BUROS
CENTER FOR TESTING

Evaluation of the Nebraska Clinical Practice Evaluation and the Nebraska First Year Teacher Survey

Report Prepared for the
Nebraska Department of Education

Brooke Milstead, BA
Tzu-Yun Chin, PhD

Buros Center for Testing

August 11, 2015

Questions concerning this report should be addressed to:
Tzu-Yun (Katherine) Chin, PhD
Buros Center for Testing,
University of Nebraska-Lincoln
21 Teachers College Hall
Lincoln, NE, 68588-0353
tchin@buros.org

Evaluation of the Nebraska Clinical Practice Evaluation and the Nebraska First Year Teacher Survey

Background

This report highlights the results of an independent psychometric review of the Nebraska Clinical Practice Evaluation (NCPE) and the Nebraska First Year Teacher Survey (NFYTS) conducted by the Buros Center for Testing (Buros) for the Nebraska Department of Education (NDE). The psychometric quality of the instruments was assessed by conducting an item analysis, reliability evaluation, and construct validation. The NCPE data used in this report were provided by several colleges and universities in Nebraska. All of the inquiries were analyzed using the statistical program IBM SPSS Statistics Version 22 with the exception of the construct validation. The construct validity investigation was conducted using the statistical package Mplus (Version 7.31).

Both the NCPE and the NFYTS were developed based on the Council of Chief State School Officers Interstate Teacher Assessment and Support Consortium (InTASC) Model Core Teaching Standards. The InTASC standards demonstrate what knowledge and abilities teachers need in order to effectively teach and improve academic achievement for K-12 students. After the initial development of the instruments, several colleges collaborated to conduct pilot studies. The results of the pilot studies and committee feedback led to modifications to the NCPE and NFYST. Buros' psychometric review was conducted for the 2015 version of the NCPE and NFYTS.

Nebraska Clinical Practice Evaluation (NCPE)

The NCPE aims to assess a teacher candidate's ability to adequately demonstrate each item representing core teaching standards by prompting raters to specify if the teacher candidate is *consistent*, *frequent*, *occasional*, or *rare* (4-point rating scale) in demonstrating the statement outlined in each item. The measure is intended to be used by two raters, including a college supervisor and a cooperating teacher (mentor), concerning a teacher candidate. The current NCPE comprises 28 questions across 11 subscales. The subscales vary in length from one to four items each.

Sample

The NCPE data from each respondent sample (college supervisors and cooperating teachers) were analyzed separately for the item analysis, reliability, and factor analysis. The results of the NCPE data are specific to the respondent population and may not generalize to other respondent groups from other populations. Some participating institutions did not provide both college supervisor and cooperating teacher ratings; therefore, those colleges and universities were only included in some parts of the investigations. In addition, Grace University and Hastings College were not included in the analysis for this report because they implemented the previous versions of the NCPE. Data provided from the colleges were collected during the 2014 fall and the 2015 spring semesters with the exception of the data from the Wayne State College and Creighton University, both of which only included the data collected during the 2015 spring semester.

Several data files provided by the colleges and universities had duplicate ratings made by cooperating teachers or supervisors for individual teacher candidates during the same course. When this occurred, randomized deletion was implemented in order to obtain only one rating associated with a teacher candidate in a single course from each type of respondents (i.e., cooperating teacher and college supervisor). These deletions are necessary for reducing interdependency of the data points, thus the independent observation assumption required for our statistical analyses could be met[1]. Table 1 summarizes the sample sizes, after deleting duplicate ratings, for each institution. The cooperating teacher and college supervisor ratings were also paired in order to investigate interrater reliability and agreement.

[1] Although a more elegant method to account for observation dependency is using a multilevel analysis approach rather than deleting observations, the current data does not allow multilevel analysis. This is because not all teacher candidates received multiple ratings, so multilevel analysis would fail because of data sparseness.

Table 1. NCPE Sample Size by Institution

Institution	Cooperating Teacher Ratings	College Supervisor Ratings	Paired Ratings
Chadron State College	120	132	106
Concordia University	178	169	157
Creighton University	17	17	17
Doane College	88	X	X
Peru State College	96	96	90
University of Nebraska-Kearney	X	288	X
Wayne State College	9	X	X
Total	*508*	*702*	*370*

X denotes data were not available.

Item Analysis

An item analysis was conducted to examine the effectiveness of items. The NCPE item response distributions and statistics are presented in Table 2 for the cooperating teachers and in Table 3 for the college supervisors.

For both cooperating teachers and college supervisors, the percentages of item omission for all items were low (under 4% missing for all items while most items had less than 1% missing). The low missing rates suggest most respondents (both cooperating teachers and college supervisors) were able to provide ratings for all items.

The item means for the 28 items ranged from 3.59 to 3.88 for the cooperating teacher ratings, and from 3.57 to 3.92 for the college supervisor ratings. Along with the frequency count for each score point presented in Table 2 and Table 3, these results demonstrate that all item responses exhibited extremely negatively skewed distributions with most of the item responses being *frequent* (3) or *consistent* (4) for both raters. Furthermore, for both raters, all 28 items had over 60% of the responses assigned to the *consistent* option (4). The *rare* option (1) was seldom used, with most items having no ratings delegated to this response choice. The concentration of item responses toward the higher end of the rating scale posed a concern that the NCPE may not be sensitive enough to differentiate between the varying levels of students, especially at the higher ranks. This concern may not be an issue if the NCPE is used as a criterion-referenced instrument rather than a norm-referenced instrument.

Table 2. NCPE Item Descriptives and Response Distribution for Cooperating Teachers Ratings (*N*=508)

Item†	Mean	Standard Deviation	1‡		2‡		3‡		4‡		Missing	
			n	%	N	%	n	%	n	%	n	%
Q01	3.76	.485	0	0.0%	13	2.6%	96	18.9%	397	78.1%	2	0.4%
Q02	3.77	.481	0	0.0%	13	2.6%	93	18.3%	401	78.9%	1	0.2%
Q03	3.72	.522	0	0.0%	18	3.5%	104	20.5%	383	75.4%	3	0.6%
Q04	3.75	.490	0	0.0%	13	2.6%	101	19.9%	390	76.8%	4	0.8%
Q05	3.77	.480	2	0.4%	7	1.4%	97	19.1%	399	78.5%	3	0.6%
Q06	3.77	.499	1	0.2%	15	3.0%	83	16.3%	406	79.9%	3	0.6%
Q07	3.77	.483	2	0.4%	8	1.6%	95	18.7%	401	78.9%	2	0.4%
Q08	3.62	.600	4	0.8%	19	3.7%	145	28.5%	339	66.7%	1	0.2%
Q09	3.75	.500	1	0.2%	13	2.6%	96	18.9%	394	77.6%	4	0.8%
Q10	3.77	.485	1	0.2%	11	2.2%	93	18.3%	400	78.7%	3	0.6%
Q11	3.80	.455	2	0.4%	6	1.2%	80	15.7%	409	80.5%	11	2.2%
Q12	3.62	.596	3	0.6%	21	4.1%	142	28.0%	339	66.7%	3	0.6%
Q13	3.59	.613	2	0.4%	28	5.5%	143	28.1%	333	65.6%	2	0.4%
Q14	3.72	.540	3	0.6%	13	2.6%	109	21.5%	381	75.0%	2	0.4%
Q15	3.64	.590	3	0.6%	21	4.1%	130	25.6%	353	69.5%	1	0.2%
Q16	3.70	.560	5	1.0%	11	2.2%	116	22.8%	375	73.8%	1	0.2%
Q17	3.69	.553	4	0.8%	11	2.2%	124	24.4%	367	72.2%	2	0.4%
Q18	3.70	.577	5	1.0%	16	3.1%	105	20.7%	380	74.8%	2	0.4%
Q19	3.73	.535	2	0.4%	17	3.3%	95	18.7%	393	77.4%	1	0.2%
Q20	3.69	.556	2	0.4%	18	3.5%	115	22.6%	371	73.0%	2	0.4%
Q21	3.71	.555	4	0.8%	14	2.8%	105	20.7%	382	75.2%	3	0.6%
Q22	3.75	.513	3	0.6%	10	2.0%	97	19.1%	394	77.6%	4	0.8%
Q23	3.88	.369	1	0.2%	5	1.0%	47	9.3%	453	89.2%	2	0.4%
Q24	3.78	.495	2	0.4%	12	2.4%	83	16.3%	410	80.7%	1	0.2%
Q25	3.82	.454	2	0.4%	9	1.8%	66	13.0%	429	84.4%	2	0.4%
Q26	3.81	.493	2	0.4%	16	3.1%	60	11.8%	429	84.4%	1	0.2%
Q27	3.63	.627	5	1.0%	25	4.9%	123	24.2%	352	69.3%	3	0.6%
Q28	3.86	.398	1	0.2%	6	1.2%	58	11.4%	443	87.2%	0	0.0%

† Item stems included in Appendix 1

‡ 1= Rare, 2= Occasional, 3= Frequent, 4= Consistent

Table 3. NCPE Item Descriptives and Response Distribution for College Supervisor Ratings (*N*=702)

Item[†]	Mean	Standard Deviation	1[‡]		2[‡]		3[‡]		4[‡]		Missing	
			n	%	N	%	n	%	n	%	n	%
Q01	3.75	.448	0	0.0%	4	0.6%	169	24.1%	525	74.8%	4	0.6%
Q02	3.75	.447	0	0.0%	5	0.7%	162	23.1%	533	75.9%	2	0.3%
Q03	3.75	.461	0	0.0%	9	1.3%	155	22.1%	535	76.2%	3	0.4%
Q04	3.71	.472	0	0.0%	6	0.9%	190	27.1%	503	71.7%	3	0.4%
Q05	3.72	.479	0	0.0%	10	1.4%	174	24.8%	517	73.6%	1	0.1%
Q06	3.80	.435	0	0.0%	10	1.4%	120	17.1%	569	81.1%	3	0.4%
Q07	3.79	.442	0	0.0%	10	1.4%	128	18.2%	564	80.3%	0	0.0%
Q08	3.71	.488	0	0.0%	11	1.6%	181	25.8%	507	72.2%	3	0.4%
Q09	3.71	.478	0	0.0%	8	1.1%	186	26.5%	505	71.9%	3	0.4%
Q10	3.73	.467	0	0.0%	7	1.0%	176	25.1%	518	73.8%	1	0.1%
Q11	3.79	.436	1	0.1%	5	0.7%	133	18.9%	553	78.8%	10	1.4%
Q12	3.57	.566	0	0.0%	26	3.7%	245	34.9%	425	60.5%	6	0.9%
Q13	3.60	.555	0	0.0%	23	3.3%	230	32.8%	435	62.0%	14	2.0%
Q14	3.65	.524	0	0.0%	16	2.3%	215	30.6%	467	66.5%	4	0.6%
Q15	3.61	.538	0	0.0%	18	2.6%	235	33.5%	446	63.5%	3	0.4%
Q16	3.72	.487	0	0.0%	12	1.7%	173	24.6%	515	73.4%	2	0.3%
Q17	3.67	.506	1	0.1%	9	1.3%	210	29.9%	480	68.4%	2	0.3%
Q18	3.64	.517	1	0.1%	10	1.4%	224	31.9%	460	65.5%	7	1.0%
Q19	3.78	.437	0	0.0%	6	0.9%	144	20.5%	549	78.2%	3	0.4%
Q20	3.71	.483	0	0.0%	9	1.3%	186	26.5%	503	71.7%	4	0.6%
Q21	3.73	.480	0	0.0%	11	1.6%	165	23.5%	510	72.6%	16	2.3%
Q22	3.76	.466	1	0.1%	8	1.1%	147	20.9%	524	74.6%	22	3.1%
Q23	3.92	.283	0	0.0%	1	0.1%	57	8.1%	644	91.7%	0	0.0%
Q24	3.77	.450	0	0.0%	9	1.3%	140	19.9%	543	77.4%	10	1.4%
Q25	3.89	.329	0	0.0%	4	0.6%	67	9.5%	622	88.6%	9	1.3%
Q26	3.83	.402	0	0.0%	6	0.9%	110	15.7%	584	83.2%	2	0.3%
Q27	3.66	.535	2	0.3%	15	2.1%	199	28.3%	473	67.4%	13	1.9%
Q28	3.88	.346	0	0.0%	5	0.7%	74	10.5%	621	88.5%	2	0.3%

[†] Item stems included in Appendix 1

[‡] 1= Rare, 2= Occasional, 3= Frequent, 4= Consistent

Reliability

Reliability analyses were conducted for the overall score and the 11 subscale scores for the NCPE. We examined both the internal consistency reliability and the interrater reliability of the NCPE scores.

Internal consistency. Cronbach's alpha (often referred to as coefficient alpha or alpha) was used to analyze the internal consistency of the scales and subscales, which evaluates how closely the items are related to each other. Cronbach's alpha coefficient ranges from 0, representing no consistency, to 1, representing perfect consistency. In general, higher reliabilities are expected for instruments that are used to make high-stakes decisions. The magnitude of alpha coefficients that have been typically sited as a minimum acceptable value for assessments that are used to make high-stakes decisions range from 0.8-0.9 (Carmines & Zeller, 1979; Nunnally, 1978).

Table 4 shows the means, standard deviations, and Cronbach's alpha coefficients for the NCPE scale and subscales for the cooperating teacher ratings; Table 5 displays these statistics for the college supervisor ratings. Reliability analysis requires complete responses for the scale or subscales; therefore, cases with one or more missing items for a given scale or subscales were not included in the analysis (i.e., listwise deletion), which resulted in varying sample sizes for each reliability analysis. In addition, because internal consistency reliability can only be computed for a scale/subscale with more than one item, the internal consistency for the subscale, *Impact on Student Learning and Development*, could not be assessed.

A high alpha coefficient was observed for the total NCPE scale scores for both the cooperating-teacher ratings (*alpha* =.973) and college-supervisor ratings (a*lpha* =.951). However, several subscales had Cronbach's alpha coefficients less than the minimum value typically accepted for high-stakes situations (i.e., *alpha* > .8).

Table 4. NCPE Scale and Subscale Descriptives and Internal Consistency Reliability for Cooperating Teacher Ratings

Scale / Subscale	# of Items	Total Score Mean	Total Score Standard Deviation	*n*	*Alpha*
NCPE	28	104.56	11.075	460	.973
Student Development	3	11.24	1.276	502	.816
Learning Differences	2	7.51	.863	502	**.733**
Learning Environments	3	11.16	1.351	502	.807
Content Knowledge	3	11.32	1.235	493	.818
Application of Content	2	7.22	1.109	503	.818
Assessment	2	7.36	1.068	505	.886
Planning for Instruction	3	11.09	1.496	503	.859
Instructional Strategies	3	11.14	1.435	503	.842
Professional Learning & Ethical Practice	4	11.41	1.171	501	.856
Leadership & Collaboration	2	7.44	1.017	504	**.771**
Impact on Student Learning & Development	1	3.86	.398	-	-

Bolded alpha values highlight values less than .8

Table 5. NCPE Scale and Subscale Descriptives and Internal Consistency Reliability for College Supervisor Ratings

Scale / Subscale	# of Items	Total Score Mean	Total Score Standard Deviation	n	Alpha
NCPE	28	104.88	8.359	614	.951
Student Development	3	11.25	1.152	694	.806
Learning Differences	2	7.43	.804	699	**.602**
Learning Environments	3	11.30	1.116	697	**.751**
Content Knowledge	3	11.23	1.092	688	**.696**
Application of Content	2	7.17	1.007	683	**.759**
Assessment	2	7.26	.986	698	.839
Planning for Instruction	3	11.03	1.224	691	**.738**
Instructional Strategies	3	11.22	1.125	680	**.736**
Professional Learning & Ethical Practice	4	11.45	.969	670	**.761**
Leadership & Collaboration	2	7.49	.800	688	**.622**
Impact on Student Learning & Development	1	3.88	.346	-	-

Bolded alpha values highlight values less than .8

Corrected item-total correlation and alpha if item deleted. The corrected item-total correlation is computed by correlating the item score with the total score of the other items on the scale. The values can range from -1 to +1 with negative correlations signifying the item may need to be reverse scored. Items with low item-total correlations (close to 0) do not distinguish respondents well, while high item-total correlations discriminate respondents adequately along the construct. Items that have low corrected item-total correlations may need to be removed, revised, or substituted by a superior item.

Alpha if item is deleted is an additional measure of item effectiveness. It reflects the impact of an item removal on the overall score reliability in terms of internal consistency. If the alpha value increases in a meaningful way when an item is deleted, it might be beneficial to remove the item. The NCPE measure of corrected item-total correlations and alpha if item is deleted are shown in Table 6 for the cooperating teacher and college supervisor ratings.

The corrected item-total correlations were adequate for both the college supervisor ratings and cooperating teacher ratings. In addition, the alpha coefficient did not improve with any item removal. Altogether, these statistics suggest all 28 items should be retained in the NCPE scale.

Table 6. NCPE Corrected Item-Total Correlation and Alpha if Item Deleted by Rater

Item[†]	College Supervisor (Alpha=.951)		Cooperating Teacher (Alpha=.973)	
	Corrected Item-Total Correlation	Alpha if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
Q01	.688	.948	.741	.972
Q02	.670	.949	.733	.972
Q03	.677	.948	.783	.972
Q05	.553	.950	.709	.972
Q06	.660	.949	.767	.972
Q07	.600	.949	.747	.972
Q08	.618	.949	.704	.972
Q09	.601	.949	.688	.972
Q10	.588	.949	.749	.972
Q11	.670	.948	.795	.971
Q12	.534	.950	.641	.972
Q13	.639	.949	.742	.972
Q14	.621	.949	.736	.972
Q15	.690	.948	.753	.972
Q16	.692	.948	.747	.972
Q17	.698	.948	.773	.972
Q18	.625	.949	.770	.972
Q19	.609	.949	.786	.971
Q20	.689	.948	.798	.971
Q21	.688	.948	.817	.971
Q22	.538	.950	.678	.972
Q23	.615	.949	.697	.972
Q24	.473	.950	.639	.972
Q25	.670	.949	.748	.972
Q26	.502	.950	.749	.972
Q27	.684	.948	.780	.972
Q28	.602	.949	.731	.972

Interrater agreement and interrater reliability. Interrater agreement and interrater reliability were assessed for the NCPE items. For this analysis, the college supervisor and cooperating teacher ratings were paired by the target of the ratings (i.e., paired by teacher candidate and by each course).

Table 7 presents the percentage of exact agreement and Cohen's Kappa for each item. The percentage of exact agreement between the cooperating teachers and the college supervisors ranged from 63.9% to 87.2%. These medium-high agreement percentages imply that the cooperating teachers and the college supervisors rated the teacher candidates' performance somewhat consistently despite the fact that each brought in different perspectives.

Additionally, Cohen's Kappa is a measure of interrater agreement that ranges in value from -1 to +1. Kappa measures the proportion of agreement of the raters while correcting for chance agreement. Values closer to 1 indicate high agreement while values close to 0 reflect low agreement. For high-stakes situations a minimum Kappa value of .5-.8 is usually desirable (Cohen, 1960; Bock & Brennan, 2002). However, Kappa coefficient has its limitations. For example, Kappa is sensitive to response distributions or base rates. When responses are heavily concentrated in one or two response categories, as is the case in what we have observed with the NCPE data, Kappa values can be lower than expected just because of the way that chance agreement is defined and calculated. Therefore, it is not unexpected to observe low Kappa values (ranging from .072 to .375) for the NCPE.

Table 7. NCPE Percentage of Exact Agreement and Cohen's Kappa

Item	n	% Exact Agreement	Kappa
Q01	364	73.35%	.177
Q02	367	73.02%	.148
Q03	365	74.25%	.280
Q04	364	71.70%	.190
Q05	366	71.31%	.188
Q06	365	75.34%	.216
Q07	368	74.46%	.195
Q08	368	68.21%	.230
Q09	365	77.53%	.289
Q10	368	74.73%	.224
Q11	359	78.83%	.240
Q12	367	66.21%	.251
Q13	358	64.80%	.227
Q14	366	63.93%	.072
Q15	367	64.58%	.184
Q16	367	71.12%	.243
Q17	368	67.66%	.186
Q18	364	71.15%	.261
Q19	366	74.86%	.247
Q20	366	72.40%	.252
Q21	355	67.89%	.113
Q22	352	70.17%	.120
Q23	368	87.23%	.265
Q24	362	80.66%	.375
Q25	364	83.52%	.307
Q26	367	80.93%	.264
Q27	355	86.41%	.192
Q28	368	86.41%	.337

In addition, Pearson product-moment correlation coefficients (*r*) and Spearman rho correlation coefficients (ρ) were calculated to assess the interrater reliability of the scale and subscale scores. These correlation coefficients have theoretical values ranging from -1 to +1 with values closer to 1 indicating high agreement. The Pearson correlation evaluates continuous variables in a linear relationship, while the Spearman correlation assesses the relationship between two continuous or ordinal variables in a rank ordered relationship.

The results of the Pearson and Spearman correlations are presented in Table 8. While all of the Pearson and Spearman correlations are statistically significant, the correlation values are low, especially for high-stakes circumstances. Typically correlation values of .7 or higher are viewed as a minimum satisfactory agreement relationship (Bock & Brennan, 2002). The Pearson correlation for the NCPE full scale was .420, while the Spearman correlation was .354. The subscale correlations ranged from .261 to .453 (Pearson *r*) and .173 to .355 (Spearman *ρ*).

Table 8. Nebraska Clinical Practice Evaluation Correlation between Raters

Scale / Subscale	# of Items	*N*	Correlation	
			Pearson *r*	Spearman *ρ*
NCPE	28	283	.420**	.354**
Student Development	3	357	.405**	.303**
Learning Differences	2	362	.370**	.270**
Learning Environments	3	363	.419**	.309**
Content Knowledge	3	355	.427**	.325**
Application of Content	2	355	.357**	.303**
Assessment	2	366	.261**	.173**
Planning for Instruction	3	359	.400**	.355**
Instructional Strategies	3	349	.361**	.265**
Professional Learning & Ethical Practice	4	350	.450**	.317**
Leadership & Collaboration	2	353	.408**	.303**
Impact on Student Learning & Development	1	368	.453**	.390**

** Indicates correlation is significant at the 0.01 level (2-tailed test).

Reliability analysis discussion. The overall NCPE score (28 items) had very high internal consistency reliability. However, some NCPE subscales, particularly for the college supervisor ratings had alpha coefficients lower than desirable. The low alpha coefficients might be due to the (1) small item number for each subscale, and (2) range restriction of the score variability.

The Cronbach's alpha, as with all reliability, is highly dependent on the number of items on the scale being assessed. Generally speaking, the more items on a scale, the higher reliability will be. The NCPE subscales only have one to four items each. One way to improve the reliability for the subscales is to add additional items to the subscales.

Furthermore, as observed earlier with the item analyses, the response distributions of both the cooperating teacher ratings and the college supervisor ratings were highly weighted toward the values of *frequently* (3) and *consistent* (4) response options. This compression of the response distributions greatly limits the variability of the scores, which affects the reliability in an adverse way. More response rating options may be necessary in order to increase response variability and the capacity to better differentiate among teacher candidates' performance. Increasing response variability, in general, may lead to higher reliability for the subscales.

Yet, the NCPE total scale alpha values do demonstrate some supportive reliability evidence across both the cooperating teacher and college supervisor ratings. The item-scale correlation and alpha if the item is deleted results also endorse the interpretation that the NCPE operates well as a unidimensional scale. The individual item-to-scale correlations were satisfactory, which indicates all of the items interrelate with the rest of the full scale reasonably. The alpha levels were essentially the same when an item was removed, aiding the conclusion that all of the items on the NCPE are highly related to each other for the total scale.

Generally speaking, the college supervisor ratings have lower internal consistency reliability than the cooperating teacher ratings. The college supervisor ratings also had slightly more limited distributions of item responses when compared to the distributions of the cooperating teacher ratings. The differences in the reliability estimates and the response distributions between the two sets of respondents are not necessarily problems because the differences might reflect true differences in different raters' observations. However, it would be prudent to examine whether other nuisance factors might have a role in causing the differences. One of the nuisance factors may be the instructions given to raters. For example, documents that help clarify the intended meaning of the NCPE items and/or help raters define the rating scale (e.g., the *Nebraska Clinical Practice Evaluation [Guidelines with Example Indicators]*) should be available to all raters in order to ensure consistent usage of the NCPE.

The interrater reliability for the NCPE items was quite low between the cooperating teacher and the college supervisor ratings. This may indicate that the college supervisors and cooperating teachers were using and understanding the NCPE in a dissimilar way or they observed different behavior samples from a given teacher candidate. The former is a nuisance factor to be eliminated, but the latter is a useful difference that should be maintained. We encourage an emphasis on ensuring the raters use the measure in a consistent way across response samples. The restriction of the response variability using primarily two options may also adversely impact the correlational values. It is likely that the interrater reliability may be improved when the number of response options increases.

Construct Validation

In order to assess the construct validity of the NCPE, the factor structure was analyzed by conducting confirmatory factor analyses (CFA). The NCPE was developed based on the 10 InTASC Model Core Teaching Standards reflecting 10 initial subscales for the measure. After pilot testing and committee feedback, an additional standard was implemented for the NCPE totaling 11 subscales. A confirmatory factor analysis was conducted to evaluate the 11-factor model of the current NCPE scale separately for the cooperating teacher ratings and for the college supervisor ratings.

Both CFA estimations converged with warnings indicating that the latent variable covariance matrices (Ψ) were not positive definite (which we will discuss later in this section). The model fit indices, including the chi-square test, comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR), were reviewed in order to evaluate whether the data support the hypothetical 11-factor latent structure of the NCPE. The results of the confirmatory factor analysis did not generate strong support for the 11-factor structure of the NCPE. Only one fit indices (SRMR) indicated that the 11-factor model fit the data well. The cooperating teacher data fit the 11-factor model slightly better than the college supervisor data. These results are presented in Table 9 for the cooperating teacher ratings and in Table 10 for the college supervisor ratings. The values suggested in the literature for adequate model-data fit (Hu & Bentler, 1999) are displayed at the bottom of the tables. It is important to note that although the values of the CFI and RMSEA indices did not provide strong support for the 11-factor model, these values did not miss the cut-off values suggested in the literature by much.

Table 9. NCPE Cooperating Teacher Respondents Confirmatory Factor Analysis (*N*=508)

Fit Indices	Results	Conclusions
Chi-Square Test of Model Fit	X^2*(296)*= 876.589 *p*=.0000	Not an exact fit
CFI	0.949	Not a close fit
RMSEA	0.062	Not a good fit
SRMR	0.029	Close fit

CFI > .95 indicates a close fit
RMSEA < .06 indicates a good fit
SRMR < .08 indicates a good fit (Hu & Bentler, 1999)

Table 10. NCPE College Supervisor Respondents Confirmatory Factor Analysis (*N*=702)

Fit Indices	Results	Conclusions
Chi-Square Test of Model Fit	X^2*(296)*= 1093.462, *p*=.0000	Not an exact fit
CFI	0.926	Not a close fit
RMSEA	0.062	Not a good fit
SRMR	0.039	Close fit

CFI > .95 indicates a close fit
RMSEA < .06 indicates a good fit
SRMR < .08 indicates a good fit (Hu & Bentler, 1999)

We further examined the parameter estimates including the factor loadings, factor correlations, and residual variances. We noted that all items loaded on their respective latent factors adequately as hypothesized. We also observed that the latent factors were highly correlated with one another where most correlations were above .8 and many above .9. These high factor correlations were the reason that the previously mentioned analysis warnings of the latent variable covariance matrices (Ψ) were not positive definite. Altogether, the results suggest the number of latent factors probably could be reduced to be less than 11. We also examined the modification indices. Although some cross-loadings and residual correlations were suggested by the modification indices, we did not observe particular patterns that suggest how the factors may be combined.

Given the results of the confirmatory factor analysis, an additional exploratory factor analysis was conducted. The scree plots, shown in Figure 1 and Figure 2, display the eigenvalues against the factor numbers for the cooperating teacher and college supervisor ratings. Examining scree plots provides an empirical revaluation regarding the number of factors that might be appropriate for the data under investigation. It is often recommended that the possible number of factors is the one that is before a sharp drop in eigenvalues. The scree plots shown in Figure 1 and 2 clearly display a distinct drop in the eigenvalues after the first factor for both respondent samples. This provides additional evidence that the NCPE is functioning as a one-dimensional measure as opposed to an instrument with distinct subscales.

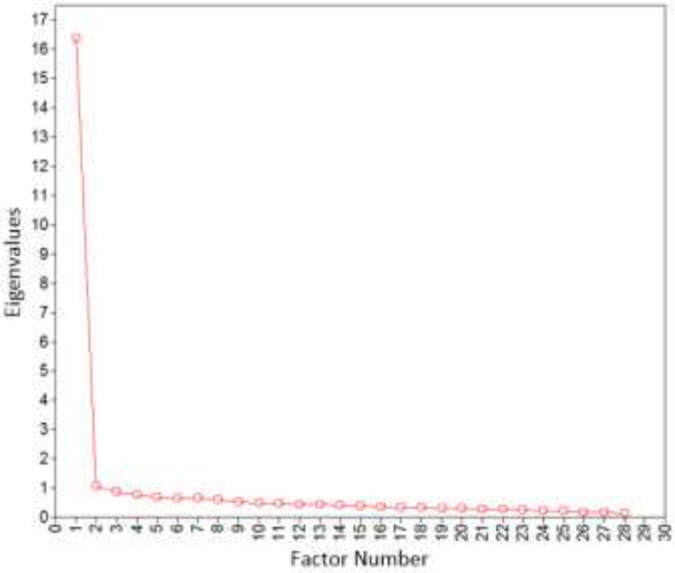

Figure 1. Scree Plot for the NCPE Cooperating Teacher Ratings

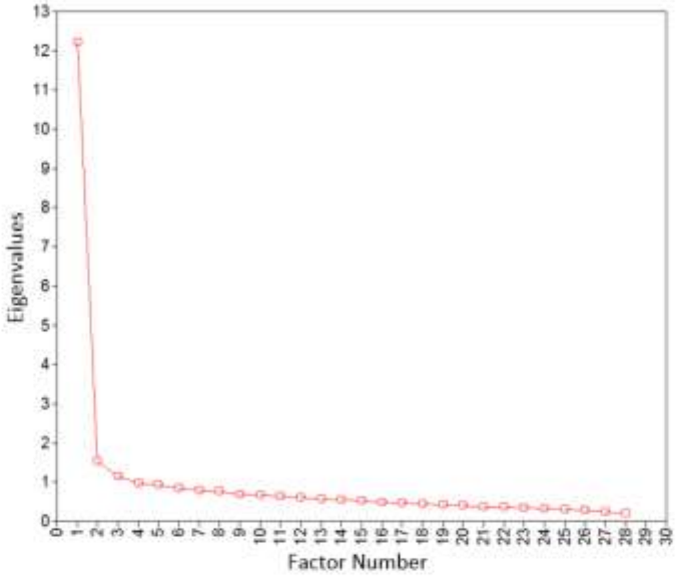

Figure 2. Scree Plot for the NCPE College Supervisor Ratings

Conclusions for NCPE

Our evaluation concludes that the NCPE survey should be viewed and used as a unidimensional scale. The factor analysis results suggest a unidimensional model for the data, and the NCPE total scores exhibited strong internal consistency reliability evidence. However, our evaluation does not support the use of the subscales on the NCPE, especially to make high-stakes decisions due to the lack of reliability. These 11 subscale scores were highly correlated, so each one did not provide unique information about teacher candidates. As a result, we discourage the use of NCPE subscale scores.

The distribution of the response options for all items was predominantly concentrated in the highest rating option (*consistent*). The lowest two response options (*rare* and *occasional*) were used extremely infrequently or not at all. This compression of the empirical usage of the response options does not permit an effective differentiation between teacher candidates across various skill levels if fine discrimination of skill levels is necessary (e.g., ranking candidates and/or programs). As mentioned previously, compression of response options can also have various adverse effects on the psychometric properties of the instrument. Additional response options may be warranted.

Furthermore, we suggest investigating the potential causes for the low interrater reliability/agreement between the cooperating teachers and college supervisors. It is important to establish that different raters can consistently evaluate and quantify performance of teacher candidates using the NCPE. Interrater reliability is often strengthened by establishing clear instructions and guidelines on how to interpret and rate the intended observations. Special care may be needed to ensure the two response samples are using and understanding the measure in a similar way.

Nebraska First Year Teacher Survey (NFYTS)

The NFYTS's purpose is to assess teachers' ability to demonstrate the InTASC teaching standards during their first year of employment. Principles were prompted by email to participate in the online survey that included the NFYTS items at the end of the 2015 spring semester. Respondents were asked to rate each first-year teacher with a 4-point rating scale consisting of *consistent*, *frequent*, *occasional*, or *rare* regarding whether the target first year teacher demonstrated each NFYTS item. The current NFYTS includes 28 items across 11 subscales. Similar to NCPE, the subscales range in length from one to four items.

Sample

Principles were contacted by NDE and prompted to fill out the NFYTS if they worked with a first-year teacher who completed their teacher preparation program at one of the Nebraska higher education institutions. Table 11 outlines the sample sizes by college/university from which the first-year teachers completed their teacher preparation program.

Table 11. NFYTS Sample Size by Institution

Institution	Sample Size
Chadron State College	28
College of Saint Mary	18
Concordia University	13
Creighton University	13
Doane College	34
Grace University	2
Hastings College	15
Midland University	12
Nebraska Wesleyan University	22
Peru State College	20
University of Nebraska- Kearney	67
University of Nebraska- Lincoln	150
University of Nebraska- Omaha	88
Wayne State College	65
York College	7
Total	*554*

Item Analysis

An item analysis was conducted to examine the effectiveness of the NFYTS items. The item response distribution is shown in Table 12. The item means for the 28 NFYTS items ranged from 3.28 to 3.68. The results indicate a negatively skewed distribution with most of the item responses being *Frequent* (3) or *Consistent* (4). Item omission was rare, which suggests most respondents were able to provide evaluations for the NFYTS items.

Table 12. NFYTS Item Descriptive Statistics and Response Distribution (*N*=554)

Item[†]	Mean	Standard Deviation	1[‡]		2[‡]		3[‡]		4[‡]		Missing	
			n	%	n	%	n	%	n	%	n	%
Q01	3.51	.615	0	0.0%	35	6.3%	201	36.3%	315	56.9%	3	0.5%
Q02	3.48	.645	1	0.2%	43	7.8%	201	36.3%	309	55.8%	0	0.0%
Q03	3.44	.666	4	0.7%	42	7.6%	211	38.1%	296	53.4%	1	0.2%
Q04	3.50	.643	2	0.4%	39	7.0%	192	34.7%	318	57.4%	3	0.5%
Q05	3.49	.657	5	0.9%	35	6.3%	194	35.0%	317	57.2%	3	0.5%
Q06	3.56	.654	3	0.5%	41	7.4%	149	26.9%	358	64.6%	3	0.5%
Q07	3.56	.664	9	1.6%	26	4.7%	165	29.8%	350	63.2%	4	0.7%
Q08	3.43	.707	7	1.3%	49	8.8%	195	35.2%	301	54.3%	2	0.4%
Q09	3.56	.599	1	0.2%	28	5.1%	183	33.0%	341	61.6%	1	0.2%
Q10	3.50	.640	2	0.4%	38	6.9%	195	35.2%	317	57.2%	2	0.4%
Q11	3.51	.623	4	0.7%	26	4.7%	206	37.2%	315	56.9%	3	0.5%
Q12	3.28	.731	5	0.9%	77	13.9%	230	41.5%	239	43.1%	3	0.5%
Q13	3.32	.723	4	0.7%	72	13.0%	219	39.5%	257	46.4%	2	0.4%
Q14	3.38	.724	6	1.1%	61	11.0%	200	36.1%	285	51.4%	2	0.4%
Q15	3.31	.736	6	1.1%	73	13.2%	220	39.7%	254	45.8%	1	0.2%
Q16	3.41	.697	5	0.9%	52	9.4%	204	36.8%	291	52.5%	2	0.4%
Q17	3.41	.715	6	1.1%	56	10.1%	197	35.6%	292	52.7%	3	0.5%
Q18	3.39	.704	6	1.1%	52	9.4%	210	37.9%	280	50.5%	6	1.1%
Q19	3.42	.706	4	0.7%	58	10.5%	191	34.5%	300	54.2%	1	0.2%
Q20	3.36	.724	5	0.9%	66	11.9%	206	37.2%	276	49.8%	1	0.2%
Q21	3.42	.706	6	1.1%	52	9.4%	196	35.4%	299	54.0%	1	0.2%
Q22	3.53	.662	1	0.2%	49	8.8%	160	28.9%	344	62.1%	0	0.0%
Q23	3.68	.584	4	0.7%	22	4.0%	121	21.8%	407	73.5%	0	0.0%
Q24	3.45	.666	4	0.7%	42	7.6%	209	37.7%	298	53.8%	1	0.2%
Q25	3.61	.626	6	1.1%	24	4.3%	151	27.3%	369	66.6%	4	0.7%
Q26	3.51	.657	1	0.2%	47	8.5%	175	31.6%	328	59.2%	3	0.5%
Q27	3.43	.690	5	0.9%	48	8.7%	202	36.5%	294	53.1%	5	0.9%
Q28	3.58	.621	3	0.5%	30	5.4%	163	29.4%	356	64.3%	2	0.4%

[†] Item stems included in Appendix 2

[‡] 1= Rare, 2= Occasional, 3= Frequent, 4= Consistent

Internal Consistency Reliability

Reliability analyses were conducted for the total NFYTS scale and 11 subscale scores. The results of the analysis are shown in Table 13. Cronbach's alpha was used to indicate how cohesive a set of items closely related to one another. High reliability coefficients were observed for the NFYTS total scale scores and subscale scores. All of the alpha coefficients were above .8 and the total scale score reliability was extremely high (alpha=.98). The results of this analysis bring support that both the total and subscale scores are reliable. Very high alpha coefficients can indicate that the items are extremely related to each other, which may indicate the scale/subscale is unidimensional. However, extremely high alpha coefficients can also bring some concern that the items are not contributing unique information to the measure or are redundant (Zinbarg, Yoval, Revelle, & McDonald, 2006).

Table 13. NFYSE Scale and Subscale Score Internal Consistency Reliability

Scale & Subscales	# of Items	Total Score		*n*	Alpha
		Mean	Standard Deviation		
NFYTS	28	97.38	14.976	519	.980
Student Development	3	10.43	1.771	550	.909
Learning Differences	2	6.99	1.205	551	.833
Learning Environments	3	10.56	1.815	549	.877
Content Knowledge	3	10.57	1.687	550	.889
Application of Content	2	6.60	1.364	551	.862
Assessment	2	6.69	1.411	551	.933
Planning for Instruction	3	10.22	1.922	548	.896
Instructional Strategies	3	10.21	1.926	553	.885
Professional Learning & Ethical Practice	4	10.66	1.640	549	.886
Leadership & Collaboration	2	6.93	1.266	549	.866
Impact on Student Learning & Development	1	3.58	.621	-	-

Cronbach's alpha is not a measure of dimensionality and is not sufficient evidence to validate subscales.

Corrected item-total correlation and alpha if item deleted. The corrected item-total correlation and alpha if item deleted were calculated for the NFYTS items and are presented in Table 14. Generally speaking, items with low corrected item-total correlation values do not distinguish respondents well along the construct, so they may need to be removed or replaced. Alpha if item is deleted is an additional measure of item effectiveness that reflects the impact of the individual item on the total scale score reliability if that item is removed. If the alpha value increases significantly following an item removal, this item may need to be reviewed and considered for removal or replacement.

As show in Table 14, the corrected item-total correlations were adequate for all items, and the alpha coefficient did not improve with any item removal. Altogether, these statistics suggest that all 28 items should be retained in the NFYTS scale.

Table 14. NFYTS Corrected Item-Total Correlation and Alpha if Item Deleted

Item[†]	Corrected Item-Total Correlation	Alpha if Item Deleted
Q01	.805	.979
Q02	.811	.979
Q03	.806	.979
Q04	.719	.979
Q05	.811	.979
Q06	.756	.979
Q07	.807	.979
Q08	.734	.979
Q09	.784	.979
Q10	.832	.979
Q11	.747	.979
Q12	.784	.979
Q13	.783	.979
Q14	.807	.979
Q15	.818	.979
Q16	.822	.979
Q17	.839	.979
Q18	.817	.979
Q19	.830	.979
Q20	.844	.979
Q21	.721	.979
Q22	.659	.980
Q23	.731	.979
Q24	.797	.979
Q25	.766	.979
Q26	.826	.979
Q27	.774	.979
Q28	.827	.979

Total Scale Alpha= .980

Overall, the NFYTS scores obtained exceptional reliability evidence. The alpha coefficients were very high for the total survey score and high for each subscale. However, the extremely high alpha coefficients observed in conjunction with the high corrected item-total correlation values may suggest that the NFYTS is most likely a unidimensional scale.

Construct Validation

The construct validity of the NFYTS scores was assessed by examining the factor structure with a confirmatory factor analysis. Like the NCPE, the NFYTS was constructed based on the InTASC Model Core Teaching Standards. A confirmatory factor analysis was conducted to evaluate the 11-factor model of the NFYTS. Table 15 presents the resulting model fit indices.

Overall, the results of the confirmatory factor analysis for the NFYTS supported the 11-factor structure of the survey. The Chi-Square test of model fit provided the only statistics that did not support that the 11-factor model fit the data sufficiently. We further noted that all items loaded on their respective latent factors adequately as hypothesized. We also observed that the latent factors were highly correlated with one another. Though these factor correlations for NFYTS were not as high as those observed for NCPE, most of the correlations were above .7, and many were above .8.

Table 15. NFYTS Confirmatory Factor Analysis (N=554)

Fit Indices	Results	Conclusions
Chi-Square Test of Model Fit	X^2*(296)*= 868.923, *p*=.0000	Not an exact fit
CFI	0.964	Close fit
RMSEA	0.059	Good fit
SRMR	0.026	Close fit

CFI > .95 indicates a close fit
RMSEA < .06 indicates a good fit
SRMR < .08 indicates a good fit (Hu & Bentler, 1999)

The scree plot (Figure 3) shows a deep drop in eigenvalues after the first factor. This suggests that the NFYTS may be represented well with just a single factor.

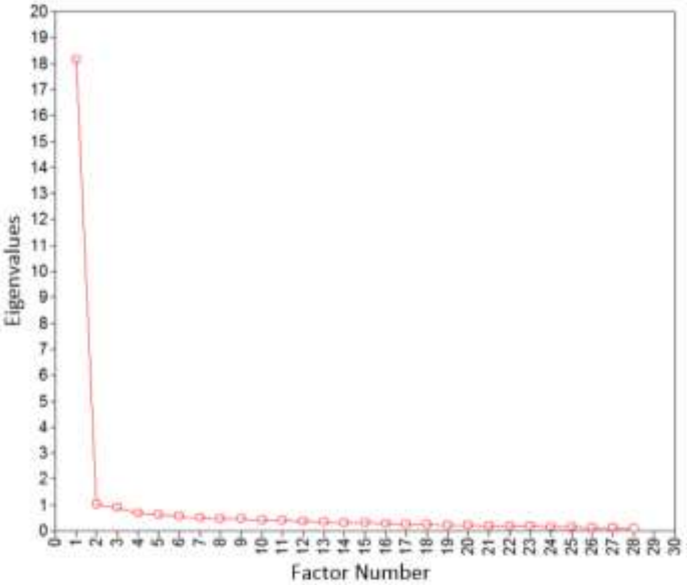

Figure 3. Scree Plot for NFYTS

Conclusions for NFYTS

The NFYTS had very high reliability for the total score and reasonable reliability for the subscale scores. Additionally, the confirmatory factor analysis in general supports the 11-factor structure for the NFYTS data. However, the extremely high alpha coefficient and the high factor correlations suggest the subscale scores were highly redundant of one another. Essentially, the entire NFYTS item set functioned as a unidimensional scale.

Similar to the NCPE, the NFYTS may benefit from expending the response scale options. Although the item response distributions spanned over the four response options slightly better than those observed for NCPE, respondents still predominantly used the highest rating option (*consistent*) to describe the first-year teachers' skills. However, we do not feel an urgent need to expand the response scale for the NFYTS although it could be beneficial. This is because our understanding is that the NFYTS results will be used at the group level (aggregate across multiple first-year teachers) in order to guide teacher preparation policy and/or program evaluation, and the NFYTS score will not be used at the individual level (e.g., rank individual teachers). Adding response option(s) will increase a respondent's cognitive load and will require more training/explanations to help respondents differentiate score points meaningfully and consistently. The cost and benefit of expanding the response scale should be carefully considered.

Future Validity Study

We would like to encourage future validity studies that examine the relationships among the scores of (1) the NCPE, (2) the NFYTS, and (3) student learning outcomes. These additional validity studies will strengthen the validity arguments for using the NCPE for teacher candidate evaluation or using either instrument for teacher preparation program evaluation. In order to conduct these validity studies, there must be a system for tracking teacher candidates and linking various data points.

References

Bock, R., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26* (4).

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons, Inc.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park, CA: Sage.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1).

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

Nunnally, J. C. (1978). *Psychometric Theory* (2nd ed.) New York, NY: McGraw-Hill.

Zinbarg, R. E., Yoval, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a universe of indicators that all have an attribute in common: A comparison of estimators for alpha. *Applied Psychological Measurement*, *30* (2).

Appendix 1

Nebraska Clinical Practice Evaluation Item Stems

Student Development	
Q01	Standard 1.1- The teacher understands how students grow and develop.
Q02	Standard 1.2- The teacher recognizes that patterns of learning and development vary individually within and across the cognitive, linguistic, social, emotional, and physical areas.
Q03	Standard 1.3- The teacher implements developmentally appropriate and challenging learning experiences.
Learning Differences	
Q04	Standard 2.1- The teacher understands individual differences and diverse cultures and communities.
Q05	Standard 2.2- The teacher ensures inclusive learning environments that enable each student to meet high standards.
Learning Environments	
Q06	Standard 3.1- The teacher works with others to create environments that support individual and collaborative learning.
Q07	Standard 3.2- The teacher creates environments that encourage positive social interaction, active engagement in learning, and self-motivation.
Q08	Standard 3.3- The teacher manages student behavior to promote a positive learning environment.
Content Knowledge	
Q09	Standard 4.1- The teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) s/he teaches.
Q10	Standard 4.2- The teacher creates learning experiences that make these aspects of the discipline accessible and meaningful for students to assure mastery of content.
Q11	Standard 4.3- The teacher integrates Nebraska Content Standards and/or professional standards within instruction.
Application of Content	
Q12	Standard 5.1- The teacher understands how to connect concepts across disciplines.
Q13	Standard 5.2- The teacher uses differing perspectives to engage students in critical thinking, creativity, and collaborative problem solving related to authentic local and global issues.
Assessment	
Q14	Standard 6.1- The teacher understands multiple methods of assessment.
Q15	Standard 6.2- The teacher uses multiple methods of assessment to engage students in their own growth, to monitor student progress, and to guide the teacher's and student's decision making.

Planning for Instruction	
Q16	Standard 7.1- The teacher plans instruction that supports every student in meeting rigorous learning goals.
Q17	Standard 7.2- The teacher draws upon knowledge of content areas, curriculum, cross-disciplinary skills, technology, and pedagogy.
Q18	Standard 7.3- The teacher draws upon knowledge of students and the community context
Instructional Strategies	
Q19	Standard 8.1- The teacher understands a variety of instructional strategies.
Q20	Standard 8.2- The teacher uses a variety of instructional strategies to encourage students to develop deep understanding of content areas and their connection and to build skills to apply knowledge in meaningful ways.
Q21	Standard 8.3- The teacher utilizes available technology for instruction and assessment.
Professional Learning and Ethical Practice	
Q22	Standard 9.1- The teacher engages in ongoing professional learning.
Q23	Standard 9.2- The teacher models ethical professional practice.
Q24	Standard 9.3- The teacher uses evidence to continually evaluate his/her practice, particularly the effects of his/her choices and actions on others (students, families, other professionals, and the community), and adapts practice to meet the needs of each student.
Q25	Standard 9.4- The teacher models professional dispositions for teaching.
Leadership and Collaboration	
Q26	Standard 10.1- The teacher seeks opportunities to take responsibility for student learning.
Q27	Standard 10.2- The teacher seeks opportunities, including appropriate technology, to collaborate with students, families, colleagues, and other school professionals, and community members to ensure student growth.
Impact on Student Learning and Development	
Q28	Standard 11.1- The teacher positively impacts the learning and development for all students.

Appendix 2

Nebraska First Year Teacher Survey Item Stems

Student Development	
Q01	Standard 1.1- The teacher understands how students grow and develop.
Q02	Standard 1.2- The teacher recognizes that patterns of learning and development vary individually within and across the cognitive, linguistic, social, emotional, and physical areas.
Q03	Standard 1.3- The teacher implements developmentally appropriate and challenging learning experiences.
Learning Differences	
Q04	Standard 2.1- The teacher understands individual differences and diverse cultures and communities.
Q05	Standard 2.2- The teacher ensures inclusive learning environments that enable each student to meet high standards.
Learning Environments	
Q06	Standard 3.1- The teacher works with others to create environments that support individual and collaborative learning.
Q07	Standard 3.2- The teacher creates environments that encourage positive social interaction, active engagement in learning, and self-motivation.
Q08	Standard 3.3- The teacher manages student behavior to promote a positive learning environment.
Content Knowledge	
Q09	Standard 4.1- The teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) s/he teaches.
Q10	Standard 4.2- The teacher creates learning experiences that make these aspects of the discipline accessible and meaningful for students to assure mastery of content.
Q11	Standard 4.3- The teacher integrates Nebraska Content Standards and/or professional standards within instruction.
Application of Content	
Q12	Standard 5.1- The teacher understands how to connect concepts across disciplines.
Q13	Standard 5.2- The teacher uses differing perspectives to engage students in critical thinking, creativity, and collaborative problem solving related to authentic local and global issues.
Assessment	
Q14	Standard 6.1- The teacher understands multiple methods of assessment.
Q15	Standard 6.2- The teacher uses multiple methods of assessment to engage students in their own growth, to monitor student progress, and to guide the teacher's and student's decision making.
Planning for Instruction	
Q16	Standard 7.1- The teacher plans instruction that supports every student in meeting

	rigorous learning goals.
Q17	Standard 7.2- The teacher draws upon knowledge of content areas, curriculum, cross-disciplinary skills, technology, and pedagogy.
Q18	Standard 7.3- The teacher draws upon knowledge of students and the community context
Instructional Strategies	
Q19	Standard 8.1- The teacher understands a variety of instructional strategies.
Q20	Standard 8.2- The teacher uses a variety of instructional strategies to encourage students to develop deep understanding of content areas and their connection and to build skills to apply knowledge in meaningful ways.
Q21	Standard 8.3- The teacher utilizes available technology for instruction and assessment.
Professional Learning and Ethical Practice	
Q22	Standard 9.1- The teacher engages in ongoing professional learning.
Q23	Standard 9.2- The teacher models ethical professional practice.
Q24	Standard 9.3- The teacher uses evidence to continually evaluate his/her practice, particularly the effects of his/her choices and actions on others (students, families, other professionals, and the community), and adapts practice to meet the needs of each student.
Q25	Standard 9.4- The teacher models professional dispositions for teaching.
Leadership and Collaboration	
Q26	Standard 10.1- The teacher seeks opportunities to take responsibility for student learning.
Q27	Standard 10.2- The teacher seeks opportunities, including appropriate technology, to collaborate with students, families, colleagues, and other school professionals, and community members to ensure student growth.
Impact on Student Learning and Development	
Q28	Standard 11.1- The teacher positively impacts the learning and development for all students.