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

Tracy Elder

Jue Wang

Stephen E. Cramer

College of Education

University of Georgia

The Intern Keys validation project is being conducted statewide and was funded through a grant from the Georgia Professional Standards Commission (GaPSC) and the Council of Chief State School Officers (CCSSO). The survey as an evaluation instrument 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 of 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 validation and reliability examination began with a statewide meeting of EPP staff in Macon. At this meeting, the issue of the appropriate "comparison group" was raised, namely, is the intern to be evaluated as an intern, or with respect to first year teachers currently employed. The consensus of the group was that the Intern Keys needed to stand as an instrument for student teachers, and not compare them to working teachers, who generally have more experience and more support.

Further discussion related to the definition of the points on the four point scale. The consensus of the group was that a score of four should be awarded based on the literal meaning of "Exemplary," to identify a performance that could be used as an example to other interns. The stated goal level of performance was three, “Proficient,” which the participants agreed meant that the candidate was worthy of an offer of a position.

At the same meeting, participants used the Intern Keys instrument to rate video representations of teaching behaviors, and discussed their ratings. The goal of this exercise was to identify Standards where fewer than 80% of participants agreed on the rating. For most standards, the 80% criterion was met initially; group discussion led to greater than 80% agreement on the rest. The worksheet used by participants is included as Appendix A.

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 Profile (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

Introduction

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.

Theoretical framework

Engelhard (2013) classified measurement theories into two research traditions, which are test-score tradition and scaling tradition. Classical test theory (CCT) and generalizability (G) theory are key models in the test-score tradition. Item response theory (IRT) is categorized into the scaling tradition. Coefficient alpha (Cronbach, 1951) has been primarily used to estimate internal consistency of an instrument based on CCT. The general formula is as below (Cronbach, 1951).

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right) \tag{1}$$

where V_i is the variance for each standard and V_t is the variance for total score of 10 standards. n refers to the number of standards.

G theory focuses on the estimation of variance components for error variances and it is developed based on CCT and analysis of variance (Brennan, 2001). Analyses based on G theory has two types including G study that estimate the variance component from each error source and decision (D) study that calculate reliability coefficients in order to facilitate making decisions. From the perspective of G theory, generalizability coefficient defined by Cronbach et al. (1972) is a reliability-like coefficient. It estimates how consistent the ratings are and how much one can generalize to other settings. Generalizability coefficient is similar to Cronbach alpha in calculation that variances explained by persons are divided by the total observed variances, and it is identical to Cronbach alpha in a crossed design study. The formula in used for calculation of Generalizability coefficient is as follows (Brennan, 2001).

$$E_{\rho^2} = \frac{\sigma^2(\mathrm{t})}{\sigma^2(\mathrm{t})+\sigma^2(\delta)} = \frac{\sigma^2(\mathrm{t})}{\sigma^2(\mathrm{t})+\left[\frac{\sigma^2(\mathrm{ts})}{\mathrm{n}'_\mathrm{s}}+\frac{\sigma^2(\mathrm{tr})}{\mathrm{n}'_\mathrm{r}}+\frac{\sigma^2(\mathrm{tsr})}{\mathrm{n}'_\mathrm{s}\mathrm{n}'_\mathrm{r}}\right]} \tag{2}$$

where t represents for the facet of candidate teachers, r is for raters' role, s stands for standards, and $\sigma^2(\delta)$ refers to the relative error variance component. The n with different subscripts refers to sample size for the corresponding facet.

IRT as in the scaling tradition aims for invariant estimation of mapping persons and items on the same scale. It computes measures or adjusted scores for person and items; differently, CCT and G theory both use raw scores directly in the analysis. Linacre (1989) introduced Many-facet Rasch model (MFRM) which is implemented into FACETS computer program. The reliability coefficient (reliability for separation) for the instrument indicates to what extent it can measure for person abilities. The MFRM model is specified as following.

$$\ln\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = \theta_n - \lambda_i - \delta_j - \tau_k \qquad (3)$$

where

P_{nijk} = probability of teacher candidate *n* receiving a rating of *k* by rater *i* on standard *j*,

P_{nijk-1} = probability of teacher candidate *n* receiving a rating of *k-1* by rater *i* on standard *j*,

θ_n = performance of candidate teacher *n*,

λ_i = severity of rater *i*,

δ_j = difficulty of standard *j*,

τ_k = difficulty of category *k* relative to category *k-1* for standard *j*.

Inter-rater agreement indicates to what degree raters agree with each other. Cohen's kappa (κ) 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 κ is calculated by using formula: (observed agreement-expected agreement)/(1- expected agreement).

The Intern Keys instrument is evaluated by multiple methods with different coefficients in this study.

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/state-evaluation-systems/intern-keys-validationproject/). Besides video training, live training methods are also available provided by Educator Preparation Provider partner (University or College who prepares candidates), RESA, school district, school, mentor or other means. Raters may take one or multiple training sessions. The rating scores are compared between two training methods - video and live. After training, raters fill in the survey as an evaluative instrument during observing candidate teachers' teaching and doing other activities with the students. The preparations of candidate teachers had before or during teaching are 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) that analyzed by SPSS software (IBM, 2013), Generalizability coefficient (G theory) done in GEANOVA (Brennan, 2001), and reliability of separation (IRT) obtained in FACETS computer program (Linacre, 1989). In addition, inter-rater agreement is evaluated by Cohen's kappa.

Results and Discussion

There are 296 intern teachers being evaluated by 304 raters (including 2 anonymous) in this study, which leads to 493 ratings in total. Eleven ratings are not complete either due to drop out or missing questions. Three intern teachers and 7 raters with these 11 incomplete ratings are removed from data analyses. Therefore, 293 intern teachers and 297 raters with 482 ratings are analyzed.

In order to examine training effects, ratings of each standard and total score of 10 standards are compared between video method and live method(s) by using independent *t*-test in SPSS (IBM, 2013). Since raters may attend one or more training sessions, we only compared raters who only watched the video (N=195 ratings) with raters who didn't watch the video (i.e., attended one or more live sessions; N=127 ratings). Independent sample *t*-tests indicated significant differences between two training methods in Standard 1 – Professional Knowledge with t=1.978, p<.05, Standard 4 – Differentiated Instruction with t=2.427, p<.05, Standard 6 – Assessment Uses with t=2.771, p<.05, Standard 9 – Professionalism with t=2.619, p<.05, Standard 10 – Communication with t=-2.091, p<.05, and Total Score with t=2.500, p<.05 (Table 1). The result shows raters who only watched video tended to rate significantly higher than those who are trained with live method(s) on those standards (Figure 1).

The frequencies of candidate teachers' preparations before teaching are shown in Table 2. Most candidate teachers had seen a copy of the Intern Keys instrument (51.66%) or Teacher Keys instrument (34.85%), and discussed the standards with raters (68.46%). 41.29% raters reported that Teacher Keys evaluation was integrated into the candidate's preparation program, and 41.08% raters provide mid-point performance feedback based on the Teacher/Intern Keys standards to the candidate teachers. However, the correlation coefficients between preparations and standard scores are all within .13, indicating very small relationship among them.

There are 170 intern teachers being rated by one mentor and one supervisor leading to 340 ratings in total. The rest intern teachers are rated only by mentor or supervisor. The reliability analysis is based on the 340 ratings with a complete design. The Cronbach alpha is .897 for 340 rating scores, .901 for ratings from mentors, and .891 for ratings from supervisors. Cronbach alpha examines the internal consistency for an instrument. The results indicate high reliability of internal consistency among all 10 standards. Also, the internal consistency of mentors and supervisors is equal, indicating that professionals in these two roles apply the instrument in a similar way.

The G study evaluated three components: candidate teachers as a random facet, raters' role as a random facet, and standards as a fixed facet. Because of a linking problem, individual rater effect is ignored; instead, effect of raters' role (i.e., mentor or supervisor) is examined. This analysis has a nested design that standards are nested within raters. The variance components for each facet obtained from G study are reported in Table 3. Person facet of candidate teachers takes a large portion of total variability, 15.79%. Raters’ role is only accounted for 1.64%, indicating small variances in ratings between supervisors and mentors. Interaction between standards and raters takes account for 4.71% of total variability. It indicates mentors and supervisors take use of standards in a relative consistent manner. Interaction between teachers and raters accounts for 27.51%, indicating mentors and supervisors rank ordered persons considerably differently. Also, interaction among teachers, standards, and raters and other undifferentiated error sources takes up to 45.85% of total variability. The generalizability coefficient that computed in the D study is .92, indicating a high reliability of the Intern Keys instrument and it is reliable for generalization.

The IRT-based MFRM has the same three facets as the G study. The rater effect cannot be attended to because of linking problem (59 subsets) as well. Measures for individual candidate teachers, raters’ role, and standards are computed and mapped on the same scale (Figure 2). Measures of raters’ role and standards are centered at 0, but measures of candidate teachers are allowed to float. The average measure for candidate teachers is 2.37 logits. Most candidate teachers are within the third rating category based on the variable map (i.e., received an average rating score of 3). The ratings from supervisors are a little lower than those of mentors, indicating supervisor group is a little more severe than mentor group in average. The differences between mentor and supervisors are relatively larger on standard 4 (Differentiated Instruction), 5 (Assessment Strategies), and 6 (Assessment Uses) than other standards (Figure 3). Among the 10 standards, Assessment Uses and Differentiated Instruction are given lower scores by raters, indicating that these are more difficult for candidate teachers to achieve. In contrast, professionalism is the easiest standard to meet.

For the 1-4 rating scale, category 3 has been mostly used by raters (Figure 4). The reliability for separation of standards is .96, showing standards spread out enough to measure the performance of candidate teachers.

Even though raters are not the focus of this study, the performance of raters is still our interest. Due to the linking problem, previous analyses couldn’t provide accurate report for rater effects. At least inter-rater agreement indices will give us a glimpse of raters’ performance. Exact agreement rates, adjacent agreement rates, Cohen’s kappa based on both exact and adjacent agreement rates are computed (Table 4). Adjacent agreement includes exact agreement cases and cases that absolute differences are within 1 rating category. The adjacent agreement rates are above 97% across 10 standards, indicating relatively good agreement among raters. After adjusting the chances, Cohen’s kappa coefficients of 6 standards (Standards 1, 2, 3, 6, 8, and 10), based on exact agreement, are smaller than .20. Based on the criterion (Landis & Koch, 1977), they are indicating a slight agreement. Cohen’s kappa coefficients of the remaining 4 standards (Standards 4, 5, 7, and 9) are between .20 and .30, showing fair agreement. Cohen’s kappas for 3 standards (Standards 1, 3, and 7), based on adjacent agreement, are above .70, which is the evidence for substantial agreement. Cohen’s kappa coefficients of standards 2, 4, and 10 are within .40 to .60, showing a moderate agreement. Standards 8 and 9 are within .20 and .40, indicating fair agreement. Standard 6 with Cohen’s kappa being .056 only has slight agreement. Standard 5 has a negative Cohen’s kappa, meaning the observed agreement is even less than the chance probability. For standard 5, there are 4 supervisors assigning 4 while mentors assigned 2 to the same candidate teachers.

Overall, the Intern Keys instrument has a high reliability. The results are consistent across different methods.

Additional activities and studies

The current GaPSC ~~guidelines for using the Intern Keys~~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. ~~at the beginning of the year, in the same way as practicing teachers are briefed on TKES at the beginning of the year.~~ For the current study, we ~~prepared~~ developed an instructional web video ~~for~~ to prepare Mentors and Supervisors for the implementation of this instrument as a summative assessment of candidate performance during the clinical practice experience. For next year, this could be re-worked and expanded to serve as guidance for ~~student~~ teacher candidates. This will have several positive effects.

First, it will be an opportunity to provide a statewide model of quality teaching across all of the EPPs. Second, it sends a clear message that use of the Intern Keys ~~will be used for student~~ can provide an educative ~~teaching evaluation~~foundation on performance expectations for beginning teachers as they enter the profession. Finally, it provides an opportunity for the EPPs to establish a language and protocol for examining teaching behaviors. Some of the EPP staff in the current study informed us that they were using the Intern Keys to provide a structure for discussions and feedback with their students.

At a minimum, w~~W~~e 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 ~~associated teacher preparation program~~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 ~~these training sessions~~preparation efforts with materials, speakers, video examples, and documents.

We would further recommend that EPPs hold orientation meetings on Intern Keys for any new ~~practice teaching staff~~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.

The data collection process for the current study asked the raters to choose the level of performance on a four point scale. This scale was used for the reliability assessments. We also collected qualitative responses from the raters when we asked what empirical data each rating was based on. As we move forward, we intend to apply qualitative data analysis techniques to these reports. This will enable us to evaluate the quality of the agreement findings. In our evaluation process, we will consider that pairs of ratings that match numerically and refer to the same empirical data to be the most reliable. Non-matching ratings (e.g., 1 and 3) may be the result of the two raters having observed different behaviors, and validly making different ratings. On the other hand, instances where raters cite the same empirical evidence but award differing ratings indicate that the rating system is not being applied in a consistent manner.

The current report was based on a volunteer group of EPP staff. It is recommended that n~~N~~ext year, 2015-2016, the instrument ~~will~~ be ~~in~~ use statewide~~, and will~~to provide additional data for validation. In 2016, all interns will have edTPA scores available, and this will serve as a solid criterion measure.

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 ~~they will be available~~ 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.

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

Table 2. Comparison of training methods

	t	df	Sig. (2-tailed)	Mean Difference*	Std. Error Difference
STD1	1.978	320	.049	.094	.047
STD2	1.439	320	.151	.081	.056
STD3	1.695	320	.091	.094	.055
STD4	2.427	246.624	.016	.146	.060
STD5	.357	320	.722	.018	.051
STD6	2.771	275.477	.006	.157	.057
STD7	1.064	289.822	.288	.066	.062
STD8	1.632	320	.104	.090	.055
STD9	2.619	299.752	.009	.134	.051
STD10	2.091	305.225	.037	.104	.050
Sum	2.500	320	.013	.983	.393

Note: *--This is video training only minus all other modes, i.e., raters who only watched the video tended to score *slightly* higher. The degrees of freedom that not even numbers are due to the violation of equal variance test. A correction has been applied for those cases (STD 4, 6, 7, 9, 10).

Table 3. Frequencies for candidate teachers' preparation

No.	Description	Frequency	Percent (%)
Q17_1	None	45	9.34
Q17_2	Candidate had seen a copy of the Intern Keys instrument	249	51.66
Q17_3	Candidate had seen a copy of the Teacher Keys instrument	168	34.85
Q17_4	Teacher Keys evaluation was integrated into the candidate's preparation program	199	41.29
Q17_5	Candidate had experience with the Teacher Keys electronic platform	36	7.47
Q17_6	Other	14	2.90
Q17_7	I don't know	75	15.56
Q18_1	None	72	14.94
Q18_2	Discussed the Teacher/Intern Keys standards with the candidate	330	68.46
Q18_3	Provided mid-point performance feedback based on the Teacher/Intern Keys standards	198	41.08
Q18_4	Reminded the candidate about the summative assessment using the Intern Keys at the end of student teaching experience	136	28.22
Q18_5	Went through the Teacher Keys electronic platform with the candidate	52	10.79
Q18_6	Shared Teacher Keys orientation materials with the candidate	102	21.16
Q18_7	Invited the candidate to attend school or district Teacher Keys trainings	30	6.22
Q18_8	Other training	14	2.90

Note: IK Prep - Intern Keys preparation. Q17: To the best of your knowledge, what kind of preparation had the teacher candidate received for the Intern Keys evaluation beginning her/his student teaching experiences? Q18: During the teacher candidate's experience with you, what kind of training did you provide to the candidate for the Intern Keys evaluation?

Table 4. G study results of variance components for each facet

Effect	Sum of squares	Mean squares	df	Variance components	Total variability (%)
Candidate teachers	272.64	1.61	169	0.040	15.79
Raters’ role	9.85	9.85	1	0.004	1.64
Standards*Raters	38.53	2.14	18	0.012	4.71
Candidate teachers*Raters	138.40	0.82	169	0.069	27.51
Candidate teachers*Standards*Raters, error	385.37	0.13	3042	0.127	50.35

Note: * represents interaction. df refers to degree of freedom.

Table 5. Inter-rater agreement indices

Standards	Exact agreement (%)	Adjacent* agreement (%)	Cohen's kappa (exact agreement)	Level of agreement	Cohen's kappa (adjacent agreement)	Level of agreement
1. Professional knowledge	72.94	99.41	.126	Slight	.604	Substantial
2. Instructional planning	66.47	98.82	.113	Slight	.551	Moderate
3. Instructional strategies	65.88	98.82	.172	Slight	.637	Substantial
4. Differentiated instructions	66.47	97.65	.232	Fair	.518	Moderate
5. Assessment strategies	76.47	97.06	.292	Fair	-.316	Less than chance
6. Assessment uses	67.65	97.06	.196	Slight	.056	Slight
7. Positive learning environment	58.82	98.82	.121	Slight	.720	Substantial
8. Academically challenging environment	65.88	97.06	.119	Slight	.222	Fair
9. Professionalism	66.47	98.82	.223	Fair	.357	Fair
10. Communication	66.47	98.82	.087	Slight	.493	Moderate

*-- Adjacent agreement includes exact agreement cases plus cases with absolute differences are within 1 category, e.g. 2 and 3.

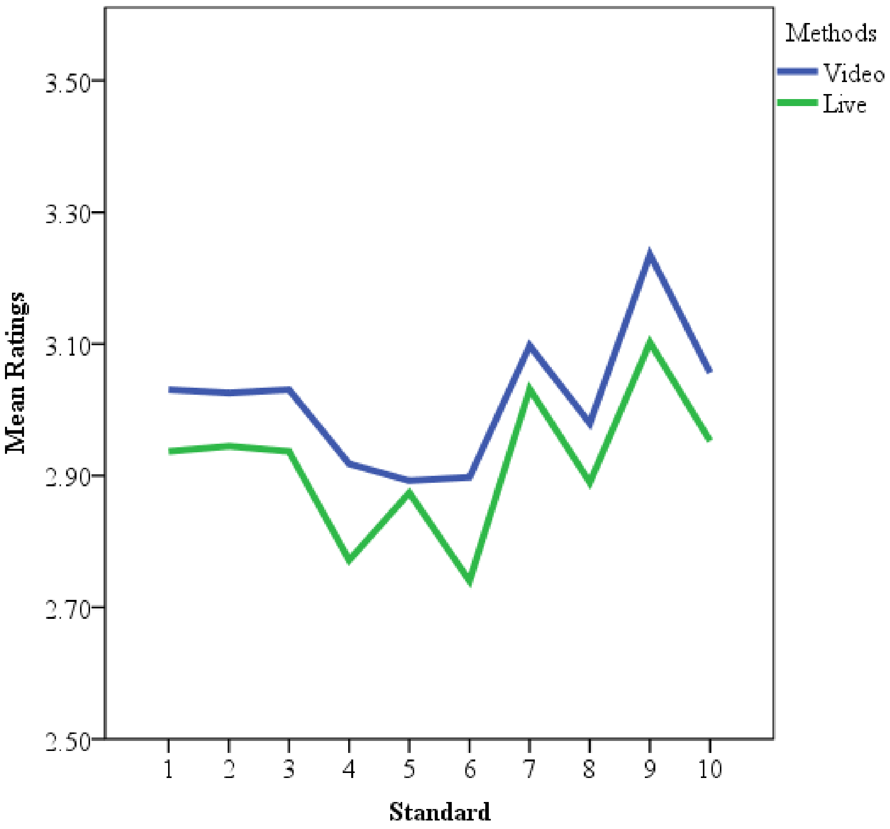
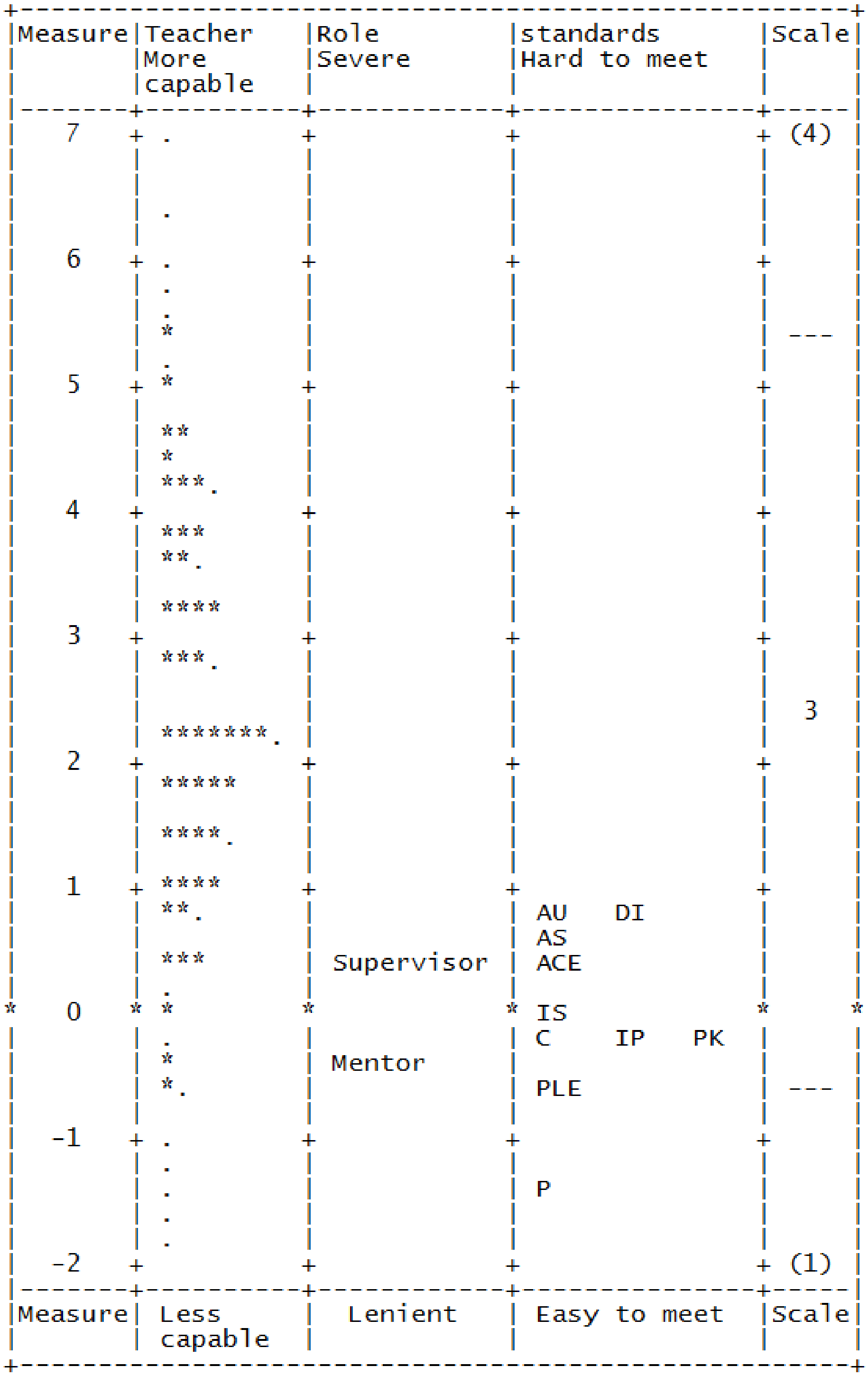


Figure 1. The average ratings across 10 standards for video method and live method(s).



```
+-----------------------------------------------------+
|Measure|Teacher   |Role        |standards      |Scale|
|       |More      |Severe      |Hard to meet   |     |
|       |capable   |            |               |     |
|-------+----------+------------+---------------+-----|
|   7   + .        +            +               + (4) |
|       |          |            |               |     |
|       |          |            |               |     |
|       | .        |            |               |     |
|       |          |            |               |     |
|   6   + .        +            +               +     |
|       | .        |            |               |     |
|       | .        |            |               |     |
|       | *        |            |               | --- |
|       | .        |            |               |     |
|   5   + *        +            +               +     |
|       |          |            |               |     |
|       | **       |            |               |     |
|       | *        |            |               |     |
|       | ***.     |            |               |     |
|   4   +          +            +               +     |
|       | ***      |            |               |     |
|       | **.      |            |               |     |
|       |          |            |               |     |
|       | ****     |            |               |     |
|   3   +          +            +               +     |
|       | ***.     |            |               |     |
|       |          |            |               |     |
|       |          |            |               |  3  |
|       | *******. |            |               |     |
|   2   +          +            +               +     |
|       | *****    |            |               |     |
|       |          |            |               |     |
|       | ****.    |            |               |     |
|       |          |            |               |     |
|   1   + ****     +            +               +     |
|       | **.      |            | AU   DI       |     |
|       |          |            | AS            |     |
|       | ***      | Supervisor | ACE           |     |
|       | .        |            |               |     |
*   0   * *        *            * IS            *     *
|       | .        |            | C    IP   PK  |     |
|       | *        | Mentor     |               |     |
|       | *.       |            | PLE           | --- |
|       |          |            |               |     |
|  -1   + .        +            +               +     |
|       | .        |            |               |     |
|       | .        |            | P             |     |
|       | .        |            |               |     |
|       | .        |            |               |     |
|  -2   +          +            +               + (1) |
|-------+----------+------------+---------------+-----|
|Measure| Less     |  Lenient   | Easy to meet  |Scale|
|       | capable  |            |               |     |
+-----------------------------------------------------+
```

Figure 2. Variable map of MFRM

Note: Professional knowledge (PK), instructional planning (IP), instructional strategies (IS), differentiated instructions (DI), assessment strategies (AS), assessment uses (AU), positive learning environment (PLE), academically challenging environment (ACE), professionalism (P), communication (C).

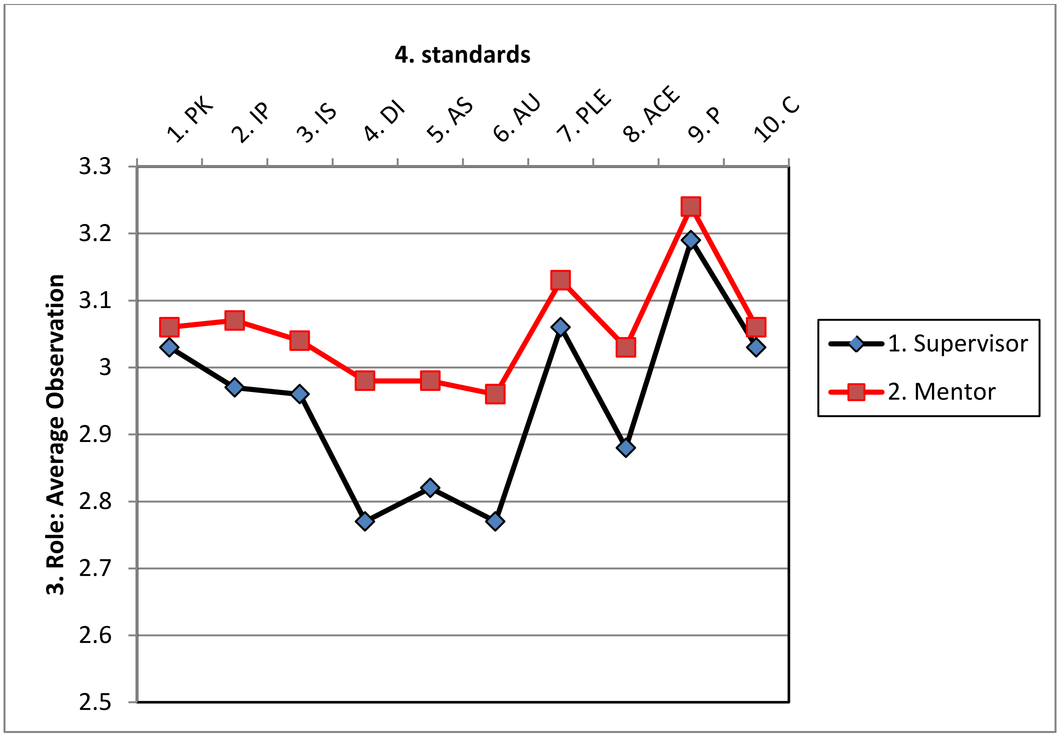


Figure 3. Interaction between raters' role and standards

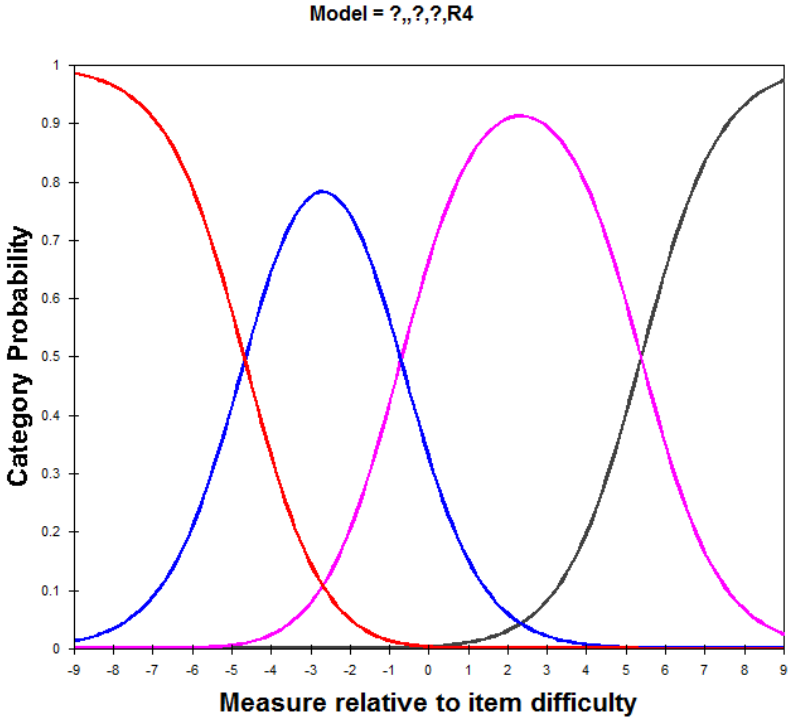


Figure 4. Category probability curve of MFRM/ **need labels of these curves /**

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*			