VALIDATION STUDY FOR THE STUDENT TEACHER OBSERVATION TOOL

ND COMMON METRICS PROJECT

May 2017

Executive Summary

As part of the ND Common Metric Project, representatives from the twelve constituent institutions of the North Dakota Association of Colleges for Teacher Education (NDACTE) recently developed the Student Teacher Observation Tool (STOT), a new instrument for assessing the performance of student teachers during the clinical experience. Pilot data were collected during the spring 2016 semester in order to conduct an exploratory factor analysis (EFA) to gauge the psychometric performance of this new instrument. Results of the EFA indicated the instrument contained measurement for two student performance constructs: the learner and learning, and professionalism. The instrument was revised using the results of the EFA by modifying instrument questions and survey design. The instrument was then administered state-wide to assess student teachers during fall semester. This report provides the results of this second validation study that was conducted for the second phase of validation, confirmatory factor analysis (CFA) and the subsequent recommendations for further instrument revision, which are to serve as a guide for improvements and further development of the instrument (i.e., "fine tuning"). The results of the CFA indicated one factor, so a subsequent EFA was run. The following is a very brief summary of the results and recommended actions and considerations needed to further develop and strengthen the instrument.

Results:

1. The instrument is able to differentiate the four hypothesized constructs. Though there are still factors that cross-load, the structure for the four factors is more distinct than the pilot EFA conducted prior to instrument-revision.
2. Though initially intended to be run as a CFA, there was not enough structure in the CFA and thus an additional EFA was conducted. Revisions to the survey instrument further warranted an additional EFA to identify areas of improvement from the first run.

Recommendations:

1. Continue to refine instrument items to further strengthen structure of the four hypothesized constructs. It is recommended that items that load onto the unintended factors be re-evaluated.
2. An additional study could be done to compute the reliabilities for the ten standards as components or formative scales.
3. Further revision of the survey design by adding a "Not Applicable" option to scale items. This was commonly requested in the open-ended comments. Also as previously recommended, utilize the validation tools within Qualtrics to reduce missing item non-response and garner a better sample size.
4. Continue to refine instrument items, distribute to cooperating teachers, and conduct future confirmatory factor analysis.

About the Instrument

Similar information as described with the initial EFA in Fall 2016's first report but note that there are now 34 items and many revisions to items and rubric descriptions. Table 1 displays the item numbers and standards along with their intended construct.

Table 1

Constructs, InTASC Standards, and Intended Alignment of Items

Construct/Areas of Knowledge	Code	InTASC Standard	Item #
The Learner and Learning	L	#1: Learner Development	1-2
		#2: Learning Differences	3-4
		#3: Learning Environments	5-9
Content Knowledge	C	#4: Content Knowledge	10-12
		#5: Application of Content Knowledge	13-16
Instructional Practice	I	#6: Assessment	17-20
		#7: Planning for Instruction	21-24
		#8: Instructional Strategies	25-28
Professional Responsibility	P	#9: Professional Learning and Ethical Practice	29-32
		#10: Leadership and Collaboration	33-34

It was noted that compared to the initial instrument, one item was removed from standard one as well as standard two. An additional item was introduced in standard number three. All other standards contained the same amount of items in the original instrument. The initial EFA suggested further item refinement, and instrument developers made adjustments accordingly. Many initially "double-barreled" items have now been simplified to measure one thing, and many rubric items have been pared down for ease of the responder and simplified with adding in descriptive words such as "inappropriate" earlier in the scale descriptor rather than a simple "not" later in the description that could be missed by the respondent.

Results

Preliminary Data Screening

The initial sample size was $n = 254$ records, but only $n = 171$ of those were usable response sets (i.e., responses were given on all 34 rating items). There were valid ratings for 155 student teachers from seven different institutions in North Dakota. Further cleaning of the data resulted in 139 usable cases. The non-response items were analyzed, and most were duplicates of a responder that began the survey and then also had a completed survey. There were two cases that were dropped due to detailed comments that the cooperating teacher desired a "Non-Applicable" option resulting in what was likely not valid or accurate use of the current rubric for the standards. Responders who indicated special education as their content area were dropped as well as responses from University supervisors.

Characteristics of the Sample

Descriptive statistics for a few important characteristics of the sample are reviewed here to confirm that it is representative of the general population as well as to check for any unusual events.

Table 2

Summary of Missing Data

Missing	Frequency	Percent	Cumulative
0	139	89.68	89.68
1	9	5.81	95.48
2	2	1.29	96.77
3	3	1.94	98.71
4	2	1.29	100.00
Total	155		100.00

Grade levels, subject areas, and student attending universities. Table 3 shows the frequencies of the grade-levels reported for the $n = 139$ valid respondents. The reported subject areas for those who had a middle- or high-school experience is provided in Table 4. Note that science was again omitted from the list of subject areas in the instrument, but respondents used the "other" option to report when science was the appropriate subject area. Table 5 shows the attending university that the student teachers attend.

Table 3

Frequencies for the Reported Grade Levels of the Student Teaching Experience

Level(s)	Freq.	Percent
Elementary	92	
Middle school	7	
High school	25	
Elementary and middle school	1	
Elementary and high school	1	
Middle and high school	6	
Elementary, middle, and high school	3	
No response	4	
Total	139	100.00

Table 4

Reported Subject Areas for Student Teachers with Middle and High School Placements

Subject Area	Freq.
Ag	1
Art	1
Business	1
English	3
Family and consumer science	1
Health	1
History	10
Math	5
Physical education	6
Science	6
Spanish	1
Special Education	3
Other	3

Table 5

Reported Attending Institution of Student Teacher

Institution	Freq.	Percent	Cum.
Mayville State University	12	8.63	8.63
Minot State University	43	30.94	39.57
North Dakota State University	27	19.42	58.99
University of Jamestown	1	0.72	59.71
University of Mary	2	1.44	61.15
University of North Dakota	35	25.18	86.33
Valley City State University	19	13.67	100.00
Total	139		100.00

Instrument Validity

Construct validation of the Student Teacher Observation Tool (STOT) was implemented via an exploratory factor analysis (EFA) using pilot data collected from a sample of n = 139 respondents that completed all 34 assessment items. These 34 rating items were used as the observed variables in the EFA.

Number of factors. Although there were four hypothesized factors, it is still necessary to confirm the number of factors based on empirical data. First, the KMO (a general measure of factorability) was .960; being greater than the recommended threshold of .6 indicates the presence of a factor structure, but it does not reveal how many factors. As is generally recommended, a few different number-of-factors (dimensionality) tests were conducted. As shown in Table 6, there was no clear consensus among the different dimensionality test for the proper number of factors to extract.

Table 6

Results from the Various Number-of-Factors Tests

Test	Number of Factors Indicated
Parallel analysis (Horn, 1965)	4
Minimum average partial correlation (MAP) test (Velicer, 1976)	3
Scree test (Cattell, 1966)	1
Kaiser rule (Kaiser, 1960)	3
Interpretability of factors	4

It should be noted that it is generally recommended in the methodological literature that parallel analysis and the MAP test are given primacy. However, with the inconclusive results in this instance, the interpretability of the factors played a key role in determining the number of factors to extract. Accordingly, different factor solutions with one to four factors were computed and examined separately. The four-factor solution emerged as the most viable and substantively meaningful solution.

Factor extraction and rotation. Four common (principal axes) factors were extracted and rotated to an oblique solution (i.e., factors were allowed to be correlated) using the oblimin rotation criterion. There were originally four hypothesized factors, and all four factors did emerge. The meanings of these four factors were determined through examination of the factor loadings on each of the items (Table 6). The first factor represents the construct *instructional practice* (I), the second factor represents the construct *content knowledge* (C), the third factor *professional responsibility* (P) and the fourth factor represents *learner and learning* (L). Table 7 displays the factor loadings.

Table 7

Factor Loadings

Construct	Standard #	Item #	Factor 1	Factor 2	Factor 3	Factor 4
Learner, learning, and diversity	1	1		.5064		
		2		.3910		
	2	1	.4431			**.4693**
		2			.4419	
	3	1				
		2				**.5517**
		3				**.7812**
		4				**.7640**
Content knowledge	4	1		**.5854**		
		2				.4414
		3	.5239			
	5	1		**.5311**		
		2	.3919	**.4439**		
		3	.4474	**.3814**		
		4	.4365			
Instructional practices	6	1				
		2	**.5341**			
		3	**.7188**			
		4	**.7039**			
	7	1		.4899		
		2	**.5494**			
		3	**.4690**			
		4			.5752	
	8	1	**.3640**			
		2		.6170		
		3	**.4099**			
		4		.3580		
Professionalism	9	1			**.7025**	
		2			**.8554**	
		3			**.4292**	
		4		.4002	**.4772**	
	10	1			**.6649**	
		2	.3649	.5088		

Non-salient loadings (< .35) appears as blanks.

Loadings (also known as *pattern coefficients*) are essentially standardized regression weights for each item with the factors as predictors (i.e., the underlying factors are used to reproduce the observed item rating scores). Thus, loadings reflect the strength of association for a factor and an item. Only salient loadings (coefficients greater than .35 in absolute value) are shown; blank cells in the table represent non-salient loadings.

A *communality* is the squared multiple correlation for an item being predicted by the factors. So, this quantity represents the proportion of variance in an item that can be accounted for by the factors. The communalities from this factor solution are quite good as all are at least moderate in magnitude (≥ .4); in fact, most are high (≥ .7). This reaffirms that the four-factor solution is indeed adequate since the factors account for a majority of the variance in all items. Table 8 display communalities by construct and Table 9 displays all communalities.

Table 8

Summary of Item Communalities by Construct

Construct	Number of Items	Mean	Min	Max
Learner, learning, and diversity	8	.665	.541	.777
Content knowledge	7	.670	.607	.730
Instructional practices	12	.653	.504	.731
Professionalism	6	.651	.548	.785

The communality for an item represents the proportion of variance accounted for by the factors. All items have very high communalities (> .500).

Table 9

All Communalities

Construct	Standard #	Item #	Communality
Learner, learning, and diversity	1	1	.6821
		2	.6400
	2	1	.6982
		2	.6129
	3	1	.5409
		2	.6726
		3	.7774
		4	.6941
Content knowledge	4	1	.6069
		2	.7036
		3	.6494
	5	1	.6915

		2	.7298
		3	.6146
		4	.6931
Instructional practices	6	1	.5468
		2	.6263
		3	.6825
		4	.7205
	7	1	.6944
		2	.7028
		3	.7100
		4	.6941
	8	1	.7008
		2	.5043
		3	.7311
		4	.5211
Professionalism	9	1	.6835
		2	.7846
		3	.5480
		4	.5557
	10	1	.6537
		2	.6805

Factor correlation. As previously mentioned, this is an oblique factor solution, meaning that the factors were allowed to be correlated. The correlation matrix can be seen in Table 10. The highest rotated factors (Factor1 and Factor 2) had a Pearson correlation of .687, which is fairly strong, coming under the recommendation against factors being correlated above .7, advising that such strongly correlated factors should be merged into a single factor. Regardless of these general guidelines, four factors were retained for two reasons: (1) the four-factor solution provided important information regarding the potential factor structure that differentiated the hypothesized constructs; and (2) a relatively smaller sample sizes can result in upwardly biased correlation estimates.

Table 10

Correlation matrix of the oblimin rotated common factors

Factors	Factor1	Factor2	Factor3	Factor4
Factor1	1			
Factor2	.6873	1		
Factor3	.6079	.6002	1	
Factor4	.595	.5625	.5555	1

Instrument Reliability

Reliability analysis typically follows validity analysis (EFA). This generally consists of computing Cronbach's alpha for each of the subscales corresponding to the factors that have been validated. Reliability analysis for the third and fourth factor (Factor3 and Factor4) were excluded because their current state of cross-loading and errant-loading items. In this study, only the L and P constructs (Factor1 and Factor2 in Table 11) show the adequate measurement. Five of the items designed to tap the L construct exhibited salient loadings on only that factor. Hence the items l_s2_1, l_s3_1, l_s3_2, l_s3_3, and l_s3_4 comprise the L subscale of the instrument, which shows very good reliability with a Cronbach's alpha of .930. Four of the items designed to tap the P construct exhibited salient loadings on only that factor. Hence the items p_s9_1, p_s9_2, p_s9_3, and p_s10_1 comprise the P subscale of the instrument, which shows relatively high internal consistency with a Cronbach's alpha of .902.

Table 11

Reliabilities of Subscales

Subscale/Construct	Number of Items	Cronbach's Alpha
Learner, learning, and diversity	8	.930
Content knowledge	7	.929
Instructional practices	12	.952
Professionalism	6	.902

All subscales have excellent reliability.

Recommendations

The recommendations for continued improvement of the STOT instrument are to continue item refinement, consider computing the reliabilities for the different standards as components or formative scales, and making additional changes to the survey design.

Item Refinement

The current EFA shows additional areas of errant and cross-loading items. It is recommended that these items continue to be refined by clarifying language and consideration of the standard and the intended construct.

Survey Form

The current dataset contained a large quantity of missing data. Future use of the STOT instrument would benefit from utilizing the validation instrument within the Qualtrics software to require the completion of each scale item. The addition of a "Not Applicable" option would also likely increase completion of each item as well as reduce the likelihood of an invalid score on the instrument to simply complete the instrument.