An Evaluation of the Validity and Reliability of the Intern Keys Assessment

Phase II

Tracy Elder

Atakan Ata

Stephen E. Cramer

College of Education

University of Georgia

September 2016

Introduction

The Intern Keys validation project is being conducted statewide and was funded through a grant from the Georgia Network for Transforming Educator Preparation (GaNTEP) and the Council of Chief State School Officers (CCSSO). The study has been conducted in two phases and the current study, Phase II, is essentially a recapitulation of a recent analysis by Elder, Wang, and Cramer (2015), also referred to as Phase I.

The Intern Keys survey is an evaluation instrument that aims to assess the performance of teacher candidate in the classrooms based on 10 standards. In addition, it also collects the information of raters' teaching background and experiences with this Intern Keys instrument. The standards for evaluating teacher candidates' performance include

(1) Professional Knowledge,
(2) Instructional Planning,
(3) Instructional Strategies,
(4) Differentiated Instruction,
(5) Assessment Strategies,
(6) Assessment Uses,
(7) Positive Learning Environment,
(8) Academically Challenging Environment,
(9) Professionalism, and
(10) Communication.

Educator Preparation Providers (EPPs) in the state of Georgia are required by the Georgia Professional Standards Commission to prepare teacher candidates for the evaluation instrument they will be subject to when they become teachers in Georgia, the Teacher Keys Effectiveness System (TKES). This report will discuss the use of the Teacher Assessment on Performance Standards (TAPS), one of the components of the TKES, to assess readiness of candidates to be a teacher in Georgia schools.

Beginning fall of 2013, many EPPs across the state of Georgia began using the TAPS as a summative assessment at the end of the clinical practice. Each EPP was encouraged by the Georgia Department of Education (GaDOE) to have at least one faculty or staff member complete the TKES credentialing. To make a clear distinction between the state's valid and reliable evaluation system, including the TAPS, and the EPPs' use of the standards and rubrics, the pre-service instrument was named the Intern Keys.

The purpose of this validation project is to assess the validity and reliability of the Intern Keys as an instrument to be used with pre-service teacher candidates. Data collected from candidate supervisors employed by the EPPs as well as the candidate's mentor are used to determine the

candidate's readiness to be certified as a teacher in Georgia. Aggregate data at the program and EPP level are used for program improvement purposes.

This report provides the findings for Phase II of the project. A detailed report that covers the findings from the Phase I can be found on the project website (http://epr.coe.uga.edu/evaluation-systems/intern-keys-validation-project/). Briefly, results from the first phase (2015) showed that the Intern Key Instrument was highly reliable. The results also indicated high level of internal consistency among all 10 standards. The internal consistency of mentors and supervisors was equal, indicating that professionals in these two roles applied the instrument in a similar way.

However, there is always room for improvement. Although the instrument had a high reliability, the differences between mentors and supervisors were relatively larger on standard 4 (Differentiated Instruction), 5 (Assessment Strategies), and 6 (Assessment Uses) than other standards. Figure 1 shows agreement levels between faculty supervisors and mentor teachers.

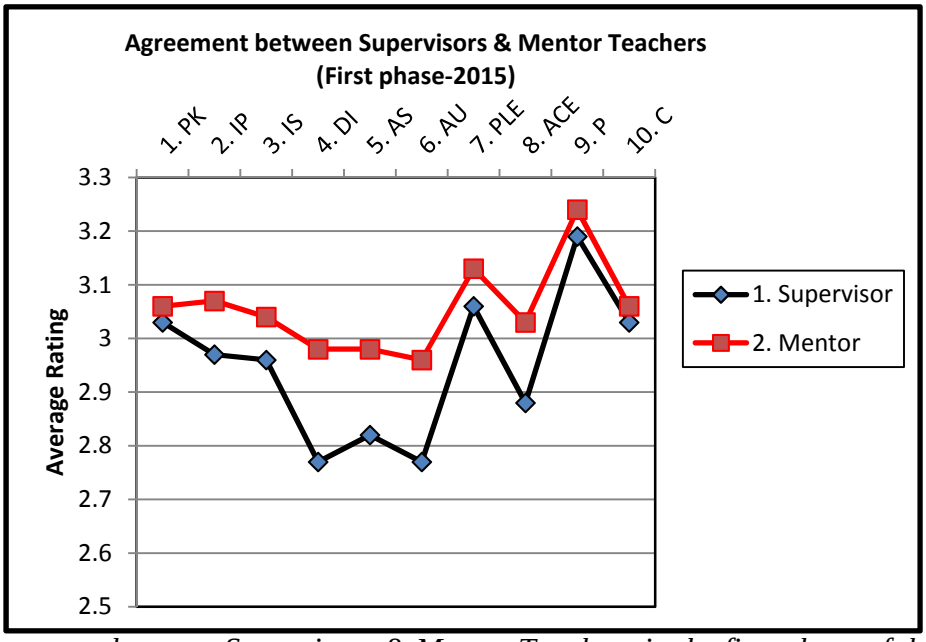


Figure 1: *Agreement between Supervisors & Mentor Teachers in the first phase of the project.*

Also, the ratings from supervisors were a little lower than those of mentors, indicating the supervisor group was a little more conservative or strict than the mentor group on average. Based on these results, some changes and adjustments were made, and a second round of analysis has been conducted.

Please see the Phase I Final Report for the Intern Keys Validation Project (2015) for more details on the preliminary work conducted with EPPs prior to use of the instrument as well as the methodology used to analyze the data collected during the study.

Based on findings from the Phase I study, changes were made to the training materials to ensure a deeper understanding of each standard with particular attention paid to standards identified as

having the lower levels of agreement between the two evaluators – mentor teachers and EPP supervisors. Additionally, changes were made to the rubric performance levels on the state of Georgia TAPS instrument; therefore, the same revisions were applied to the Intern Keys instrument. These changes were not expected to negatively influence the findings of the phase II study.

Validity Support for Intern Keys

The essential validity aspect of the Intern Keys instrument is its close relationship to the Teacher Assessment on Performance Standards (TAPS) process of the Teacher Keys. Indeed, the Standards were adopted almost word for word. This is seen as appropriate, since the TAPS is Georgia's attempt to define the process of classroom teaching and student development, which is precisely what the student teacher is working to internalize. The Teacher Keys Handbook makes this explicit: "Performance standards refer to the major duties performed by a teacher." The Handbook further states that "Evaluators should always refer to the Performance Standards when rating a teacher." The participants endorsed this concept as a means of doing a valid and reliable evaluation of candidates across a variety of settings and content areas.

The TAPS were validated extensively before their inclusion in the TKES process. James Stronge, of the College of William and Mary, has provided extensive references on this process. A more recent validation study was performed at the Georgia Center for Assessment (GCA) at the University of Georgia. That study dealt primarily with the construct validity of the TAPS, not its content, and found that the internal consistency (ordinal alpha) of the TAPS was .95, a very high value. A similar analysis of the Intern Keys internal reliability appears later in this report, and replicates the GCA finding.

Validity approaches

The validity of an instrument can be assessed in at least four ways. In ascending order of rigor these are: Face Validity, Content Validity, Criterion-related Validity, Construct Validity, and Faith Validity.

The **face validity** of the Teacher Keys instrument is obvious. This is supported by the Teacher Keys Handbook. Since the Intern Keys instrument is essentially identical, its face validity is likewise obvious.

To establish **content validity**, the content and language of the Intern Keys can be compared to a variety of documents that are widely accepted by the Georgia Department of Education and other groups. Perhaps the most complete of these is the *InTASC Model Core Teaching Standards and Learning Progressions for Teachers* (InTASC). This publication of the Council of Chief State School Officers (CCSSO, 2013) lays out ten Standards which largely mirror the ten Standards of the Intern Keys, although the match-up is not precise. This document is supported by a literature review on the CCSSO website (www.ccsso.org/intasc) that undergoes continual updating. A

side-by-side listing of the InTASC and Intern Keys Standards appears as Table 1. The individual standard matching from one set to the other is left to the reader, but it is clear that both sets deal with the same underlying construct.

The **criterion-related Validity** of the Intern Keys instrument still needs to be established. There was no opportunity in the current study to collect criterion data to use for comparison. Indeed, the selection of the data to be used as a criterion itself requires considerable consideration, to be sure that it actually reflects the construct of Teacher Performance that we want to define. One possibility might be the grade point average in methods courses for candidates. This may not be terribly useful, however, since (a) there may be a restriction of range in the grades, due to grade inflation and (b) many of the behaviors identified in the Intern Keys are more predispositions than teachable skills. The idea remains a possibility for further research.

The Teacher Effectiveness Measure, which contains the Teacher Assessment on Performance Standards (TAPS), also includes measures of student performance, the Student Growth Percentile (SGP) and Student Learning Objectives (SLO). Although it might seem attractive to use student scores as a criterion measure for Intern Keys, it should be noted that in a study by the Georgia Center for Assessment (GCA, 2014) the correlations of TAPS with SGP and SLO were both significant but very low (0.24 and 0.17). The inclusion of student data in the Teacher Effectiveness Measure (TEM) is well established, but grounding a validity argument on a measure that explains such a small proportion of variance (3%-6%) is not very persuasive.

Finally, we shall consider **construct validity** evidence for the Intern Keys. We again refer to the GCA study of TKES, where the question of the relationship of the TEM (not TAPS) score and years of experience was tested. The correlation between TEM and years of experience was 0.01, essentially zero. In another analysis, GCA determined that the correlation between experience and SGP and SLO was likewise vanishingly small (-.003 and .06), again not accounting for a meaningful amount of variance. These findings can be seen as providing divergent support for the use of Intern Keys, since candidates with very little experience are not at a major disadvantage compared with more experienced teachers in the application of the Standards.

GCA went on to examine the TEM score with a multiple regression analysis. This tested the hypothesis that TEM score was related to the demographics of the class. Although the percent of economically disadvantaged (ED), disabled (SWD), and limited English proficiency (LEP) were significant individually, the model explained only 9% of the variance in TEM score, indicating that the measure is providing a clear evaluation of the teacher. A separate GCA study (GCA, 2013) did find a correlation of -.43 between TEM score and percent ED, although the same sample found very small correlations between TEM and SWD or LEP. Since the Intern Keys instrument is essentially identical with the TAPS, this is indirect evidence that the Intern Keys measure is based mainly on the performance of the candidate.

The GCA study also examined the internal consistency of the TAPS, which was reported as .95, a value in line with the reliabilities noted later in this report for the Intern Keys.

Finally, the GCA study performed an iterative principal factor analysis of the TAPS instrument. Application of the Kaiser-Guttman rule and the scree plot both clearly indicated that the TAPS has only one factor, which they called Teacher Performance. The conclusion that an identical instrument, Intern Keys, is likewise unitary is clear.

Finally, the term **faith validity** was coined at the annual CCSSO meeting in Boulder, CO in 1990. It is used when you sincerely believe that you are measuring what you say you are measuring. It is appropriate to place your right hand on your heart when invoking faith validity. We sincerely believe that the Intern Keys measure important qualities of the teaching candidate.

Reliability Analysis of the Intern Keys Instrument

Two types of raters score the performances of candidate teachers on a rating scale of 1-4: mentors and supervisors. Each rating category is corresponding to each level of achievement. Level 4 (with a rating score of 4) is the highest and level 1 (with a rating score of 1) is lowest. Each standard has different descriptions for each rating category.

In order to ensure that the survey is reliable before making decisions upon candidate teachers, reliability evidence needs to be gathered for this instrument. The theoretical framework for the study can be found in the Phase I report and will not be discussed in this report.

Methods

The raters are trained in order to use the instrument in a consistent manner and evaluate candidate teachers accurately. Intern Keys Validation Project Orientation learning module as a video training method is available on the University of Georgia's Educator Preparation Resources (http://epr.coe.uga.edu/evaluation-systems/intern-keys-candidate-assessment-on-performance-standards/). Besides video orientation to the instrument, live training methods are also available through Educator Preparation Provider (EPP) partners (University or College who prepares candidates), RESA, school district, school, mentor or other means. Raters fill in the Intern Keys instrument as an evaluative instrument based on observation of candidate teachers' teaching and other activities in the classroom. Information on the preparation of teacher candidates on the Intern Keys instrument completed by the mentor teacher and/or EPP supervisor before or during the student teaching experience is also gathered by the instrument. The frequencies of preparations are reported in the result section. The reliability coefficients of the instrument include Cronbach alpha (CCT). In addition, inter-rater agreement is evaluated by Cohen's kappa.

Results and Discussion

In the first phase of this study, there were 293 teacher candidates and 297 raters with 482 ratings analyzed. This current analysis for Phase II had 480 teacher candidates from 12 institutions, and each teacher candidate was rated by one mentor public school teacher AND one faculty supervisor yielding to 960 ratings (after adjustments for incompletes and duplicates). 1,336 mentor teachers or faculty supervisors at 12 different institutions were invited via email, and a total of 1,177 responses were recorded.

Reliability analysis

Cronbach's alpha is a measure of how well the items included measure the same construct, or "hang together."

Sample chosen

The Phase II sample size was 480, to include candidates rated by one mentor and one supervisor.

Reliability Statistics

Cronbach's Alpha	N of Items
.898	10

This is an excellent alpha.

Now, comparing the mentors and the supervisors, we see that they are equal in terms of same construct.

Mentors:

Reliability Statistics

Cronbach's Alpha	N of Items
.901	10

Supervisors:

Reliability Statistics

Cronbach's Alpha	N of Items
.891	10

Agreement

Two professionals rated each candidate. We compared these ratings. In order to control for random agreements, we also calculated Cohen's Kappa. Cohen's kappa measures the inter-rater agreement after adjusting for the chance agreement (Cohen, 1960). The chance agreement is the probability when the occurrences of two events are independent. This is especially an issue in a short rating scale. The coefficient is calculated by using formula:

(observed agreement-expected agreement) / (1- expected agreement).

According to Landis & Koch (1977), Kappa might be interpreted as in the table below.

Kappa	Interpretation
< 0	Poor agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

In this study;

1) Cohen's Kappa for Standard 2 was 0.26. It is fair agreement.
2) Cohen's Kappa for Standard 3 was 0.17. It is slight agreement.
3) Cohen's Kappa for Standard 4 was 0.19. It is slight agreement.
4) Cohen's Kappa for Standard 5 was 0.12. It is slight agreement.
5) Cohen's Kappa for Standard 6 was 0.18. It is slight agreement.
6) Cohen's Kappa for Standard 7 was 0.23. It is fair agreement.
7) Cohen's Kappa for Standard 8 was 0.16. It is slight agreement.
8) Cohen's Kappa for Standard 9 was 0.24. It is fair agreement.

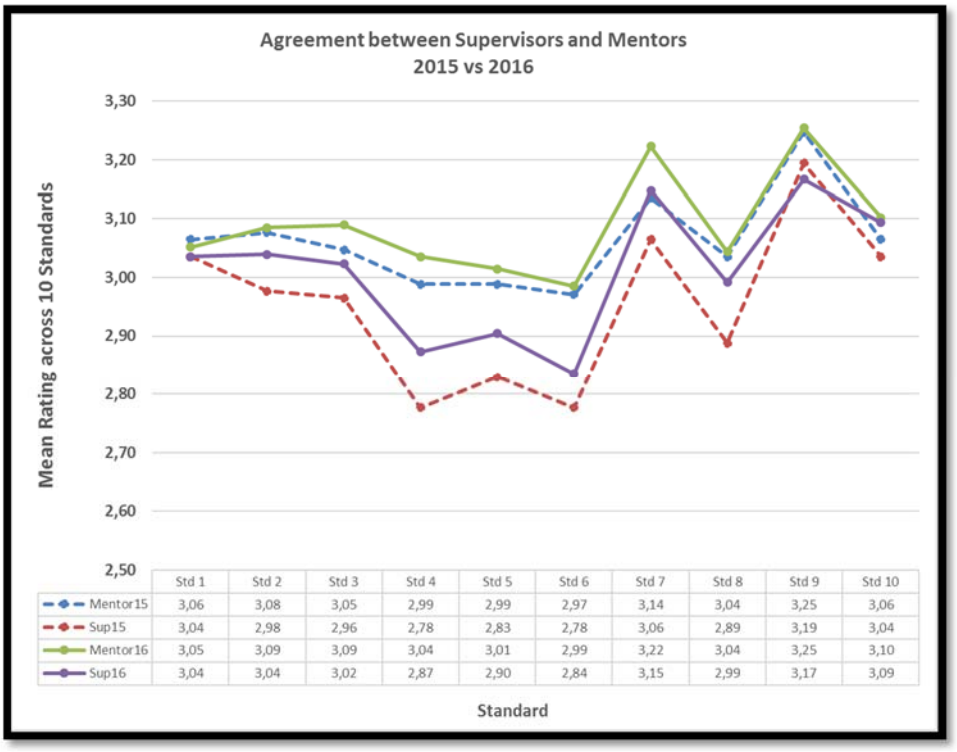
Summary of 10 standards

Standard	Exact Agreement	Percent (%)	Adjacent Agreement	Percent (%)	Exact + Adjacent	Discrepant Percent
1	340	70.8%	136	28.3%	99.2%	0.83%
2	344	71.7%	130	27.1%	98.8%	1.25%
3	322	67.1%	154	32.1%	99.2%	0.83%
4	295	61.5%	170	35.4%	96.9%	3.13%
5	334	69.6%	141	29.4%	99.0%	1.04%
6	312	65.0%	164	34.2%	99.2%	0.83%
7	304	63.3%	168	35.0%	98.3%	1.67%
8	317	66.0%	159	33.1%	99.2%	0.83%
9	308	64.2%	156	32.5%	96.7%	3.33%
10	341	71.0%	136	28.3%	99.4%	0.63%

Exact agreement: mentor and supervisor gave exact same ratings.
Adjacent agreement: the absolute difference between ratings given by mentor and supervisor is 1.
Discrepant: the absolute difference between ratings given by mentor and supervisor is greater than 1.

This table gives the level of agreement of Mentors and Supervisors for each standard and for the total score. We use here the same agreement definition as is used in the Georgia Writing Assessment, that inter-rater differences of one are indicative of the raters using different definitions or observing on different days, but generally do not indicate disagreement. Thus the agreement rate is Exact plus Adjacent Agreement. The table shows that the level of agreement was very high. For all Standards, Exact plus Adjacent Agreement is over 95%. Exact Agreement ranges from 61% to 71% across the Standards.

It is clear that Supervisors and Mentor Teachers agree very well on the extent to which the teacher candidates they evaluate meet the Intern Keys Standards.



	Std 1	Std 2	Std 3	Std 4	Std 5	Std 6	Std 7	Std 8	Std 9	Std 10
Mentor15	3,06	3,08	3,05	2,99	2,99	2,97	3,14	3,04	3,25	3,06
Sup15	3,04	2,98	2,96	2,78	2,83	2,78	3,06	2,89	3,19	3,04
Mentor16	3,05	3,09	3,09	3,04	3,01	2,99	3,22	3,04	3,25	3,10
Sup16	3,04	3,04	3,02	2,87	2,90	2,84	3,15	2,99	3,17	3,09

c

Figure 2: *Agreement between Supervisors & Mentor Teachers: 2015 and 2016*

Here, figure 2 shows the comparison between agreement levels for the first (2015) and second (2016) phases across 10 standards. Dotted lines are ratings from 2015 and straight lines are from 2016. As the figure displays, there is a general increase in ratings, but most importantly, there is a decrease in mentor and supervisor differences. This means that efforts to improve the Intern Keys assessment over the past year have yielded positive results.

Additional Activities and Studies

The current GaPSC rule 505-3-.01 for teacher preparation stipulates that the candidate be made aware of the content and expectations of the State's performance evaluation system. The Intern Keys evaluation has been developed to mirror the Teacher Assessment on Performance Standards (TAPS), a component of the Teacher Keys. For the current study, we developed an orientation video to prepare Mentors and Supervisors for the implementation of this instrument as a summative assessment of candidate performance during the clinical practice experience.

The following recommendations, originally presented in the Phase I report, are worth mentioning again in this final report for Phase II of the study. At a minimum, we recommend that all EPPs organize beginning of the year meetings with raters and candidates to move toward a common understanding of the standards and the performance levels. In many cases, these meetings will be somewhat redundant, particularly if the EPP/program is using the Intern Keys structure in its methods classes. We see this as the preferred method of preparation for teacher candidates and can support preparation efforts with materials, speakers, video examples, and documents.

We would further recommend that EPPs hold orientation meetings on Intern Keys for any new supervisors that they may hire. These should include practice on rating video examples. The current report lists agreement levels for the ten standards in Table 5. We encourage EPPs to refer to this list to see which standards appear to have the lowest levels of agreement, and to focus training efforts on them. Table 5 shows that Professional Knowledge and Assessment Strategies have the best exact agreement between the two raters. However, if adjacent agreement is considered, Cohen's kappa shows lower levels of agreement for Assessment Strategies, Assessment Uses, Academically Challenging Environment, and Professionalism. If training resources are limited, we would suggest focusing on these standards rather than those standards with substantial levels of agreement. Our support efforts—video and text-- will be planned with this in mind. Video examples that we produce will highlight candidate behaviors that appear to lead to discrepant ratings.

Validation efforts so far have mainly dealt with the content of the instrument. With wider use, we will have access to other data related to these teachers. For the state TEM, student data forms a significant part, both student reports and test scores. In the current study, we did not have access to these data, but in the future access to these data will allow us to assess the predictive validity of the instrument.

The GCA studies mentioned above made extended use of multiple regression techniques to examine the predictors of the TEM and TAPS scores. The predictors used in those studies were student and school characteristics. With greater access to candidates' data, we will be able to test the hypotheses that various candidate variables—gender, ethnicity, content specialty, level of degree, geographic locality, etc.—may predict some of the variance in Intern Keys scores. In this same analysis, hypotheses about the relationship of Intern Keys scores to the evaluators'

demographic and professional characteristics may be tested. We have no evidence at this point that there are any issues of bias in the application of the Intern Keys, but studies that examine the demographic characteristics of evaluators and candidates will help to settle any such concerns.

Predictive validity can also be assessed by comparing The Intern Keys scores with the eventual Teacher Keys TAPS score. Applying the same standards to the same professionals two consecutive years should certainly display strong correlations.

References

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Council of Chief State School Officers. (2013). Interstate Teacher Assessment and Support Consortium *InTASC Model Core Teaching Standards and Learning Progressions for Teachers* (InTASC). Washington, D.C.: CCSSO.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Elder, Wang, and Cramer (2015). *An Evaluation of the Validity and Reliability of the Intern Keys Assessment*.

Engelhard G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge Academic.

Georgia Center for Assessment. (2014). *Assessing the validity and reliability of the Teacher Keys Effectiveness System (TKES) and the Leader Keys Effectiveness System (LKES) of the Georgia Department of Education.* Athens, GA: University of Georgia

Georgia Center for Assessment. (2013). *Statistical analysis of the teacher effectiveness measure.* Athens, GA: University of Georgia

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: Mesa Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Stronge & Associates (2013). *Stronge teacher evaluation system: A validation report*. Williamsburg, VA: The College of William and Mary.

Table 1: Comparison of InTASC and Intern Keys Standards

InTASC Standards	Intern Keys Standards
Standard #1: Learner Development	Standard 1--Professional Knowledge
Standard #2: Learning Differences	Standard 2—Instructional Planning
Standard #3: Learning Environments	Standard 3--Instructional Strategies
Standard #4: Content Knowledge	Standard 4--Differentiated Instruction
Standard #5: Application of Content	Standard 5--Assessment Strategies
Standard #6: Assessment	Standard 6--Assessment Uses
Standard #7: Planning for Instruction	Standard 7--Positive Learning Environment
Standard #8: Instructional Strategies	Standard 8--Academically Challenging Environment
Standard #9: Professional Learning and Ethical Practice	Standard 9--Professionalism
Standard #10: Leadership and Collaboration	Standard 10—Communication

Appendix A: EPP Instrument validation worksheet

Intern Keys Validation Session
December 8-9, 2014

Your ID _____

1. Evidence can be long; note exactly which part you used to make your decision. Use the time code for video. Indicate page and location for text.
2. Which Standard are you rating?
3: Which level of performance do you observe? The online application uses "Grade."
4: What do you want to remember about this rating to share in the discussion? What was especially useful? What would have made the task easier?

Artifact	**Evidence Location[1]**	**Standard[2] (1-10)**	**Grade[3] (4-1)**	**Comment[4]**
Vid._____	*Minutes: seconds*			
Art._____	*Page, location*			