and challenging learning experiences based on each learner's strengths, interests, and needs.	0.873	0.112		
1.2 - Collaborates with families, colleagues, and other professionals to promote student growth and development	0.769	-0.020		
2.1 - Allows learners multiple ways to demonstrate learning sensitive to diverse experiences, while holding high expectations for all.	0.796	0.104		
3.1 - Develops learning experiences that engage and support students as self-directed learners who internalize classroom routines, expectations, and procedures.	0.783	0.331		

3.2 - Collaborates with students to establish a positive learning climate of openness, respectful interactions, support, and inquiry.	0.834	0.178		
3.3 - Utilizes positive classroom management strategies, including the resources of time, space, and attention effectively.	0.720	0.583		
4.1 - Bases instruction on accurate content knowledge using multiple representations of concepts.	0.834		-0.112	
5.1 - Uses data sources to assess the effectiveness of instruction and to make adjustments in planning and instruction.	0.776		0.164	
5.2 - Engages students in understanding and identifying the elements of quality work.	0.776		0.131	
5.3 - Documents student progress and provides descriptive feedback to student, parent, and other stakeholders in a variety of ways.	0.824		0.170	
6.1 - Demonstrates knowledge of the Core Standards and references it in short- and long-term planning.	0.766		-0.065	
6.2 - Integrates cross-disciplinary skills into instruction to purposefully engage learners in applying content knowledge.	0.806		0.114	
7.1 - Practices a range of developmentally, culturally, and linguistically appropriate instructional strategies to meet the needs of individuals and groups of learners.	0.829		-0.004	
7.2 - Provides multiple opportunities for students to develop higher-order and meta-cognitive skills.	0.824		-0.223	
7.3 - Supports and expands learner's communication skills through reading, writing, listening, and speaking.	0.802		0.148	
7.4 - Uses a variety of available and appropriate technology and resources to support learning.	0.703		0.536	
7.5 - Develops learners' abilities to find and use information to solve real-world problems.	0.813		0.116	
8.1 - Adapts and improves practice based on reflection and new learning.	0.815			0.071
9.1 - Participates actively in decision-making processes, while building a shared culture that affects the school and larger educational community.	0.835			0.467
9.2 - Advocates for the learners, the school, the community, and the profession.	0.889			0.177

References

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on the teaching and learning of mathematics* (pp. 83–104). Westport, CT: Ablex.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80,* 796-846.

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. (2014). Strengthening clinical preparation: The holy grail of teacher education. *Peabody Journal of Education, 89*, 547-561.

Gignac, G.E. (2015). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment, 30*(2), 130-139.

Grossman, P. L., & Stodolsky, S. S. (1995). Content as context: The role of school subjects in secondary school teaching. *Educational Researcher, 24*(8), 5–11, 23.

Hammerness, K., Darling-Hammond, L., & Shulman, L. (2002, August). Toward expert thinking: How case-writing contributes to the development of theory-based professional knowledge in student-teachers. *Teaching Education, 13*, 221–245.

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

Kline, R.B. (2005). *Principles and Practices of Structural Equation Modeling*, Guilford Press: New York.

McDonald, R.P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum.

McDonald, R.P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum.

Muthen, L.K. & Muthen, B.O. (2017). *Mplus User's Guide.* Los Angeles: CA: Muthen & Muthen.

Reise, S.P. (2012). The rediscovery of the bifactor measurement models. *Multivariate Behavioral Research*, *47,* 667-696.