VALIDATION STUDY FOR THE STUDENT TEACHER OBSERVATION TOOL
ND COMMON METRICS PROJECT

August 2016

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. This report provides the results of this validation study and the subsequent recommendations for instrument revision, which are to serve as a guide for improvements and further development of the instrument (i.e., "fine tuning"). After revision, the instrument can then be used statewide by all twelve of the NDACTE institutions during the fall 2016 semester for the purposes of collecting data for the second phase of validation—specifically, confirmatory factor analysis (CFA).

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 *professional responsibility* area of knowledge (construct) from the others (*the learner and learning*, *content knowledge*, and *instructional practice*); however, it needs further development and fine-tuning to differentiate those three from each other.
2. The *professional responsibility* subscale shows very good reliability.
3. Caveat: The results of the EFA may have limited stability (replicability) due to the relatively small sample of usable responses (n = 80).

Recommendations:

1. A number of items are candidates for revision or deletion.
2. The response scale rubrics may be in need of some clarification and fine-tuning.
3. The online data collection form (in Qualtrics) is in need of a few additional modifications.
4. After revision of the instrument, continue with the second phase of validation (i.e., full-scale distribution and confirmatory factor analysis).

About the Instrument

The STOT pilot form consisted of 35 rating items in rubric format and was administered online (Qualtrics). A cooperating teacher would use this instrument to rate the performance of a student teacher during the clinical experience. The rubrics were designed to provide a set of rating scores (1 to 4 by increments of .5) with detailed descriptions, which would allow for a rater to choose which level best fit the performance of the student teacher. Stevens and Levi (2013) describe this type of rubric design with detailed choices as a way to convey effective feedback in a manner that shows the specific expectations for each level of performance.

Each item in the STOT was designed to correspond to one of the ten InTASC standards, each of which belonging to one of four fundamental areas of knowledge (more generally known as *latent constructs*). Accordingly, each item in the instrument was initially designed to tap only one standard and

subsequently, only one construct. Table 1 shows the four hypothesized constructs, the corresponding letter-code abbreviations used in this report, and the construct-standard alignment.

It should be noted that the focus of this validation study was on these four hypothesized factors rather than the ten standards because the standards do not necessarily represent truly autonomous constructs; rather, each standard represents a specific feature (or *facet*) of a construct. Even if the ten standards were stand-alone constructs, it would require at least five items per standard (thus a total of 50 or more items) to obtain a reliable measurement. Such a lengthy instrument would likely not be practicable.

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	1-3
		#2: Learning Differences	4-6
		#3: Learning Environments	7-10
Content Knowledge	C	#4: Content Knowledge	11-13
		#5: Application of Content Knowledge	14-17
Instructional Practice	I	#6: Assessment	18-21
		#7: Planning for Instruction	22-25
		#8: Instructional Strategies	26-29
Professional Responsibility	P	#9: Professional Learning and Ethical Practice	30-33
		#10: Leadership and Collaboration	34-35

Note that in this paper, items have been named using the following naming convention which references three pieces of information—namely, the construct code letter, the standard number, and the item number. So, for instance, P-S9-3 refers to the third item (3) that was originally intended to reflect standard 9 (S9) as part of the construct *professional responsibility* (P). These item identifiers are shown with each item stem in the Appendix.

Results

Preliminary Data Screening

The initial sample size was $n = 133$, but only $n = 80$ (60.2%) of those were complete and usable response sets (i.e., responses were given on all 35 rating items). This means that 50 cases were missing responses on at least one of the 35 rating items. Table 2 shows the breakdown of frequency counts for the number of items without a response. There were 80 cases with no missing response data, thus the usable sample size of $n = 80$. It would appear from Table 2 that only a few inadvertently failed to respond to a small number of items, but many stopped well before being finished.

Table 2

Frequencies for Total Item Nonresponse

Missing Responses	Freq.	Percent	Cum. Pct.
0	80	60.15	60.15
1	2	1.50	61.65
3	1	0.75	62.41
18	1	0.75	63.16
22	1	0.75	63.91
24	1	0.75	64.66
29	1	0.75	65.41
32	1	0.75	66.17
34	2	1.50	67.67
35	43	32.33	100.00
Total	133	100.00	

It is important to note here that the methodological literature (see Gorsuch, 1983) generally recommends a minimum of at least three (some say five) respondents per item in order to achieve stable (replicable) factor analysis results. With a final sample size of $n = 80$, this ratio is just over two cases per item in this study. Therefore, the replicability of these results may be limited due to the low respondent-to-item ratio.

Characteristics of the Sample

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

Grade levels and subject areas. Table 3 shows the frequencies of the grade-levels reported for the $n = 80$ 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 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 3

Frequencies for the Reported Grade Levels of the Student Teaching Experience

Level(s)	Freq.	Percent
Elementary	42	52.50
Middle school	15	18.75
High school	16	20.00
Elementary and high school	1	1.25
Middle and high school	3	3.75
Elementary, middle, and high school	1	1.25
No response	2	2.50

Table 4

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

Subject Area	Freq.
Art	2
English	5
Family and consumer science	1
German	1
Health	2
History	12
Math	3
Physical education	8
Science	5
No response	2

Completion times. The time taken for each respondent to complete the instrument was also recorded. The completion times were strongly positively skewed for the $n = 80$ valid cases, which is typical especially when respondents are allowed to stop and continue at a later time (which is a feature in Qualtrics). Basic descriptive statistics for completion times (reported in minutes) are given in Table 5. Notice that while the mean is fairly large (104.8 minutes), this is due to the positive skew of the observed distribution. The median (10.9 minutes) is a more robust indicator of central tendency for these timespan data. Additionally, the frequencies for completion times (in 10-minute increments) are given in Table 6, which shows that many completed the instrument in less than 10 minutes. Although not apparent in Table 6, it may be worth noting that nine respondents finished in less than five minutes. Such cases with "rushed" completion times are typically candidates for removal because the veracity of such responses are dubious; however, this is simply not an option with this dataset given the relatively small sample size.

Table 5

Descriptive Statistics for Instrument Completion Time (Minutes)

Mean	Minimum	First Quartile	Median	Third Quartile	Maximum
104.8	3.4	6.7	10.9	19.1	5810.3

Table 6

Frequencies of Instrument Completion Times in 10-Minute Increments

Interval (Minutes)	Freq.	Percent	Cum. Pct.
$T < 10$	36	45.00	45.00
$10 \leq T < 20$	25	31.25	76.25
$20 \leq T < 30$	4	5.00	81.25
$30 \leq T < 40$	6	7.50	88.75
$40 \leq T < 50$	3	3.75	92.50
$50 \leq T < 60$	2	2.50	95.00
$T \geq 60$	4	5.00	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 $n = 80$ respondents that completed all 35 assessment items. These 35 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 .940; 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 7, there was no clear consensus among the different dimensionality test for the proper number of factors to extract.

Table 7

Results from the Various Number-of-Factors Tests

Test	Number of Factors Indicated
Parallel analysis (Horn, 1965)	1
Minimum average partial correlation (MAP) test (Velicer, 1976)	3
Scree test (Cattell, 1966)	2
Kaiser rule (Kaiser, 1960)	3
Sequential KMO (Hill, 2011)	1
Interpretability of factors	2

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 two-factor solution emerged as the most viable and substantively meaningful solution. Further, the three- and four-factor solutions were shown to be "Heywood" (pathological) cases, meaning they returned nonviable estimates for one or more parameters (likely due to the relatively small sample size).

Factor extraction and rotation. Two 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. Although there were originally four hypothesized factors, only two factors emerged. The meanings of these two factors were determined through examination of the factor loadings on each of the items (Table 8). The first factor represents an amalgamation of the constructs *learner and learning* (L), *content knowledge* (C), and *instructional practice* (I), while the second factor represents the construct *professional responsibility* (P).

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 .3 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 ($\geq .4$); in fact, most are high ($\geq .7$). This reaffirms that the two-factor solution is indeed adequate since the factors account for a majority of the variance in all items.

Table 8

Rotated Pattern (Loading) Matrix with Communalities from the Two-Factor Solution

Item	Loadings		Communality
	Factor 1	Factor 2	
L-S1-1	.8010		.7919
L-S1-2	.8615		.7818
L-S1-3	.9065		.6954
L-S2-1	.8307		.7228
L-S2-2	.6692		.7129
L-S2-3	.6153		.6412
L-S3-1	.6935		.6133
L-S3-2*	.3875	.5387	.7582
L-S3-3	.5608		.6500
L-S3-4	.9056		.5852
C-S4-1	.8158		.6105
C-S4-2	.8487		.7680
C-S4-3	.8913		.7091
C-S5-1	.8273		.7502
C-S5-2	.8120		.6706
C-S5-3	.8150		.5902
C-S5-4	.8952		.7739
I-S6-1	.7446		.7572
I-S6-2	.6216		.7067
I-S6-3*	.5665	.3264	.7090
I-S6-4	.7851		.7056
I-S7-1	.6730		.6316
I-S7-2	.7775		.7500
I-S7-3	.7521		.7659
I-S7-4**		.6042	.7277
I-S8-1	.8030		.7198
I-S8-2	.8984		.5989
I-S8-3	1.0107		.8565
I-S8-4	.7689		.7579
P-S9-1		.8799	.8572
P-S9-2		.8259	.8670
P-S9-3		.8992	.6568
P-S9-4		.7138	.7160
P-S10-1		.6423	.7214
P-S10-2**	.5481		.6122

* Cross-loading items

** Errant-loading items

Factor correlation. As previously mentioned, this is an oblique factor solution, meaning that the factors were allowed to be correlated. The two rotated factors had a Pearson correlation of .761, which is fairly strong. In fact, some sources recommend against factors being correlated above .7, advising that such strongly correlated factors should be merged into a single factor. Regardless of these general guidelines, two factors were retained for two reasons: (1) the two-factor solution provided important information regarding the potential factor structure that differentiated the P construct from L, C, and I; and (2) smaller sample sizes can result in upwardly biased correlation estimates.

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 first factor (Factor 1 in Table 8) was excluded because in its current state, that factor represents an undifferentiated composite of three hypothesized constructs (L, C, and I). In this study, only the P construct (Factor 2 in Table 8) appears to be adequately measured. Five 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, P-S9-4, and P-S10-1 comprise the P subscale of the instrument, which shows very good reliability with a Cronbach's alpha of .938.

Recommendations

The following recommendations broadly deal with two areas: psychometric issues and form design issues. The psychometric issues focus on the revision of the wording in certain components of the instrument in order to achieve better measurement of and differentiation among the intended constructs. The form design issues deal with the general content, layout, and functionality of the online data collection form.

Psychometric Issues

Revision of item stems. Fowler (2014) described the careful crafting of questions as being paramount to the construction of a valid and reliable instrument that is accurately understood and used by respondents. Accordingly, the first and most important recommendation is for the revision of the wording of a few particular questions so that they more clearly convey the concepts that the raters should be assessing in those items. The specific items in need of revision are discussed below.

Cross- and errant-loading items. After an examination of the rotated loading matrix (Table 8), there were two items that crossed-loaded on both factors (L-S3-2 and I-S6-3), and two that loaded on a factor for which they were not designed to measure (I-S7-4 and P-S10-2). These four items should be scrutinized and revised to better reflect their intended constructs. Alternatively, these item can simply be dropped from the instrument.

Double-barreled items. A double-barreled item is a question that is phrased in such a way that it is essentially inquiring about multiple distinct issues or characteristics simultaneously. Such ambiguously constructed items are a common cause of inaccurate measurements. Consider, for example, item L-S1-3: "Sequences lessons to ensure coherence with curriculum and account for student's prior knowledge." This item is clearly asking about two separate things: (1) coherence with curriculum, and (2) accounting for prior knowledge. A rater is limited to only one response for a question, so an ambiguous item such as this leads to ambiguous data. For example, a rater may observe one aspect addressed in this question as "Distinguished" according to the rubric, yet the other aspect in this same question may be judged as only "Emerging." Additional inconsistency arises when different raters base their responses on completely different parts of the question. Table 9 contains a complete list of all items in the STOT that appear to be double-barreled.

Table 9

List of Potentially Double-Barreled Items

Item	Stem
L-S1-2	Implements developmentally appropriate instructional strategies and practices to support student learning
L-S1-3	Sequences lessons to ensure coherence with curriculum and account for students' prior knowledge
L-S3-2	Develops and maintains a classroom environment that promotes student engagement
L-S3-4	Uses technologies to enhance learning and guide learners to apply them in appropriate, safe, and effective ways
C-S5-3	Knows where and how to access resources, including technologies, to build global awareness and understanding
C-S5-4	Engages learners in critical/creative thinking, and collaborative problem solving experiences
I-S6-3	Uses multiple and appropriate data sources to identify student learning needs
I-S7-4	Plans and works collaboratively with other teachers and/or specialists to design instruction that supports individual student learning
I-S8-4	Uses effective communication skills and strategies to convey ideas and information to students
P-S9-1	Seeks and accepts feedback to improve teaching effectiveness
P-S10-2	Works effectively with parents, families, and the community

Refine the rating scale rubrics. The next psychometric-related recommendation is to review and the fine-tune the rubrics attached to each of the rating items, particularly for the "Proficient" and "Distinguished" rating levels. Since the rating levels are translated to a numerical scale with equal increments (1 to 4 by increments of .5), the rubrics should describe approximately equivalent-spaced gradations of developmental sophistication (semantically and pragmatically speaking). That is, the set of criteria should "feel" like they reflect roughly equivalent developmental steps. Such semantic-numeric inconsistencies can adversely impact validity and reliability because such irregular spacing may impose an unnatural restriction of range on the distribution of rating scores for an item. Stevens and Levi (2013) provide a more in-depth and detailed set of guidelines for the creation and refinement of rubrics.

As an example of this issue, consider the rubric for item L-S1-3 shown in Table 10. The stem of this item reads, "Sequences lessons to ensure coherence with curriculum and account for students' prior knowledge." In this rubric, the levels "Undeveloped," "Emerging," and "Proficient" seem to make even, gradual steps towards increased proficiency. However, there appears to be a disproportionately larger substantive leap from "Proficient" to "Distinguished." There are many additional and new skills that need to appear here that must be distinguished and were not accounted for as developing throughout. In this case each level is defined differently which can add confusion and inconsistent scoring.

Table 10

Rubric for Item L-S1-3

Rating Level	Description
Undeveloped	Lessons are not sequenced to align with standards and do not account for students' prior knowledge
Emerging	Sequences lessons that address students' prior knowledge as a class, but does not consider individual development differences
Proficient	Sequences lessons that consider students' prior knowledge and leads students toward mastery of standards in a coherent manner
Distinguished	Sequences lessons and practice toward mastery of standards for all students in a coherent manner. Lessons access and expand on students' prior knowledge and build on each lesson in preparation for future learning

Form Design Issues

Online form layout. If feasible, consider redesigning the layout of the Qualtrics survey form so that respondents can view an item with the specific related standard simultaneously. Many of the instrument items and rubric definitions were very similar (differing by only a few words), so being able to see all relevant information would help mitigate any potential confusion for respondents.

Use the item nonresponse alert functions in Qualtrics. It is not uncommon for a respondent to overlook a single item. It is also possible for entire blocks of items to be overlooked when using online forms with multiple pages (such as the STOT). To help avoid item nonresponse, use the error-checking feature in Qualtrics. This will issue an alert should a respondent attempt to submit a form with missing input.

Include all of the most common subject areas. The list of subject areas should be reviewed to ensure that all of the common subject areas are given. Of note, science was omitted from this list. Of course, respondents can use the "other" option to report when science was the appropriate topic. However, consider one of the more colorful "other" responses: "Science- seriously- no science on the list?!". Seemingly trivial mistakes such as this can compromise face validity from the perspective of the respondent, which may inadvertently lead to a less-than-serious attitude when completing the rating form.

References

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245-276.

Fowler, F. J. (2014). *Survey research methods* (5th ed.). Thousand Oaks, California: Sage Publications, Inc.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hill, B. D. (2011). *The sequential Kaiser-Meyer-Olkin procedure as an alternative for determining the number of factors in common-factor analysis: A Monte Carlo simulation* (Doctoral dissertation). Oklahoma State University, Stillwater, OK.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141-151.

Stevens, D. D., & Levi, A. J. (2013). *Introduction to rubrics* (2nd ed.). Sterling, Virginia: Stylus Publishing.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327.

Appendix

This appendix contains a listing of all 35 item stems used in the instrument (Table A1). Also shown are the item identifier codes used within this report. As previously mentioned, items have been assigned identifiers using the construct code letter, the standard number, and the item number. So, for instance, P-S9-3 refers to the third item (3) that was originally designed to reflect standard 9 (S9) as part of the construct *professional responsibility* (P).

Table A1

Item Identifiers and Stems from the Instrument

Order	Item Identifier	Stem
1	L-S1-1	Designs developmentally appropriate instruction to support student learning
2	L-S1-2	Implements developmentally appropriate instructional strategies and practices to support student learning
3	L-S1-3	Sequences lessons to ensure coherence with curriculum and account for students' prior knowledge
4	L-S2-1	Effectively teaches students from various socioeconomic backgrounds, culturally and ethnically diverse backgrounds and communities
5	L-S2-2	Plans differentiated instruction for a variety of learning needs
6	L-S2-3	Exhibits fairness and belief that all students can learn
7	L-S3-1	Fosters a safe and respectful environment that promotes learning
8	L-S3-2	Develops and maintains a classroom environment that promotes student engagement
9	L-S3-3	Clearly communicates expectations for appropriate student behavior
10	L-S3-4	Uses technologies to enhance learning and guide learners to apply them in appropriate, safe, and effective ways
11	C-S4-1	Effectively teaches subject matter
12	C-S4-2	Creates meaningful learning experiences to assure mastery of content
13	C-S4-3	Integrates culturally relevant content to build on learners' background knowledge
14	C-S5-1	Designs instruction and learning tasks that connect core content to relevant, real-life experiences for students
15	C-S5-2	Designs activities where students engage with subject matter from a variety of perspectives
16	C-S5-3	Knows where and how to access resources, including technologies, to build global awareness and understanding
17	C-S5-4	Engages learners in critical/creative thinking, and collaborative problem solving experiences
18	I-S6-1	Designs and modifies formative and summative assessments to match learning targets

19	I-S6-2	Engages learners in critical /creative thinking, and collaborative problem solving experiences
20	I-S6-3	Uses multiple and appropriate data sources to identify student learning needs
21	I-S6-4	Engages students in self-assessment strategies
22	I-S7-1	Connects lesson goals with school curriculum and state standards
23	I-S7-2	Uses assessment data to inform planning for instruction
24	I-S7-3	Adjusts instructional plans to meet students' needs
25	I-S7-4	Plans and works collaboratively with other teachers and/or specialists to design instruction that supports individual student learning
26	I-S8-1	Varies instructional strategies to engage learners
27	I-S8-2	Uses technology appropriately to enhance instruction
28	I-S8-3	Integrates differentiated instruction for a variety of learning needs
29	I-S8-4	Uses effective communication skills and strategies to convey ideas and information to students
30	P-S9-1	Seeks and accepts feedback to improve teaching effectiveness
31	P-S9-2	Uses self-reflection to improve teaching effectiveness
32	P-S9-3	Upholds legal responsibilities as a professional educator and student advocate
33	P-S9-4	Demonstrates commitment to the profession
34	P-S10-1	Collaborates with colleagues to improve student performance
35	P-S10-2	Works effectively with parents, families, and the community