Reliability and Validity of the KTIP Assessment

Terry Hibpshman
Martin School
February 2017

Executive Summary

This paper is the result of a request for analysis of the reliability and validity of the new PGES[1]-aligned KTIP[2] IPR[3], first rolled out statewide during the 2015-2016 school year. Information about reliability and validity are necessary to assure that the instrument is an acceptably accurate measure, and that it, in fact, is principally a measure of intern performance. The IPR is currently used to determine whether interns will be recommended for professional certification. We also anticipate using the IPR in aggregate as a measure of provider performance in the KEPAS[4] accountability system.

Data for this study include the ratings of all interns collected in the IMS data system during the 2015-2016 school year. The sample includes 2331 K-12 interns, and 88 IECE interns. Analysis of IECE data were deferred pending the collection of additional data.

Analysis of the available data led to five findings:

Finding 1: *KTIP IPR composite committee ratings are sufficiently reliable to be used to determine which interns should be granted professional certification*. *Scoring was applied accurately and consistently, at least in the final cycle, as indicated by the generalizability study.*

Finding 2: *The scoring rules are appropriate. Some improvement in rater training might be indicated to assure that the rules are always appropriately applied*.

Finding 3: *Because of only limited support for the 4 domain model, we do not recommend use of domain scores for decision-making about interns or EPPs*.

Finding 4: *Although the new IPR is an improvement over the old procedure, scoring is still too lenient. Interns would not be harmed by spreading out the scores a bit. This could be done in rater training by reminding raters that precision in scoring in the long term is better for the intern*.

Finding 5: *The IPR is a reasonably good measure of intern performance. Because it is, it is appropriate for use both as a measure of eligibility for professional certification and of EPP program performance*.

In addition, we make a few recommendations for future data collection and improvements in rater training.

---

[1] PGES is the Professional Growth and Effectiveness teacher evaluation system of the Kentucky Department of Education.

[2] KTIP is the Kentucky Teacher Internship Program administered by EPSB.

[3] The IPR is the Internship Performance Record, a system developed to measure the performance of teacher interns.

[4] KEPAS is EPSB's accountability system for approved educator preparation programs.

Introduction

Since its first year of operation in 1985, The Kentucky Teacher Internship Program (KTIP) has been the gateway to professional teacher certification in the Commonwealth (Brockman, 2015). Originally intended as a support system for new teachers in their first year post-training, it was seen principally as a means of reducing the perceived high levels of attrition by early-career teachers[5], although it soon also became a means of assuring the quality of teachers newly admitted to teaching in Kentucky. This latter goal may not have been contemplated by the developers of KTIP, but it is clear from early EPSB Board minutes that it became important soon after the program was implemented. Board members, beginning in 1991, engaged in extensive discussion of the need for "teacher valued outcomes," i.e., capacities that new teachers would be expected to have at the point of preservice completion (Hibpshman, 2005). These capacities were codified as a set of "new teacher standards," which were incorporated, in the early 1990's, into the KTIP Internship Performance Record (IPR), a measurement system that required all teacher interns to fully meet the New Teacher Standards by the end of their internship. It is clear from a reading of the Board minutes during this period that the New Teacher Standards, and the IPR, were originally conceived as a means of measuring the performance of teacher preparation *programs*, although they also met the additional purpose of quality control of Kentucky teacher performance by denying professional certification to interns who could not fully meet the New Teacher Standards.[6]

This approach to the measurement of new teacher quality was never satisfactory, because the IPR, as originally conceived, was never an adequate measurement system. Because internship is a gateway to professional certification, and all interns must fully meet all the standards (i.e., achieve the maximum possible score on each standard) there was never much range in scores, especially for the final, consequential, rating.[7] As a result, it was impossible to ever establish validity or reliability for the IPR, since nearly everyone, in effect, received the same score. Yet it seems likely that new teachers, even among those whose performance is at least minimally satisfactory, should exhibit a range of performance. That we could not demonstrate this range of performance, even though we were sure on theoretical grounds that it existed, obviated the value of the IPR system as a measurement of either teacher quality or preparation program performance.

The persons who support the intern in his or her first year of teaching comprise a committee of three, the principal of the school to which the intern is assigned, a "resource teacher," and a "teacher educator" supplied by the educator preparation program (EPP) in that region. The resource teacher is an experienced teacher who, to the extent possible, practices at the same level (e.g., elementary) and in the same content area as the intern. Resource teachers are required to spend at least 60 hours with the intern over the course of the year, 20 in class and 40 outside of class. Resource teachers, principals, and teacher educators are required to undergo training in the measurement process and the process of supporting the intern (EPSB, 2015b). The committee members conduct an orientation meeting and three conferences with the intern during the

[5] Teacher attrition, especially among early career teachers, has been seen as a problem nationally (Ingersoll, 2002), and has generally been seen as a consequence of working conditions in the schools. KTIP was intended as a remedy for this problem, based on the belief that support in the first year of practice could shield new teachers from some of the shocks associated with the first year of teaching. Whether this belief was true or not, we found in studies of teacher attrition in 2001 that Kentucky has lower attrition rates for new teachers than have been widely reported in the literature.

[6] It is important to keep in mind that the internship program as originally developed was not conceived as a measurement of either new teacher or provider performance. Its function was to provide support to new teachers, and this continues to be its principal function today.

[7] Few interns (less than 1%) have ever failed their internships. As a result, even the percent passing rate could not be used as a measure of program performance.

year, and each member of the committee rates intern performance three times. The first two ratings are formative, the third is summative. If an intern receives a rating of "ineffective" for any component for the third cycle, the committee can opt for a fourth assessment cycle.[8] At the end of the internship year, a final joint rating is assembled by the committee. This final rating constitutes the score on which the intern's performance – and eligibility for professional certification – depends.

The numerous problems with the New Teacher Standards, and with the IPR, had been the subject of extensive internal discussion within EPSB for some time, starting about 2000. The Standards were not unidimensional, and were framed in philosophical terms that made it difficult to develop clear exemplars of adequate performance. Because nearly every intern received a perfect or near-perfect score at the end of the internship, the IPR could not be used for its original purpose, evaluating the performance of EPP's, and the reliability and validity of the IPR could never be established. EPSB staff for several years advocated the replacement of this system with a better, more sensitive approach to measurement of intern performance (Hibpshman, 2005).

This change took place in 2015. The Kentucky Department of Education (KDE) had recently developed a new teacher evaluation system (The Professional Growth and Effectiveness System – PGES) based on the Danielson Framework, and during discussion at a Board retreat in 2013 EPSB Board members questioned whether the measurement of intern performance ought to be consistent with the PGES (KACI, 2013a). There are numerous advantages to this approach: the intern would have his or her performance measured in the same terms as would be applied in subsequent years, and to the extent that the PGES were a reliable and valid measure, would make it possible to measure career progression from the beginning, on a single scale.[9] After considerable discussion, EPSB agreed to align the KTIP program with the PGES, and work began in fall 2013 for a 2014-2015 pilot of the new approach (Brockman, 2015). This resulted in a new IPR, nearly identical in form to the teacher evaluation used by PGES. We note, however, that the administration procedure of the IPR is much more complex than that of the PGES, as it is applied to every intern every year, and involves multiple raters and multiple ratings.

In the course of the year-long internship, new teachers assemble a body of "evidence" which serves, along with classroom observation, as the basis for the ratings by the committee. There are seven sources of evidence:

- Lesson plans
- Post-observation reflections
- Professional growth
- Records and communications
- Professional involvement
- Observations of teaching
- A student voice survey

[8] Our dataset includes results from cycle 4 interns, but the small number of such cases prevents us from including them in the analysis.

[9] *A fortiori*, it would fulfill the original intent of the IPR, to provide a measure of EPP performance.

EPSB provides templates on its website to assist interns in developing their sources of evidence. Templates identify the particular component of the Framework that is addressed by the evidence supplied by the intern. Some of the templates require that additional information, such as copies of teacher-created assessments, be supplied by the intern. Committees may also require additional information when needed. The first assessment period is from 1-60 instructional days after the orientation meeting; the second is 61-110 instructional days after the orientation meeting; and the third is from 111 instructional days after the orientation meeting to the end of the school year. The professional growth plan is particularly interesting because it contains a self-assessment by the intern on each of the components rated by the committee (EPSB, 2015a).

The KTIP assessment permits, for each component, a range of scores from 1 (ineffective) to 4 (exemplary). Unlike the old IPR, where interns had to demonstrate mastery of each of the standards to pass the internship, the PGES-aligned KTIP IPR requires only that the intern be rated as "developing" (the second-lowest rating) on each of the standards (Brockman, 2015). This is a fundamental and positive change in system philosophy, as it allows for a range of scores, both for interns who pass the internship and those who do not.

The purpose of the present report is to evaluate the metrical characteristics – reliability and validity – of this new, PGES-aligned, approach to intern performance measurement, to the extent that available data make it possible to do so. To this end, the report is divided into four sections. In section 1, we discuss the Danielson Framework and its relationship to the PGES and IPR. In section 2, we discuss reliability and validity issues relevant to a performance assessment such as the IPR. In section 3, we provide data analysis and reasoning relevant to evaluating the reliability and validity of the IPR. In section 4, we make recommendations for further studies necessary to the ongoing evaluation of this measurement procedure.

Section 1: The Danielson Framework

PGES, and by extension, the KTIP IPR, are based on Charlotte Danielson's Framework for Teaching (FFT). The FFT had its genesis in the development of Praxis III by the Educational Testing Service (ETS) from 1987 to the early 1990's. Praxis III was designed as a performance assessment system to complement Praxis I (measurement of candidate academic skills prior to admission to teacher training) and Praxis II (measurement of content and pedagogical knowledge at the completion of a preservice program) (Myford et al., 1993). The development team – of which Danielson was a member – engaged in an extensive and ingenious effort to define what it meant to be an effective teacher, incorporating literature reviews, job analyses, collaboration with practicing teachers, and consultation with experts in the field of teacher preparation and licensing, including state certification staff (Dwyer, 1998). Praxis III was intended for use as a licensing test: as such, it would be administered principally to persons undergoing their first year of teaching, before a professional certificate had been issued. Praxis III was adopted by a few state certification agencies, but never was widely used, probably because it constitutes an expensive and time-consuming assessment model. Although 39 states currently use the Praxis II tests for initial licensure or certification[10], only 6 have ever adopted the Praxis III. Some states, such as Kentucky, have had their own performance evaluation processes for new teachers.

Danielson used the research base created during the development of Praxis III as a springboard for the development of the FFT. An inspection of the domains and components for the Praxis III and the FFT makes it clear (see appendix Chart 1) that the two are closely related. The

[10] Some states require the Praxis II tests (including the PLT tests); some allow Praxis II tests as alternatives, among other permissible examinations.

principal difference between the two frameworks is that while the Praxis III contains 19 components, the FFT has 22. FFT domains and components are labeled in somewhat more elegant terms, and some Praxis III components have been either split or combined to create the FFT components. It is not clear from the promotional materials available from Teachscape or other sources how Danielson went about making these changes, or the basis for inclusion of some additional material to the original Praxis III model.

The FFT has been well-received in education nationally. It serves as the basis for teacher evaluation systems in numerous states and school districts (Milanowski, 2011). The Kentucky PGES, although it claims to be based on the Danielson Framework only in part, taking into account also the Kentucky Teacher Standards, the Kentucky Program of Studies, the Kentucky Core Academic Standards, and the KDE Characteristics of Highly Effective Teaching and Learning (KDE, 2014), as a formal measurement system is almost indistinguishable from the materials published by Danielson. The rollout of the PGES involved development of rubrics for scoring teacher performance against each of the FFT components, and training for evaluators (principally school principals) who would be expected to rate teachers using the new system. As developed, the PGES calls for ratings of experienced teachers every second year by a single rater.

Using the rubrics and other materials developed for the PGES, EPSB staff began the process in 2013 of adapting the PGES rating system for use by KTIP. Three possible adaptations of the PGES were contemplated: continuing to use the existing IPR during the internship, with the PGES serving as the final, consequential measurement; evaluating interns with both the existing IPR and the PGES; and adopting the PGES as the internship instrument, training internship committee members in its administration (Hibpshman, 2013). The latter option was chosen, principally because it would be much easier to administer than the other two.

Two groups were influential in the development of the PGES-aligned KTIP assessment, the KTIP/PGES work group, empaneled ad hoc for the purpose of planning the change, and the Kentucky Advisory Council for Internship (KACI), a standing committee of the EPSB. These groups, along with EPSB staff, began the work of aligning the KTIP assessment with the PGES late in 2013, completing their work in time for a 2014-2015 school year pilot. The pilot was conducted at 21 public school districts across the state, covering the range from large to small and including county and independent districts, as well as at one nonpublic school. Based on experience with the pilot, some adjustments were made to rubrics and to the training process. 184 interns were involved in the pilot project (Brockman, 2016).

Any discussion of reliability and validity of any instrument based on the FFT must begin with a discussion of the nature of Danielson's Framework. Although a number of investigators (Benjamin, 2002; Holtzapple, 2003; Kane et al., 2010; Milanowski, 2011; Keesler & Howe, 2012; Güerere, 2013; Ho & Kane, 2013; Lazareth & Newman, 2013; Hood et al., 2015; Johnson & Semmelroth, 2015; Lash et al., 2016; Roegman et al., 2016) have conducted reliability and validity studies on instruments based in whole or in part on the FFT, the Danielson Framework cannot, by its nature, be the subject of reliability or validity studies.[11] The FFT is not a test, or even strictly a measurement procedure, although as written it contains strong measurement components.[12] It would be more accurate to describe it as a theory about the nature of teaching effectiveness. Although Charlotte Danielson does not herself describe it as a theory, her own words make it clear that this is what she intended in its development:

[11] Because reliability and validity depend on the intended use of an instrument and the circumstances of its administration and scoring, topics we will discuss in section 2.

[12] Teachscape, Danielson's company, provides training and certification for evaluators using the FFT.

> The framework aims to describe all of teaching, in all its complexity. It is comprehensive, referring not only to what occurs in the classroom but also to what happens behind the scenes and beyond the classroom walls. The comprehensive nature of the framework for teaching sets it apart from other, earlier attempts to describe teaching. (Danielson 2007, location 464)

She claims that the FFT "derives as much as possible from sound educational research," although she acknowledges that in some areas, formal research has not yet been conducted, and those portions of the framework are based on the recommendations of experts. That some of the FFT is based on widely-accepted ideas in education that have yet to be proven was also noted by Hazi (Hazi, 2014), who commented for example that the idea that constructivist approaches to teaching are superior – a fundamental principle embraced by the FFT – has never been demonstrated by empirical research.

That the FFT is based substantially on research, but also substantially on accepted but unproven ideas about the nature of good teaching is not problematic if the FFT is viewed as a theory of effective teaching. Theories are always based on collections of such ideas, and are subject to later revision as additional research evidence accumulates. It can be argued that the various ideas on which the FFT are based constitute what Cronback and Meehl (Cronbach and Meehl, 1955, p. 290) described at the "nomological network" of theory necessary for the construct validation of a test.

It is measurements based on the FFT, not the FFT itself, that require reliability and validity studies. To the extent that the theory of effective teaching proposed by Danielson proves to be useful, then we would expect well-developed measurements based on the FFT to exhibit the reliability and validity that will in the end lend support to Danielson's Framework. How we go about demonstrating these things will be the subject of the next section.

Section 2: Reliability and Validity

Much has been written about reliability and validity that will not be recapitulated here. We will direct the interested reader to one of the many excellent textbooks on the subject (e.g., Nunnally & Bernstein, 1994) that have been published over the course of the past century. In general, reliability and validity can be framed as two simple ideas:

1. How precise are the scores produced by the measurement?
2. How well does the measurement serve as a proxy for the thing we want to measure?

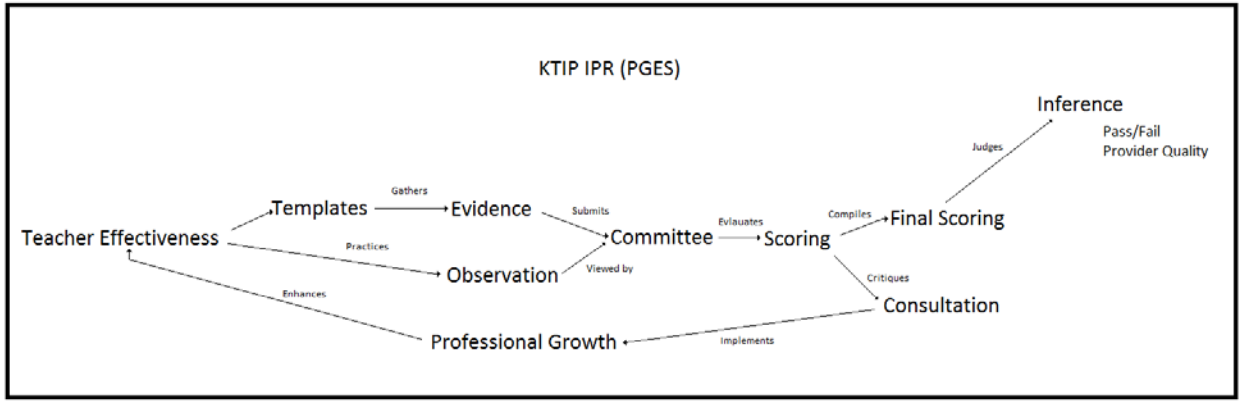
In terms of the current project, there is a theoretical construct which we wish to quantify, which me might call "teacher performance," or perhaps "teacher effectiveness." We wish to quantify the strength of this construct as it exists with respect to individual teachers because we believe based on our experience – either through research or personal observation – that this construct is related in some predictable way to a goal we wish to maximize, providing a quality education to public school children.[13] The construct is not directly observable, in part because it is

[13] We intentionally eschew "student achievement" as the sole criterion because the history of both the FFT and the PGES clearly do not take the view that it is the only outcome of interest in the evaluation of teaching or teacher quality.

complex and multidimensional, and in part because we are constrained by time and funding. We assume that if the construct exists, then it will influence, directly or indirectly, any exemplar[14] of the construct we might choose, so that a judicious choice of exemplars, and a carefully constructed procedure for rating the exemplars, will serve as a proxy measure of the extent to which particular teachers can be expected to be effective. Judicious choice of exemplars and careful construction of scoring procedures then constitute the basis for assuring that the measurement is sufficiently precise for the use for which it is intended, and has a strong enough relationship to the construct to assure that we can accurately sort individuals on the basis of the measurement. We desire to demonstrate that the measurement is both precise enough (reliable) and strongly enough related to the construct (valid) so that we can use it as a proxy for both the quality of individual teachers and the quality of the programs that prepare them.

Evaluation of the precision and validity of the IPR begins with an understanding of the process of assembling an IPR measurement, which is illustrated in Chart 2.

Chart 2
The KTIP IPR Process



At each stage of the process (directed lines in the diagram) some event occurs, and the process of evaluating the reliability and validity of the IPR depends on our ability to evaluate the effect of these events on the subsequent IPR score(s).

The process of conducting reliability and validity studies of performance assessments has been the subject of numerous articles in the psychometrics research literature for some time (Jaeger, 1993; Messick, 1994; Delandshere & Petrosky, 1998; Clauser et al., 1999; Barrett, 2001; Bachman, 2002; Stemler, 2004; Denner et al., 2008; Hill et al, 2012; Hill, Charalambous & Kraft, 2012; Ho & Kane, 2013; Lazarev & Newman, 2013; Johnson & Semmelroth, 2015; Lane & Stone, 2002; Lash et al., 2016), and our process mirrors what has been written by the various authors of this

[14] Such as, for example, observation of classes taught by a teacher, or the technical quality of assessments they create for their classes.

literature. For the sake of consistency, we follow here a recommendation for the evaluation of a performance instrument given by Bell et al. in 2012:

1. Scoring
    1.1. The scoring rule is appropriate.
    1.2. The scoring rule is applied accurately and consistently.
    1.3. The scoring is bias free.
    1.4. The data fit the scoring model.
2. Generalization
    2.1. The sample adequately represents the quality of all relevant lessons.
    2.2. Unexpected error is sufficiently accounted for.
3. Extrapolation
    3.1. The score on all lessons is related to the teaching quality teachers and students are able to enact.
    3.2. There are not systematic errors that undermine the extrapolation to teaching quality.
4. Implication
    4.1. The implications associated with teaching performance are appropriate.
    4.2. The properties of the observed scores on the lessons support the implications associated with the judgments of teaching performance. (p. 66)

This framework imposes requirements for how we should go about demonstrating each of its elements, as given in Table 1.

Table 1
Analysis Elements and Strategy

Element	**Analysis strategy**
1.1. The scoring rule is appropriate	• Analysis of scoring rule development • Analysis of the extent to which raters use the entire range of available score points
1.2. The scoring rule is applied accurately and consistently	• Inter-rater reliability studies • Generalizability analysis
1.3. The scoring is bias free	• Generalizability analysis
1.4. The data fit the scoring model	• Component/domain correlations

	• Confirmatory factor analysis
2.1. The sample adequately represents the quality of all relevant lessons	• Analysis of Templates
2.2. Unexpected error is sufficiently accounted for	• Generalizability analysis • Score distribution analysis
3.1. The score on all lessons is related to the teaching quality teachers and students are able to enact	• Inference
3.2. There are not systematic errors that undermine the extrapolation to teaching quality	• Generalizability analysis
4.1. The implications associated with teaching performance are appropriate	• Inference
4.2. The properties of the observed scores on the lessons support the implications associated with the judgments of teaching performance	• Inference • Confirmatory factor analysis • Component/domain correlations

An important distinction between reliability and validity analysis is that while reliability is principally a mathematically-demonstrable attribute of a measurement procedure, validity is much more subject to inference, and might rely on reasoning in addition to demonstration of mathematical features of the measurement and its relationship to other measures (Cizek, 2012). That is, validity relies on human judgement, and requires a defense on inferential grounds of various sources of evidence which may not be internal to the measurement procedure itself (Crooks, 1996).

With respect to reliability, our task in evaluating the KTIP IPR is governed by the fact that it is a performance assessment, and involves ratings by judges. Some traditional methods of demonstrating reliability, such as internal consistency, are not available to us for this type of measurement, as they are likely to produce inflated results. What is at issue for us is the extent to which the committee members who rate each intern are in agreement, and the extent to which the interns' scores are a function of intern performance rather than to differences among raters.[15] In technical terms, teacher quality is the "target domain" which we are attempting to measure, and the particular exemplars on which ratings are based come from a "universe of generalization" (Webb & Shavelson, 2005), all possible exemplars of teacher practice that would serve to indicate whether a teacher intern were effective. To the extent that our chosen

[15] That is, the extent to which the ratings are a function of some underlying measureable attribute of teachers.

exemplars fully represent the domain of generalization, and to the extent that rater judgements consistently represent the domain, then we can expect our measurement to be more reliable.

We note here that modern concepts of test development, especially generalizability theory, tend to blur the distinction between reliability and validity somewhat, because generalizability studies attempt to estimate the relationship between the sampled tasks used by the measurement procedure and the target domain, which can be seen as a content validity issue (Cronbach et al., 1972; Feldt & Brennan, 1983; Brennan, 1989). The relevant reliability issue with such studies is whether sufficient tasks of the right type are included in the measurement to provide stable estimates of the intern's capacity as a teacher (Chapelle, 2010). To the extent that this is true, then we should expect the measurement to be a stable indicator of the target domain.

For our purposes, reliability amounts to the extent to which the three raters of each intern's performance agree on the quality of the performance. Because this is a performance assessment and raters are involved, the appropriate measures of reliability include some measure of rater agreement, and an intraclass correlation drawn from a generalizability study. There are several possible measures of rater agreement. We use the percent agreement between raters for components, which has the advantage of simplicity. For domains and the total score, percent agreement is problematic because the data are additive rather than categorical; we will use Pearson product-moment correlations for those measures. The intraclass correlation will be drawn from an ANOVA of the intern X rater X cycle data available to us (Naizer, 1992), using Edug™, an application written specifically for this purpose (Cardinet et al., 2010). Each of the variables rater and cycle is known in generalizability theory as a "facet," a source of variance to which we are indifferent, but which might affect the score magnitude (Webb & Shavelson, 2005).

A number of factors might affect the reliability of the KTIP IPR, including the nature of the task, the specificity of the scoring rules, the level of training of the raters, and the conditions under which the scoring occurs (Clauser, 2000). None of these can be evaluated directly with the data available to us. What we can do is estimate the level of consistency among the three raters assigned to each intern. If the resulting estimate is unsatisfactory, then investigation of the particular cause and remedies of the problem will require additional study.

The nature of validity has been an ongoing argument in the psychometrics literature since the 1970's, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) notwithstanding. The accepted view, as described in the Standards, is that validation of a measurement involves a demonstration that its use supports inferences about the interpretation of resulting scores, and that it is the inferences that are validated, not the test itself (Chapelle, 2010; Cizek, 2012). A given measurement procedure would then have to be validated for each different inference to be drawn from it. In this view, a test developer makes one or more claims about the meaning of test scores (Zarębski, 2009; Chapelle, 2010), which are then supported by evidence of various types. The interpretive arguments that serve to validate a test cannot be verified in any absolute sense, and must be accepted or rejected based on their plausibility (Kane, 1992, 1999).

Not all psychometricians accept this view. An alternative point of view suggests that a test is valid if:

> (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure (Borsboom et al., 2004, p. 1061).

Regardless of the disagreement about exactly what validity means, almost all modern authors in the psychometrics literature agree that all validity is construct validity (Scriven, 1987). Particular approaches to validation of tests – such as analysis of the fit of its content to the domain it purports to measure – are in this view just approaches to marshaling evidence to demonstrate that the test score is an exemplar of some construct of interest to the developer.

In the case of the KTIP IPR, the construct of interest is teacher performance. By this we mean a relatively enduring attribute of persons that will cause them to be more or less effective in fulfilling the role of classroom teacher. Some persons, for reasons of ability, training, or motivation, will exhibit more of this attribute than others, and we expect that persons with more of the attribute will be of greater value to the educational enterprise than will those with less. Although student achievement is an important consequence of teachers' levels of performance, it is not the only consequence of interest. To the extent that we can measure this attribute, we would expect teachers who score higher on our assessment to be better at promoting student achievement, but also at such things as promoting a healthy school climate, working with parents and other professionals, managing student conduct, and the like. If our measurement is valid, it will serve both as a measure of the quality of individual teachers, and of the EPPs that prepare them.

Section 3 will discuss our efforts to demonstrate the validity of the KTIP IPR, as well as our efforts to demonstrate its reliability.

Section 3: Analysis of the reliability and validity of the KTIP IPR

The data

The new, PGES-aligned KTIP IPR was piloted in the 2014-2015 school year, and was implemented as a statewide system in the 2015-2016 school year. Our analysis is based on a dataset containing the IPR records of interns for this first statewide internship year. Our dataset consists of 2410 internship records for which there are complete data for all components (these comprise all interns who completed the internship that year), as well as admissions and exit records from the EPP programs where they completed their preservice programs. There are two types of interns in the dataset: K-12 interns, and interdisciplinary early childhood (IECE) interns. There are 2331 of the former, and 88 of the latter. Our analysis is based only on the K-12 interns, as we felt that the number of IECE internships was too small to produce stable results, given the methods we would be using. We note that the EPP preparation records available to us are only for persons trained by Kentucky-approved programs: about 200 of the interns lack these data. In addition to the IPR and admissions/exit records, a wealth of information about the development and administration processes of the IPR was available to us. We also have the certification test scores of the interns from the certification data system.

Description of the sample

Of the interns in the sample for whom demographic data are available, 76% were female, 24% male. 94% were White not-Hispanic, the other six percent distributed across other ethnicities. 85% were placed in public schools. 166 of the Commonwealth's 174 public school

districts had at least one intern, and two of them – Jefferson and Fayette County – accounted for 23% of all interns. 14 public districts had only one intern. 50% of the sample were placed in elementary schools, 27% in high schools, 19% in middle schools, 4% in schools with mixed grade levels, and less than 1 % in preschools, including Head Start centers. Interns were evenly divided between rural and urban districts. Two thirds were prepared at state-operated institutions of higher education. One independent institution of higher education (IHE) accounted for about 10% of the interns.

Appendix Table 2 shows the distribution of content areas for which persons in the sample were prepared. Note that these are not unduplicated: many persons in teacher preparation programs are simultaneously enrolled in more than one program. 86% of the program enrollments were in traditional programs, 14% in alternative programs. A few of the interns were listed as enrolled in "advanced" programs, due to a reporting problem in the admissions and exit data system.

Data analysis

Our reliability results begin with an evaluation of the relationships among the 22 components that make up the KTIP IPR. Table 3 shows product-moment correlations between each pair of components for the final committee assessment. All of the components are strongly correlated, with the minimum correlation at .44 and the maximum at .71. We found that the strength of the correlations among items grew with each succeeding cycle, so that the weakest correlations were found in cycle 1. Table 4 shows intercorrelations of the components for the resource teacher assessment in cycle 1, for contrast with the intercorrelations in the final composite measurement. Similar data tables are available for all of the measurement cycles. Although the correlations are somewhat weaker in the first cycle, they are nonetheless substantial, with a minimum of .32 and a maximum of .67.

Table 5 shows the correlations between raters for domain scores – the sum of component scores in each domain. As with component scores, the strength of the correlations increases markedly from cycle 1 to cycle 3. Even in cycle 3, the correlations between raters are weak, showing limited agreement, and are not evidence of adequate inter-rater reliability for domain sums. The correlation between resource teacher and principal ratings and the committee composite score, however, indicates that the domain scores for the composite is a much stronger measure than any of the rater domain scores. The correlation between the teacher educator and committee composite domain scores is somewhat weaker.

An issue that arose in the course of the analysis was whether it made sense to calculate a single, total score for the KTIP IPR. The FFT and all measures that derive from it assume four relatively disjoint domains of teacher practice, and if these hold up on analysis, a total score for the entire assessment would make little sense. To test whether IPR scores recapitulate this model, we conducted two analyses.

Table 6 gives correlations between each of the components and each of the domain totals, for the final committee rating. All components correlated substantially with all of the domain totals, although the components of each domain correlated more strongly to the domain to which they belonged. To make this clear, we offer Table 7, which shows the median correlation between components and domains. The entries on the diagonal are of considerably greater magnitude than the off-diagonal entries, although the relationships with the off-diagonal entries are quite strong. As with the component intercorrelations, we found that the strength of relationship between

components and domains increased with each additional cycle, and reached its peak with the cycle 3 committee assessment. The stronger correlations on the diagonal lends at least weak support for the domain structure of the assessment.

When we conducted a confirmatory factor analysis, however, as shown in Table 8, it became clear that the domain structure of the assessment is at best weakly supported. The factor analysis produced a first principal component that explained 88% of the variance among IPR components, and the factor loading of each component on the first factor was quite large. This is hardly surprising given the strong correlations among the 22 components, but it indicates – along with the weak rater reliabilities for domain scores – that we would not be on firm ground trusting domain scores and would be better advised to emphasize the total score of all components as a measure of teacher quality.[16]

It is expectable that scores would increase over time on the IPR, given that the first two assessments are formative, and are expected to engage the intern in a developmental process leading to improvement in performance.[17] Table 9 shows the progression of mean domain scores by rater for each cycle. By way of comparison, note that the minimum-possible domain score for domains 1 and 4 is 12, and the minimum for domains 2 and 3 is 10; while the maximum for domains 1 and 4 is 30 and the maximum for domains 2 and 3 is 25. The mean domain scores in every case are not far below the maximum, an issue will we take up at some length later. Table 10 shows similar results for total scores. The minimum possible total score is 44 and the maximum is 110.

Comparisons of total scores show the strongest evidence for reliability of the IPR. Table 11 shows correlations of total scores for each of the ten IPR assessments, and Table 11 isolates the correlations between raters for each cycle. Correlations are modest at best in the first cycle, but become considerably stronger by the second cycle, and reach acceptable levels by the third. Especially encouraging are the correlations between raters and the committee score for the final cycle, where the results show impressive levels of agreement.[18]

In addition to correlational measures of agreement, we calculated the percent agreement for each pair of raters for each component, as shown in Tables 13 – 15. These were calculated by counting the number of times in which each pair of raters were in exact agreement, then dividing by the number of ratings.[19] As with all other measures, the magnitude of the agreement between raters grows with each cycle. Agreement in the third cycle is acceptable.

It is always good practice to plot and inspect the distribution of scores. The result was Charts 3-6, which show the distribution of domain and total scores for the three cycles. Domain score distributions are initially bimodal, but by cycle 2 become approximately normal. The total score distributions appear in cycle 1 to be multimodal, but by cycle 2 become approximately normal. It is obvious from inspecting the charts that after the first cycle, kurtosis becomes a problem. Kurtosis is a measure of the "peakedness" of a distribution. If the scores of a distribution tend to cluster around a single point, the kurtosis will be relatively high. Tables 16 and 17 show the kurtosis for the domain scores and total scores respectively. Kurtosis for a standard normal distribution would be 0 – these kurtosis values are quite above that.

[16] A related issue that we are not prepared to evaluate at present, given that all of the components seem to measure a single factor, is whether it might be advisable to reduce the number of components in the IPR. This would require a generalizability D-study, which we lack the data to do at the moment. It would also require a change in system philosophy, since the IPR as currently constructed is based solely on the FFT.

[17] i.e., increases in scores over the course of the internship are evidence of the validity of the measurement process.

[18] The high level of agreement might be due in part to the pressure to assure that interns meet a minimum standard, and the tendency of raters to rate generously.

[19] It would not be useful to use this methodology for either domain scores or totals, since the data are not categorical, and any result would be misleading.

Skewness, the extent to which scores accumulate to the right or left of the center of the distribution, does not seem to be a problem with either the domain or total scores.

We conducted a generalizability study using Edug™ (Cardinet et al., 2010), to determine how well total scores could be generalized to the target domain, and to obtain estimates of the proportion of variance due to the three factors of interest, interns, raters, and cycles. This analysis produced Table 18. This analysis used data for all raters for all three cycles, except for the committee composite rating. The generalizability coefficient of .87 is acceptable, and indicates that the ratings are probably good indicators of the target domain. The proportion of variance attributable to differences in interns was 37%, with 23% due to cycles and 16% due to raters.

This proportion of intern variance is weak, and it occurred to us to ask what would happen if just the second and third cycles were subjected to a generalizability analysis. The result was Table 19. Here, 75% of the variance is attributable to differences between interns, a more than acceptable value. We also calculated an intraclass correlation from the first generalizability analysis of .77. This is a bit weak, but acceptable, and we note that it is based on all three cycles and all raters.

We computed correlations between IPR total scores and Praxis and other tests administered prior to certification. These produced a significant result only for the *Mathematics: Concepts and Problems* examination. This parallels what has been found in other studies of instruments based on the FFT. We did find (Table 20) small but significant correlations between the IPR committee total rating and interns' GPA at admission into teacher preparation programs.

Because it is probable that we will eventually use the results of IPR assessments to evaluate the performance of EPPs, we developed a chart (Chart 7) of the mean IPR total score for each of the providers from which the interns had completed. The chart is generally consistent with impressions of EPSB staff of the strength of the various Kentucky-approved EPPs.

In order to evaluate the fit between the assessment model and the KTIP IPR procedures, we conducted a qualitative analysis of documents relating to the process of developing the new method, administration procedures, and rater training. We also examined information about the development of the KDE PGES. The development of the IPR did not involve construction of new elements or rubrics for rating intern responses: these were adopted without substantial amendment from the PGES. Most of the development effort went into aligning the PGES with the existing EPSB Kentucky Teacher Standards, in developing the Sources of Evidence used in ratings, and in developing the training program for raters.

The Sources of Evidence Templates are detailed and demanding, requiring the intern to make a substantial effort to assemble characteristic exemplars of their practices in each of the areas rated by the IPR. The templates provide information to both the intern and rater of the IPR component(s) related to the information in the template. Comments collected from interns and from internship committee members made it clear that using the templates required a lot of detailed work and careful thinking about what to place there and the implications for assessment.

The training process spent more time on administrative procedures than any other subject. This is not surprising given that this was a first rollout, but may have detracted somewhat from more substantive issues. The training does include some excellent examples of the different levels of the classroom observation component, but examples for the other components were somewhat weaker.

Analysis

It is clear from the various measures of rater agreement that the IPR is sufficiently reliable for its intended use, identifying which interns should be granted professional certification. We would want the intraclass correlation from the generalizability study to be a bit higher, but the obtained result of .77 is at least minimally sufficient. Given that this was based on the entire set of ratings, including all three raters and all three cycles, it is probable that this value would be considerably higher if we were to confine ourselves to the composite committee rating in cycle 3. Happily, that is what we want to do: the first two cycles are formative only, and the consequential rating is only the committee rating. So we state our first finding:

Finding 1: *KTIP IPR composite committee ratings are sufficiently reliable to be used to determine which interns should be granted professional certification*. *Scoring was applied accurately and consistently, at least in the final cycle, as indicated by the generalizability study.*

We are concerned by the much lower values for measures of reliability in earlier cycles, especially the first. There is no reason, in theory or practice, why raters should be less precise in identifying effective teaching practice in the first two cycles than they are in the third. Ratings are based on observations and standardized documents collected from interns, and practice of any given quality should be as amenable to measurement early in the process as later. It is possible that the weaker performance of the raters in the early cycles might be due to the fact that this is the first year of the system, requiring some adjustment by the raters early in the process. Although a problem of this type is probably unavoidable, it should be subject to amelioration by a sufficiently rigorous training program.

There is no doubt that the scoring rule is appropriate. The IPR is based on the FFT via the PGES, both of which have a long history with the identical rubrics. Given the high quality of the templates and the fact that they are carefully anchored in the components, there is no reason to believe that the scoring rules do not, in fact, capitulate the FFT model on which the IPR is based. This brings us to a second finding:

Finding 2: *The scoring rules are appropriate. Some improvement in rater training might be indicated to assure that the rules are always appropriately applied*.

The FFT is organized into a set of four domains, each of which has either 5 or 6 components. We were able to adduce some weak evidence for this pattern in the data available to us, but the evidence was not strong enough for us to be comfortable using domain scores in any meaningful way. The factor analysis, and the much stronger reliabilities for the total score than for domain scores, make it clear that a single, total score for the IPR is a more meaningful measure. This suggests a third finding:

Finding 3: *Because of only limited support for the 4 domain model, we do not recommend use of domain scores for decision-making about interns or EPPs*.

Analysis of score distributions is problematic. Cycle 1 scores follow a different pattern than do the latter two cycles, and for both domain scores and the total score, kurtosis is a serious problem. The cause of this problem is apparent if we consider the score that predominates in the total score distributions in the second and third cycles: 88. This is the score an intern would receive if she or he scored at the next-to-highest category on each of the 22 components.[20] We believe that this results from lenient scoring by raters, who are reluctant to give a maximum score in most cases, but will err on the side of generosity if given a choice. It seems reasonable to suppose – given that all interns must score at least the next-to-lowest level on each component to be recommended for professional certification – that there are at least some who are minimally capable who are being scored too generously. This suggests a fourth finding:

Finding 4: *Although the new IPR is an improvement over the old procedure, scoring is still too lenient. Interns would not be harmed by spreading out the scores a bit. This could be done in rater training by reminding raters that precision in scoring is in the long term better for the intern*.

We lack the capacity at present to make comparisons between interns' IPR scores and student achievement, but we note that almost all studies of the FFT in the past have found a relationship between FFT scores and student outcomes (Noell et al, 2014). We should make every effort to conduct such studies in the future, but given the high fidelity of the PGES and IPR to the FFT model, and the FFT's excellent development history, we are relatively safe in assuming that the IPR is, indeed, a pretty good measure of teacher performance. Because it is likely to be so, we give our fifth finding:

Finding 5: *The IPR is a reasonably good measure of intern performance. Because it is, it is appropriate for use both as a measure of eligibility for professional certification and of EPP program performance*.

We have some weak evidence for the quality of the IPR in our analysis of its relationship to admissions GPA and the relationship of mean scores to EPP quality. This is very weak evidence and will require future analysis.

Section 4: Recommendations

A number of issues arose during the current investigation. Although the IPR as it was implemented in this first rollout has proven to be reliable and valid for its intended use, some improvements are indicated, both for the purposes of further reliability and validity studies, and for improvement of the management of the system.

[20] The scoring of components on the protocol used by raters ranges from 1-4, but the data file records these as 2-5. Had the scoring and data been completely consistent, then the typical score would have been 66. This score transformation creates a bit of confusion, but does not affect the analysis.

First, it would be very helpful to collect additional data. The data we have at present consist of ratings by intern committees on the 22 components, and what can be gleaned from EPSB's existing data systems. The system actually produces much more data than this, and it would be helpful to capture these data for further use. Sources of data that would prove helpful to future studies include:

- Some measure of rater performance during training. This would probably involve collecting data that do not exist, but would be useful for identifying training problems and sources of systematic variability in rater performance.
- A sample of intern Sources of Evidence templates would be helpful in evaluating the quality of the rating process.
- Interns produce self-ratings at the end of each cycle. These would be useful in validity studies.
- Each cycle involves ratings by the interns' students. These would also be helpful in validity studies.

In addition to additional data collection a few other recommendations are in order:

- A bit more detail, including characteristic examples, for rating non-observational components would be helpful.
- It would be helpful to stress to raters the importance of precision in rating.

Appendix
Tables and Charts

Chart 1
Comparison of Praxis III and Danielson Framework Domains

Praxis III Domains and Components	Danielson Domains and Components
Domain A: Organizing Content Knowledge for Student Learning A1: Becoming familiar with relevant aspects of students' background knowledge and experiences A2: Articulating clear learning goals for the lesson that are appropriate for the students A3: Demonstrating an understanding of the connections between the content that was learned previously, the current content, and the content that remains to be learned in the future A4: Creating or selecting teaching methods, learning activities, and instructional materials or other resources that are appropriate for the students and that are aligned with the goals of the lesson A5: Creating or selecting evaluation strategies that are appropriate for the students and that are aligned with the goals of the lesson	Domain 1: Planning and Preparation Component 1a: Demonstrating Knowledge of Content and Pedagogy Component 1b: Demonstrating Knowledge of Students Component 1c: Setting Instructional Outcomes Component 1d: Demonstrating Knowledge of Resources Component 1e: Designing Coherent Instruction Component 1f: Designing Student Assessments
Domain B: Creating an Environment for Student Learning B1: Creating a climate that promotes fairness B2: Establishing and maintaining rapport with students B3: Communicating challenging learning expectations to each student B4: Establishing and maintaining consistent standards of classroom behavior B5: Making the physical environment as safe and conducive to learning as possible	Domain 2: The Classroom Environment Component 2a: Creating an Environment of Respect and Rapport Component 2b: Establishing a Culture for Learning Component 2c: Managing Classroom Procedures Component 2d: Managing Student Behavior Component 2e: Organizing Physical Space
Domain C: Teaching for Student Learning C1: Making learning goals and instructional procedures clear to students C2: Making content comprehensible to students C3: Encouraging students to extend their thinking C4: Monitoring students' understanding of content through a variety of means, providing feedback to students to assist learning, and adjusting learning activities as the situation demands C5: Using instructional time effectively	Domain 3: Instruction Component 3a: Communicating with Students Component 3b: Using Questioning and Discussion Techniques Component 3c: Engaging Students in Learning Component 3d: Using Assessment in Instruction Component 3e: Demonstrating Flexibility and Responsiveness
Domain D: Teacher Professionalism D1: Reflecting on the extent to which the learning goals were met D2: Demonstrating a sense of efficacy D3: Building professional relationships with colleagues to share teaching insights and to coordinate learning activities for students D4: Communicating with parents or guardians about student learning	Domain 4: Professional Responsibilities Component 4a: Reflecting on Teaching Component 4b: Maintaining Accurate Records Component 4c: Communicating with Families Component 4d: Participating in a Professional Community Component 4e: Growing and Developing Professionally Component 4f: Showing Professionalism
Source: Dwyer C (1998). Psychometrics of Praxis III: Classroom Performance Assessments. Journal of Personnel Evaluation in Education 12(2): 163–187, 1998.	Source: Danielson, C (2007). Enhancing Professional Practice: A Framework for Teaching 2nd edition. Association for Supervision & Curriculum Development. Kindle Edition.

Table 2
Intern Content areas

PROGRAM_NAME	N	pct
Agriculture	17	0.74
Art	45	1.96
Biological Science	91	3.97
Business and Marketing Education	46	2.00
Chemistry	39	1.70
Chinese	6	0.26
Communication Disorders	26	1.13
Director of Special Education	1	0.04
Earth Science	6	0.26
Elementary	1166	50.81
Engineering and Technology Education	8	0.35
English	263	11.46
English as a Second Language	33	1.44
Family and Consumer Science	19	0.83
French	19	0.83
Gifted Education (Grades P-12)	4	0.17
Health	73	3.18
Hearing Impaired	7	0.31
Instructional Computer Technology	7	0.31
Instrumental Music	6	0.26
Integrated Music	77	3.36
Interdisciplinary Early Childhood Education	31	1.35
Learning and Behavior Disorders	787	34.29
Literacy Specialist	33	1.44
Mathematics	145	6.32
Middle School English	208	9.06
Middle School Mathematics	242	10.54
Middle School Science	170	7.41
Middle School Social Studies	197	8.58
Moderate and Severe Disabilities	87	3.79
Physical Education	100	4.36
Physics	17	0.74
Planned Program for Rank I	7	0.31
Reading	4	0.17
School Guidance Counselor P-12	18	0.78
School Media Librarian	6	0.26
School Principal (Grades P-12)	2	0.09
Social Studies	232	10.11
Spanish	66	2.88
Teacher Leader	61	2.66
Teacher Leader (Special Ed)	2	0.09
Theater	1	0.04
Visually Impaired	8	0.35
Vocal Music	5	0.22

Table 3

Component Intercorrelations

Final Committee Ratings

	1B	1C	1D	1E	1F	2A	2B	2C	2D	2E	3A	3B	3C	3D	3E	4A	4B	4C	4D	4E	4F	Total score
1A	0.58	0.61	0.61	0.62	0.53	0.49	0.55	0.51	0.48	0.49	0.58	0.53	0.59	0.54	0.58	0.57	0.52	0.45	0.52	0.55	0.52	0.75
1B		0.59	0.57	0.55	0.50	0.59	0.60	0.55	0.53	0.53	0.62	0.51	0.58	0.53	0.62	0.58	0.55	0.56	0.52	0.56	0.53	0.76
1C			0.63	0.68	0.60	0.51	0.58	0.53	0.50	0.51	0.59	0.59	0.59	0.61	0.59	0.60	0.55	0.51	0.50	0.55	0.50	0.78
1D				0.61	0.51	0.50	0.56	0.51	0.47	0.52	0.55	0.53	0.60	0.50	0.59	0.59	0.56	0.56	0.58	0.59	0.55	0.76
1E					0.61	0.53	0.58	0.57	0.54	0.50	0.57	0.58	0.65	0.58	0.60	0.60	0.54	0.52	0.51	0.55	0.52	0.78
1F						0.47	0.52	0.50	0.49	0.44	0.53	0.59	0.56	0.71	0.53	0.54	0.51	0.50	0.45	0.52	0.46	0.73
2A							0.70	0.64	0.64	0.55	0.61	0.48	0.57	0.49	0.59	0.51	0.50	0.51	0.50	0.53	0.54	0.75
2B								0.64	0.65	0.55	0.65	0.56	0.64	0.54	0.60	0.57	0.53	0.53	0.53	0.56	0.56	0.80
2C									0.69	0.56	0.60	0.51	0.60	0.53	0.55	0.50	0.50	0.49	0.47	0.53	0.51	0.76
2D										0.51	0.59	0.52	0.59	0.50	0.55	0.49	0.45	0.50	0.45	0.51	0.48	0.73
2E											0.57	0.46	0.54	0.47	0.59	0.53	0.54	0.50	0.50	0.52	0.54	0.71
3A												0.56	0.63	0.56	0.64	0.58	0.55	0.53	0.54	0.58	0.54	0.79
3B													0.61	0.63	0.53	0.52	0.47	0.48	0.45	0.54	0.44	0.73
3C														0.61	0.62	0.55	0.52	0.50	0.51	0.57	0.51	0.79
3D															0.56	0.53	0.51	0.50	0.46	0.53	0.46	0.74
3E																0.61	0.58	0.57	0.55	0.60	0.58	0.79
4A																	0.58	0.58	0.58	0.64	0.57	0.77
4B																		0.65	0.57	0.60	0.59	0.74
4C																			0.59	0.61	0.58	0.73
4D																				0.68	0.67	0.73
4E																					0.66	0.78
4F																						0.74

Minimum = .44

Maximum = .71

Table 4

Component Intercorrelations

Cycle 1 Resource Teacher

	1B	1C	1D	1E	1F	2A	2B	2C	2D	2E	3A	3B	3C	3D	3E	4A	4B	4C	4D	4E	4F	Total score
1A	0.50	0.52	0.50	0.54	0.47	0.39	0.46	0.40	0.38	0.42	0.49	0.40	0.46	0.45	0.49	0.45	0.41	0.34	0.39	0.40	0.40	0.68
1B		0.51	0.47	0.50	0.47	0.47	0.50	0.45	0.44	0.43	0.49	0.40	0.49	0.45	0.52	0.44	0.41	0.43	0.38	0.40	0.39	0.70
1C			0.49	0.59	0.52	0.40	0.48	0.45	0.40	0.41	0.49	0.46	0.50	0.47	0.48	0.43	0.41	0.42	0.38	0.41	0.37	0.70
1D				0.47	0.42	0.36	0.40	0.35	0.35	0.42	0.41	0.38	0.43	0.39	0.42	0.40	0.42	0.39	0.38	0.40	0.38	0.64
1E					0.50	0.43	0.48	0.45	0.44	0.44	0.49	0.46	0.53	0.49	0.51	0.46	0.42	0.39	0.37	0.41	0.39	0.72
1F						0.38	0.43	0.41	0.41	0.38	0.42	0.42	0.46	0.61	0.44	0.42	0.41	0.39	0.35	0.39	0.36	0.67
2A							0.67	0.57	0.61	0.50	0.55	0.39	0.50	0.38	0.53	0.38	0.35	0.36	0.35	0.36	0.41	0.69
2B								0.59	0.61	0.50	0.59	0.46	0.58	0.44	0.56	0.43	0.39	0.38	0.36	0.38	0.40	0.74
2C									0.67	0.51	0.49	0.41	0.50	0.43	0.49	0.36	0.34	0.34	0.33	0.35	0.35	0.69
2D										0.49	0.47	0.41	0.51	0.43	0.48	0.38	0.35	0.36	0.33	0.36	0.36	0.69
2E											0.49	0.37	0.43	0.39	0.48	0.38	0.42	0.40	0.39	0.39	0.43	0.67
3A												0.45	0.53	0.44	0.54	0.43	0.40	0.38	0.35	0.39	0.39	0.71
3B													0.51	0.47	0.45	0.36	0.34	0.36	0.32	0.34	0.33	0.63
3C														0.49	0.55	0.40	0.37	0.39	0.34	0.35	0.34	0.71
3D															0.49	0.42	0.42	0.38	0.35	0.38	0.37	0.67
3E																0.46	0.40	0.39	0.38	0.40	0.41	0.73
4A																	0.49	0.46	0.45	0.51	0.49	0.67
4B																		0.60	0.51	0.51	0.54	0.66
4C																			0.54	0.53	0.49	0.65
4D																				0.64	0.62	0.63
4E																					0.59	0.66
4F																						0.65

Minimum = .32

Maximum = .67

Table 5

Domain Intercorrelations

	domain 1				domain 2				domain 3				domain 4			
cycle 1	rt	pr	te		rt	pr	te		rt	pr	te		rt	pr	te	
rt	1.00	0.54	0.39		1.00	0.55	0.44		1.00	0.53	0.41		1.00	0.56	0.42	
pr		1.00	0.44			1.00	0.47			1.00	0.43			1.00	0.45	
te			1.00				1.00				1.00				1.00	
	domain 1				domain 2				domain 3				domain 4			
cycle 2	rt	pr	te		rt	pr	te		rt	pr	te		rt	pr	te	
rt	1.00	0.61	0.48		1.00	0.62	0.51		1.00	0.61	0.50		1.00	0.62	0.50	
pr		1.00	0.49			1.00	0.53			1.00	0.52			1.00	0.50	
te			1.00				1.00				1.00				1.00	
	domain 1				domain 2				domain 3				domain 4			
cycle 3	rt	pr	te	cm	rt	pr	te	cm	rt	pr	te	cm	rt	pr	te	cm
rt	1.00	0.69	0.55	0.80	1.00	0.69	0.58	0.80	1.00	0.69	0.58	0.81	1.00	0.69	0.56	0.80
pr		1.00	0.55	0.85		1.00	0.59	0.84		1.00	0.57	0.84		1.00	0.55	0.85
te			1.00	0.73			1.00	0.75			1.00	0.75			1.00	0.72

Table 6

Component/Domain correlations

Final Committee Ratings

domain	1A	1B	1C	1D	1E	1F	2A	2B	2C	2D	2E	3A	3B	3C	3D	3E	4A	4B	4C	4D	4E	4F
1	0.81	0.78	0.85	0.80	0.84	0.78	0.64	0.70	0.65	0.62	0.61	0.71	0.68	0.73	0.72	0.72	0.71	0.66	0.64	0.68	0.69	0.63
2	0.61	0.67	0.63	0.61	0.66	0.58	0.85	0.86	0.85	0.85	0.75	0.72	0.61	0.71	0.61	0.69	0.62	0.60	0.60	0.59	0.68	0.63
3	0.69	0.69	0.72	0.67	0.73	0.71	0.66	0.73	0.68	0.67	0.64	0.82	0.82	0.85	0.82	0.81	0.68	0.64	0.68	0.61	0.68	0.61
4	0.64	0.67	0.65	0.70	0.66	0.60	0.63	0.66	0.61	0.59	0.63	0.67	0.59	0.64	0.61	0.71	0.80	0.80	0.81	0.83	0.85	0.83
total	0.75	0.76	0.78	0.76	0.78	0.73	0.75	0.80	0.76	0.73	0.71	0.79	0.73	0.79	0.74	0.79	0.77	0.74	0.73	0.73	0.78	0.74

Table 7

Median Domain Correlations

Committee Ratings

	Domain 1	Domain 2	Domain 3	Domain 4
Domain 1median	0.81	0.64	0.72	0.65
Domain 2median	0.62	0.85	0.69	0.61
Domain 3median	0.70	0.67	0.82	0.63
Domain 4median	0.65	0.63	0.64	0.82
Total score median	0.76	0.75	0.79	0.74

Table 8
Confirmatory Factor Analysis

Principal Components Analysis

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	19.28	18.96	0.88	0.88
Factor2	0.32	0.02	0.01	0.89
Factor3	0.30	0.12	0.01	0.90
Factor4	0.18	0.03	0.01	0.91
Factor5	0.15	0.01	0.01	0.92
Factor6	0.15	0.01	0.01	0.93
Factor7	0.14	0.00	0.01	0.93
Factor8	0.13	0.01	0.01	0.94
Factor9	0.12	0.00	0.01	0.94
Factor10	0.12	0.00	0.01	0.95
Factor11	0.12	0.01	0.01	0.96
Factor12	0.10	0.00	0.00	0.96
Factor13	0.10	0.00	0.00	0.96
Factor14	0.10	0.00	0.00	0.97
Factor15	0.10	0.01	0.00	0.97
Factor16	0.09	0.00	0.00	0.98
Factor17	0.09	0.00	0.00	0.98
Factor18	0.09	0.00	0.00	0.99
Factor19	0.08	0.00	0.00	0.99
Factor20	0.08	0.00	0.00	0.99
Factor21	0.08	0.01	0.00	1.00
Factor22	0.07		0.00	1.00

Rotated Factor Loadings

Component	Factor1	Uniqueness
1A	0.9363	0.1233
1B	0.9399	0.1166
1C	0.9444	0.1082
1D	0.946	0.1051
1E	0.9444	0.1081
1F	0.9246	0.1451
2A	0.9279	0.139
2B	0.9382	0.1198
2C	0.9284	0.1382
2D	0.9197	0.1541
2E	0.9376	0.1209
3A	0.9488	0.0997
3B	0.9194	0.1547
3C	0.9369	0.1222
3D	0.9272	0.1404
3E	0.9488	0.0999
4A	0.9395	0.1174
4B	0.9403	0.1157
4C	0.9349	0.1259
4D	0.9339	0.1278
4E	0.9432	0.1105

Table 9

Domain Means and Standard Deviations by Cycle

	cycle 1			
	rt	pr	te	
domain 1	20.53	20.89	20.49	
	2.63	2.55	2.60	
domain 2	13.99	14.37	14.17	
	1.98	1.88	1.92	
domain 3	20.60	21.10	20.76	
	2.56	2.40	2.51	
domain 4	21.31	21.53	20.32	
	2.51	2.58	2.66	

	cycle 2			
	rt	pr	te	
domain 1	22.93	22.72	22.86	
	2.48	2.33	2.59	
domain 2	15.48	15.44	15.51	
	1.92	1.76	1.83	
domain 3	22.89	22.74	22.96	
	2.52	2.29	2.53	
domain 4	23.45	23.23	22.81	
	2.26	2.24	2.62	

	Cycle 3			
	rt	pr	te	cm
domain 1	24.30	23.81	23.81	23.89
	2.23	2.32	2.27	2.17
domain 2	16.44	16.09	16.16	16.14
	1.88	1.81	1.68	1.68
domain 3	24.27	23.78	23.86	23.86
	2.49	2.27	2.25	2.05
domain 4	24.68	24.22	24.03	24.24
	2.23	2.24	2.16	2.04

Table 10

Total Score Means by Cycle by Rater

Rating	Mean	Standard Deviation
cycle 1 rt	76.44	8.44
cycle 1 pr	77.88	8.26
cycle 1 te	85.38	8.24
cycle 2 rt	84.74	8.12
cycle 2 pr	84.13	7.61
cycle 2 te	84.13	8.45
cycle 3 rt	89.69	8.00
cycle 3 pr	87.91	7.92
cycle 3 te	87.86	7.51
cycle 3 cm	92.58	7.32

Table 11

Total score Intercorreletions

	cycle 1 pr	cycle 1 te	cycle 2 rt	cycle 2 pr	cycle 2 te	cycle 3 rt	cycle 3 pr	cycle 3 te	committee
cycle 1 rt	0.67	0.55	0.54	0.45	0.40	0.39	0.36	0.31	0.35
cycle 1 pr		0.62	0.53	0.57	0.45	0.40	0.43	0.35	0.40
cycle 1 te			0.47	0.45	0.56	0.33	0.34	0.38	0.35
cycle 2 rt				0.80	0.75	0.72	0.67	0.63	0.67
cycle 2 pr					0.76	0.66	0.71	0.61	0.67
cycle 2 te						0.64	0.63	0.71	0.66
cycle 3 rt							0.91	0.88	0.95
cycle 3 pr								0.88	0.95
cycle 3 te									0.92

Table 12

Total score Intercorrelations

Detail table

cycle 1	pr	te	cycle 2	pr	te	cycle 3	pr	te
rt	0.67	0.55	rt	0.80	0.75	rt	0.91	0.88
pr		0.62	pr		0.76	pr		0.88

Table 13

Rater Agreement for Cycle 1 Components

	RT/PR	RT/TE	PR/TE
1A - Demonstrating Knowledge of Content and Pedagogy	0.67	0.61	0.62
1B - Demonstrating Knowledge of Students	0.66	0.60	0.62
1C - Selecting Instructional Outcomes	0.65	0.59	0.61
1D - Demonstrating Knowledge of Resources	0.64	0.60	0.62
1E - Designing Coherent Instruction	0.63	0.59	0.61
1F - Designing Student Assessment	0.65	0.61	0.64
2A - Creating an Environment of Respect and Rapport	0.65	0.62	0.63
2B - Establishing a Culture of Learning	0.65	0.61	0.62
2C - Managing Classroom Procedures	0.61	0.57	0.60
2D - Managing Student Behavior	0.62	0.60	0.61
2E - Organizing Physical Space	0.66	0.64	0.64
3A - Communicating with Students	0.63	0.61	0.60
3B - Using Questioning and Discussion Techniques	0.64	0.61	0.62
3C - Engaging Students in Learning	0.63	0.61	0.61
3D - Using Assessment in Instruction	0.63	0.60	0.61
3E - Demonstrating Flexibility and Responsiveness	0.62	0.58	0.60
4A - Reflecting on Teaching	0.66	0.60	0.61
4B - Maintaining Accurate Records	0.65	0.58	0.61
4C - Communicating with Families	0.67	0.59	0.62
4D - Participating in a Professional Community	0.66	0.57	0.58
4E - Growing and Developing Professionally	0.68	0.60	0.61
4F - Demonstrating Professionalism	0.69	0.59	0.59

Table 14

Rater Agreement for Cycle 2 Components

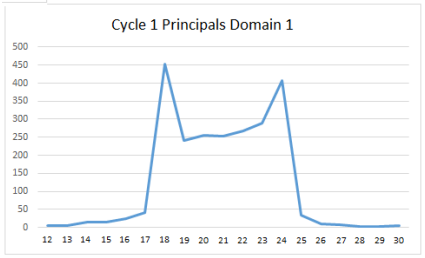
1A - Demonstrating Knowledge of Content and Pedagogy	0.74	0.73	0.71
1B - Demonstrating Knowledge of Students	0.73	0.68	0.71
1C - Selecting Instructional Outcomes	0.72	0.70	0.70
1D - Demonstrating Knowledge of Resources	0.73	0.70	0.70
1E - Designing Coherent Instruction	0.72	0.70	0.69
1F - Designing Student Assessment	0.68	0.64	0.66
2A - Creating an Environment of Respect and Rapport	0.71	0.68	0.70
2B - Establishing a Culture of Learning	0.72	0.69	0.71
2C - Managing Classroom Procedures	0.68	0.67	0.67
2D - Managing Student Behavior	0.68	0.66	0.66
2E - Organizing Physical Space	0.74	0.72	0.74
3A - Communicating with Students	0.72	0.69	0.70
3B - Using Questioning and Discussion Techniques	0.65	0.61	0.62
3C - Engaging Students in Learning	0.67	0.65	0.66
3D - Using Assessment in Instruction	0.67	0.65	0.65
3E - Demonstrating Flexibility and Responsiveness	0.72	0.71	0.72
4A - Reflecting on Teaching	0.74	0.70	0.71
4B - Maintaining Accurate Records	0.75	0.71	0.73
4C - Communicating with Families	0.73	0.68	0.69
4D - Participating in a Professional Community	0.75	0.69	0.70
4E - Growing and Developing Professionally	0.73	0.68	0.67
4F - Demonstrating Professionalism	0.76	0.69	0.71

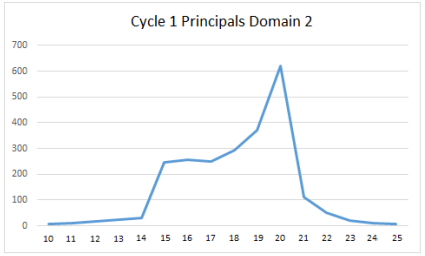
Table 15

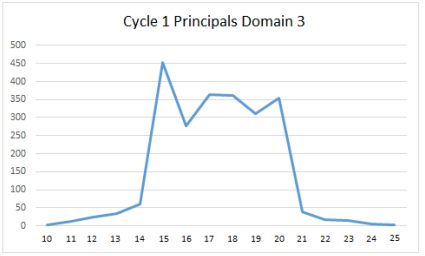
Rater Agreement for Cycle 3 Components

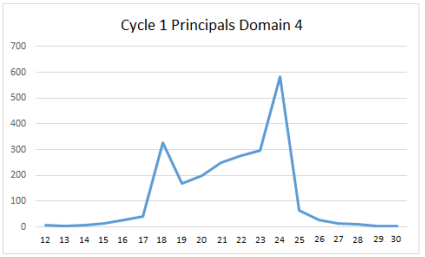
	RT/PR	RT/TE	PR/TE
1A - Demonstrating Knowledge of Content and Pedagogy	0.79	0.73	0.75
1B - Demonstrating Knowledge of Students	0.78	0.73	0.76
1C - Selecting Instructional Outcomes	0.79	0.75	0.76
1D - Demonstrating Knowledge of Resources	0.77	0.74	0.76
1E - Designing Coherent Instruction	0.77	0.73	0.74
1F - Designing Student Assessment	0.77	0.72	0.73
2A - Creating an Environment of Respect and Rapport	0.74	0.69	0.72
2B - Establishing a Culture of Learning	0.75	0.71	0.73
2C - Managing Classroom Procedures	0.74	0.71	0.73
2D - Managing Student Behavior	0.73	0.71	0.73
2E - Organizing Physical Space	0.75	0.71	0.74
3A - Communicating with Students	0.76	0.73	0.74
3B - Using Questioning and Discussion Techniques	0.73	0.69	0.71
3C - Engaging Students in Learning	0.74	0.70	0.72
3D - Using Assessment in Instruction	0.76	0.73	0.74
3E - Demonstrating Flexibility and Responsiveness	0.75	0.72	0.76
4A - Reflecting on Teaching	0.78	0.72	0.75
4B - Maintaining Accurate Records	0.79	0.75	0.78
4C - Communicating with Families	0.79	0.75	0.76
4D - Participating in a Professional Community	0.79	0.74	0.77
4E - Growing and Developing Professionally	0.77	0.73	0.76
4F - Demonstrating Professionalism	0.77	0.71	0.74

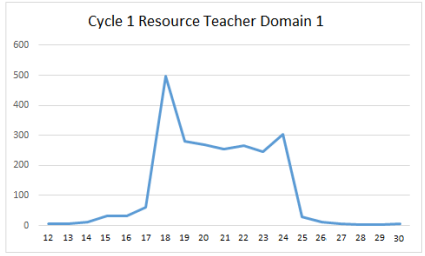
Chart 3
Cycle 1 Domain Score Distribution by Rater

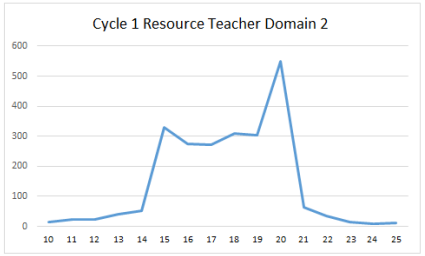


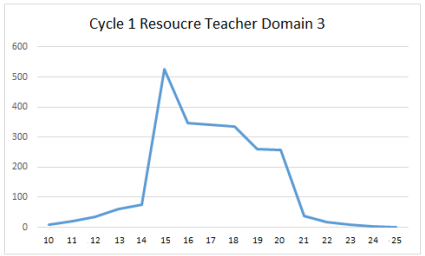


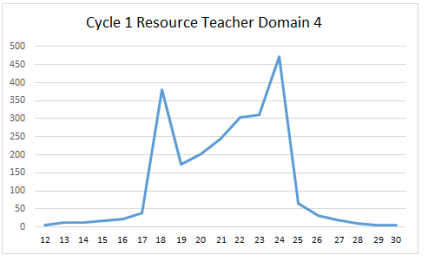


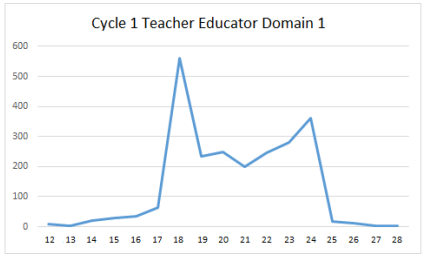


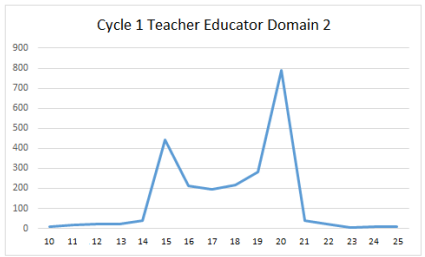


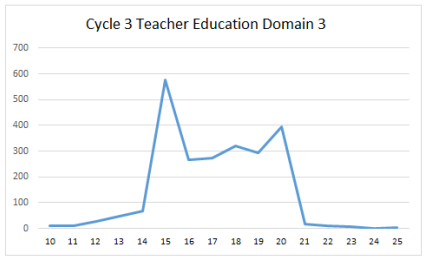












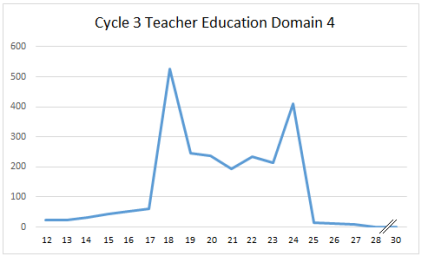
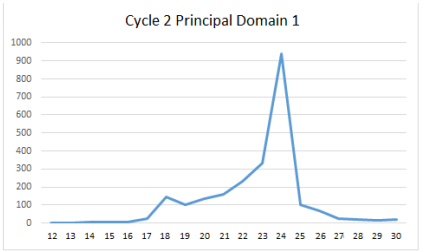
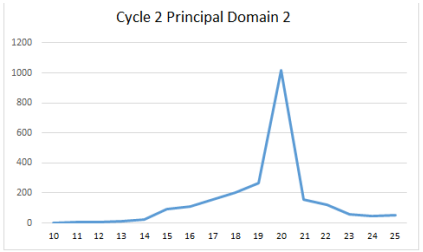
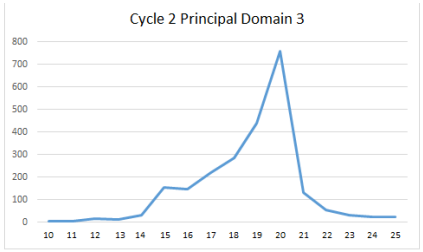
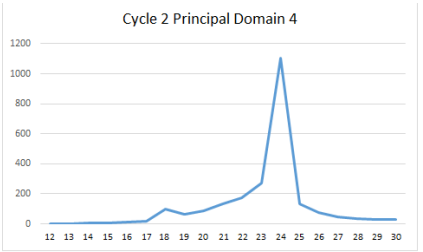


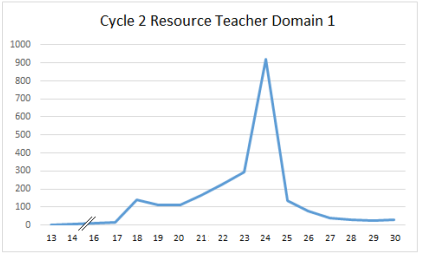
Chart 4
Cycle 2 Domain Score by Rater

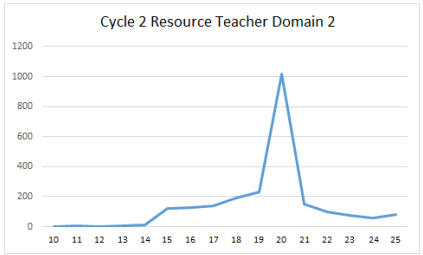


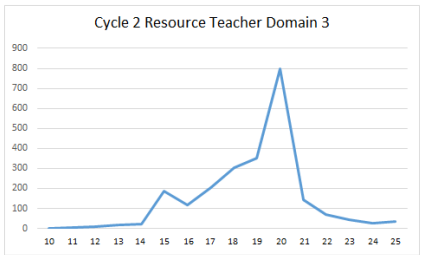


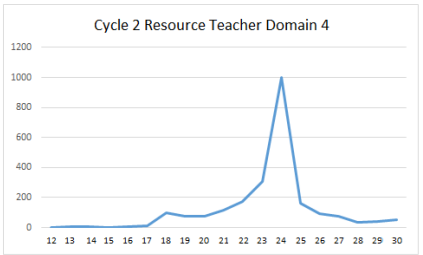


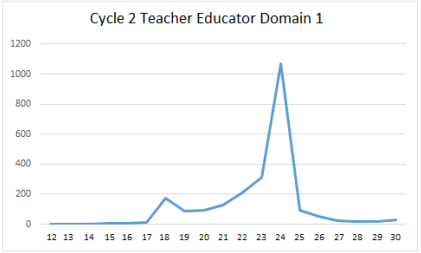


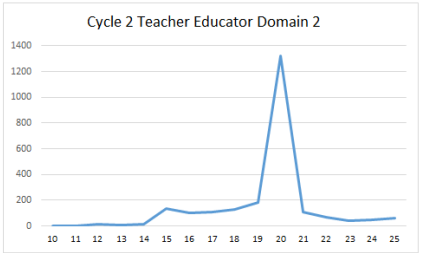


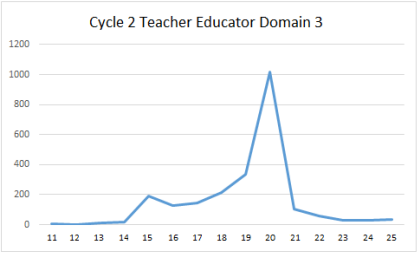












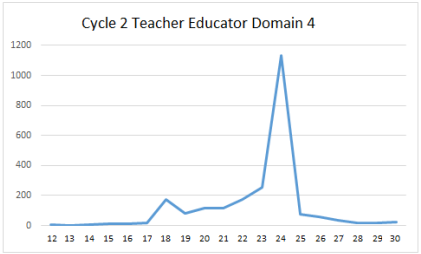
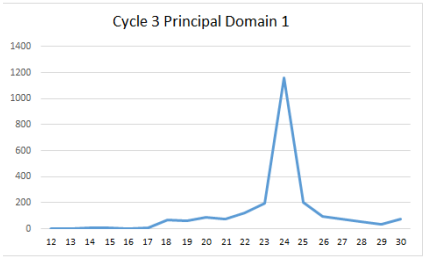
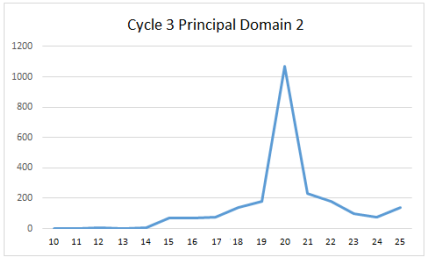
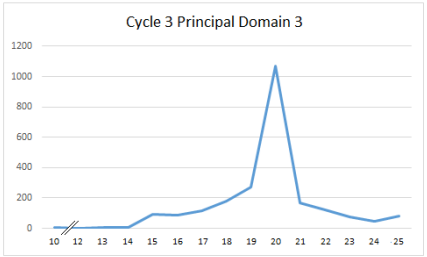
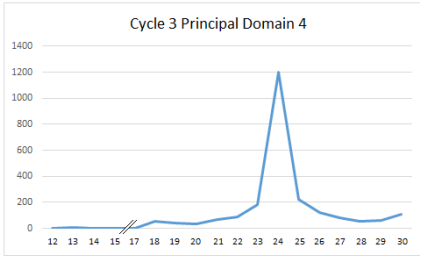


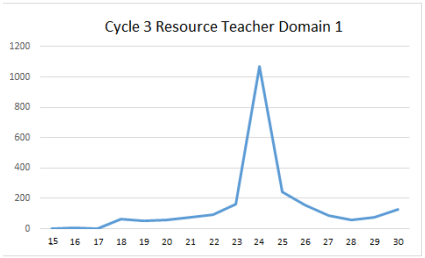
Chart 5
Cycle 3 Domain Score by Rater

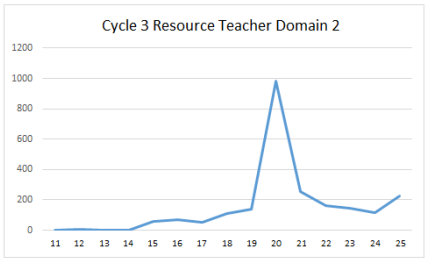


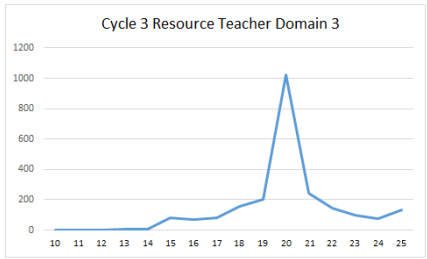


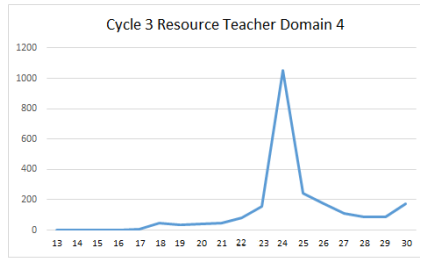


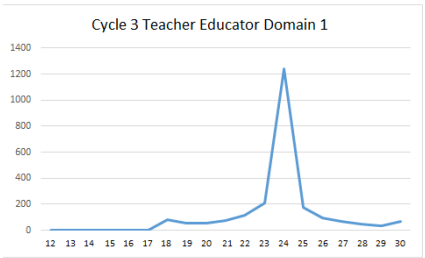


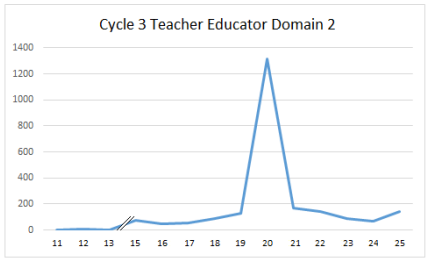


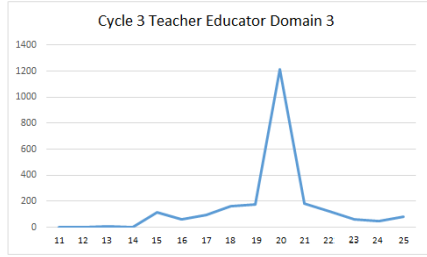


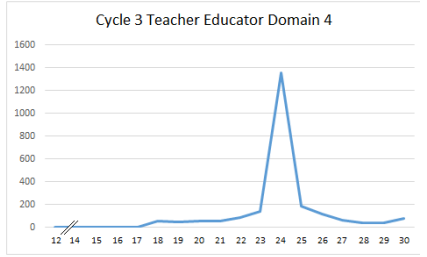


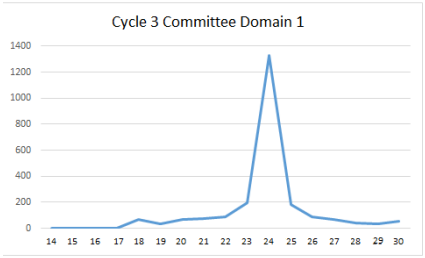


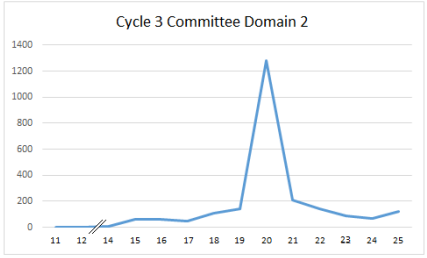


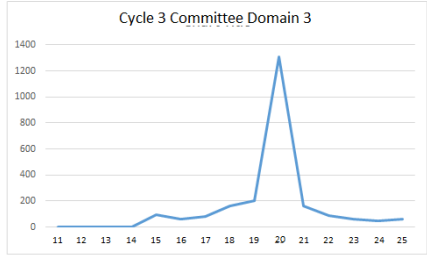












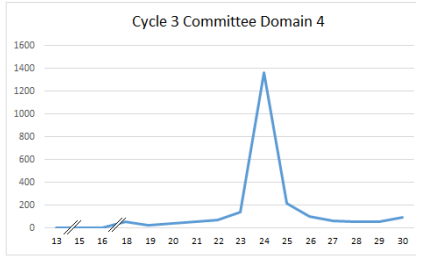
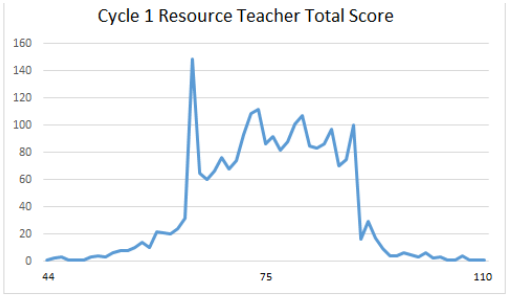
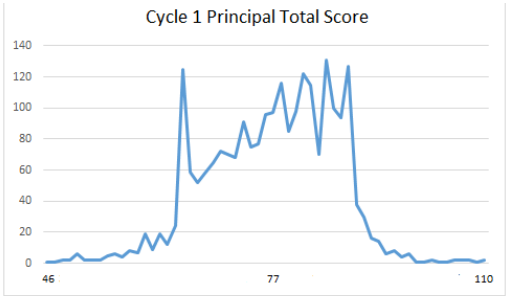
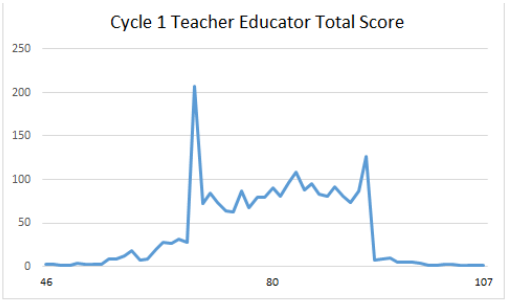
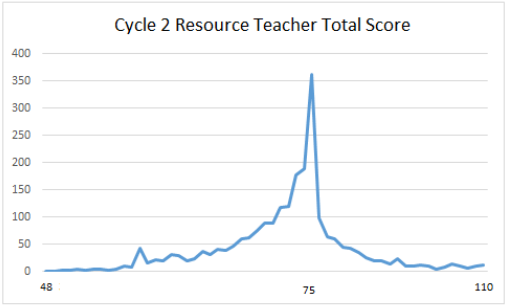


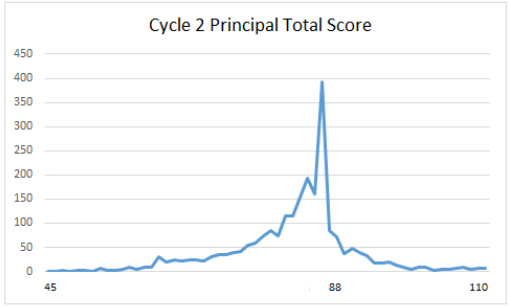
Chart 6
Total Score Distribution by Cycle

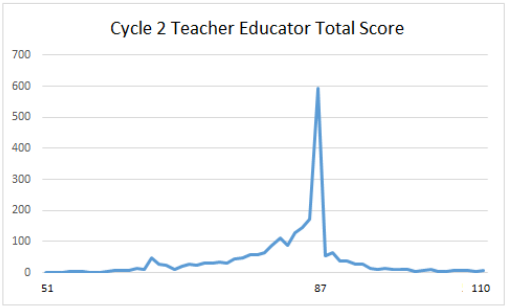


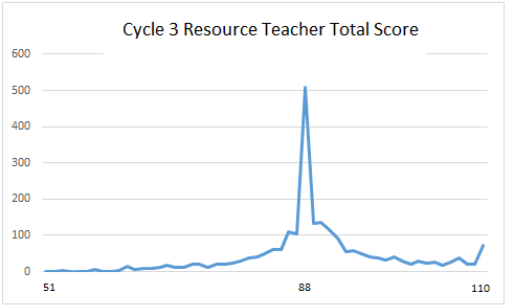


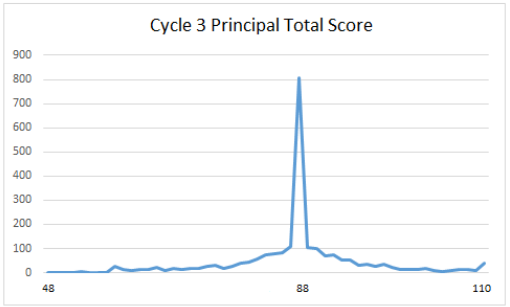


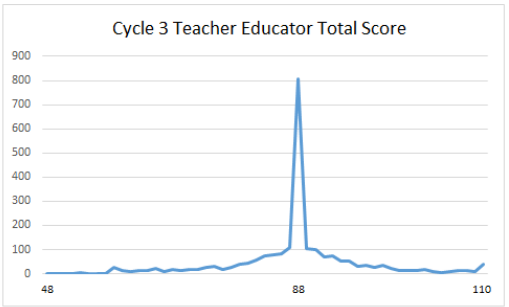












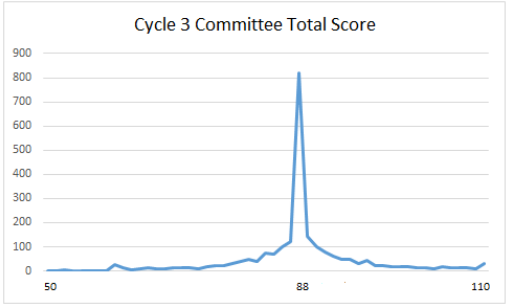


Table 16

Skewness and Kurtosis of Domain Score by Cycle

	cycle 1		
	rt	pr	te
domain 1	0.09	-0.11	-0.14
	-0.01	-0.22	-0.55
domain 2	-0.29	-0.36	-0.47
	0.11	0.11	-0.14
domain 3	-0.12	-0.17	-0.15
	0.17	0.16	-0.27
domain 4	-0.30	-0.39	-0.30
	0.00	-0.02	-0.24

	cycle 2		
	rt	pr	te
domain 1	-0.34	-0.58	-0.61
	0.86	1.20	1.41
domain 2	-0.26	-0.42	-0.55
	0.86	1.18	1.55
domain 3	-0.12	-0.17	-0.15
	0.17	0.16	-0.27
domain 4	-0.41	-0.66	-0.82
	1.89	2.08	1.66

	Cycle 3			
	rt	pr	te	cm
domain 1	0.02	-0.32	-0.28	-0.26
	1.08	2.13	2.37	2.88
domain 2	-0.18	-0.22	-0.16	-0.11
	0.61	1.12	1.78	1.60
domain 3	-0.03	-0.15	-0.14	-0.14
	0.99	1.70	1.61	2.26
domain 4	0.01	-0.19	-0.28	-0.07
	1.24	2.74	3.17	3.34

Table 17

Skewness and Kurtosis of Total Scores by Cycle

Rating	Skewness	Kurtosis
cycle 1 rt	-0.15	0.10
cycle 1 pr	-0.30	0.15
cycle 1 te	-0.26	0.15
cycle 2 rt	-0.26	1.22
cycle 2 pr	-0.52	1.82
cycle 2 te	-0.55	1.50
cycle 3 rt	-0.02	1.07
cycle 3 pr	-0.21	2.29
cycle 3 te	-0.20	2.40
cycle 3 cm	-0.23	2.41

Table 18

Generalizability Analysis

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
I	909910.5	2330	390.5195	35.14889	37.75238	37.75238	35.9	1.28017
R:I	236587.6	4662	50.74808	10.60449	16.91603	16.91603	16.1	0.36228
C	501302.9	2	250651.5	35.83714	35.83714	23.89142	22.7	25.34497
IC	197426	4660	42.36609	7.81049	7.81049	7.81049	7.4	0.30676
RC:I	176546.4	9324	18.93463	18.93463	18.93463	18.93463	18	0.27728
Total	2021773	20978	100%					

Generalizability coefficients

Coef_G relative	0.87
Coef_G absolute	0.87

Table 19

Analysis of Variance Cycles 2 and 3

Analysis of variance

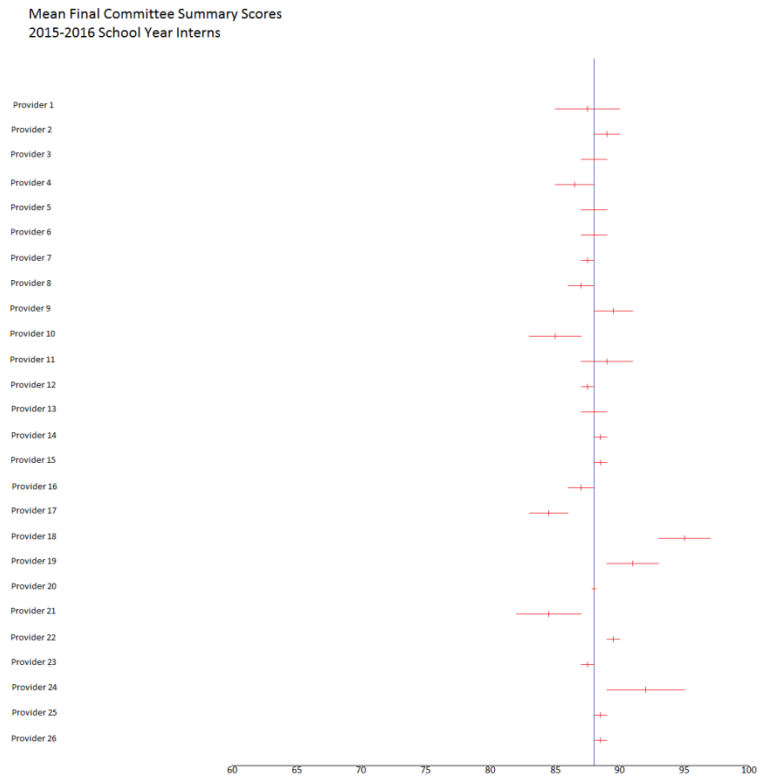
				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
I	1052378.88560	2330	451.66476	75.13058	75.12852	75.12852	99.0	2.20453
R:I	4166.00000	4662	0.89361	0.28621	0.44680	0.44680	0.6	0.00983
C	0.00114	1	0.00114	-0.00004	-0.00004	-0.00002	0.0	0.00000
IC	719.66552	2330	0.30887	-0.00410	-0.00410	-0.00410	0.0	0.00374
RC:I	1497.33333	4662	0.32118	0.32118	0.32118	0.32118	0.4	0.00665
Total	1058761.8856	13985					100%	

Table 20

Correlation Between Total Score and Admissions GPA

	r
cycle 1 rt	0.14
cycle 1 pr	0.16
cycle 1 te	0.12
cycle 2 rt	0.15
crcle 2 pr	0.15
cycle 2 te	0.11
cycle 3 rt	0.13
cycle 3 pr	0.11
cycle 3 te	0.12
cycle 3 cm	0.14

Chart 7
Trial Provider Mean Performance Chart



References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Bachman, L (2002). Alternative Interpretations of Alternative Assessments: Some Validity Issues in Educational Performance Assessments. Educational Measurement: Issues and Practice, Fall 2002: 5-18.

Barrett P (2001). Assessing the Reliability of Rating Data. Downloaded from http://www.pbarrett.net/presentations/rater.pdf on December 7, 2016.

Bell C, Gitomer D, McCaffrey D, Hamre B & Pianta R (2012). An Argument Approach to Observation Protocol Validity. Educational Assessment, 17:62–87, 2012.

Benjamin, W (2002). Development and Validation of Student Teaching Performance Assessment Based on Danielson's Framework for Teaching. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).

Borsboom D, Mellenbergh G & van Heerden J (2004). The Concept of Validity. Psychological Review, 111(40: 106101071, 2004.

Brennan R (2001). An Essay on the History and Future of Reliability from the Perspective of Replications. Journal of Educational Measurement, 38(4): 295-317, 2001.

Briesch A, Swaminathan H, Welsh M & Chafouleas S (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. Journal of School Psychology, 52: 13-35, 2014.

Brockman D (2015). KTIP article. Frankfort, Kentucky: Unpublished manuscript, 2015.

Brockman D (2016). Personal communication.

Cardinet J, Johnson S & Pini G (2010). Applying Generalizability Theory Using Edug. New York: Taylor and Francis, 2010.

Chapelle C, Enright M & Jamieson J (2010). Does an Argument-Based Approach to Validity Make a Difference? Educational Measurement: Issues and Practice, 29(1), pp. 3–13, 2010.

Cizek G (2012). Defining and Distinguishing Validity: Interpretations of Score Meaning and Justifications of Test Use. Psychological Methods, 17(1): 31–43, 2012.

Clauser B (2000). Recurrent Issues and Recent Advances in Scoring Performance Assessments. *Applied Psychological Measurement,* 24(4): 310–324, 2000.

Clauser B, Clyman S & Swanson D (1999). Components of Rater Error in a Complex Performance Assessment. Journal of Educational Measurement, 36(1): 29-45, 1999.

Cronbach L, Gleser G, Nanda H & Rajaratnam N (1972). The Dependability of Behavioral Measurements. New York: Wiley, 1972.

Cronbach L & Meehl P (1955). Construct Validity in Psychological Tests. Psychological Bulletin, 52(4): 281-302, 1955.

Crooks T, Kane M & Cohen A (1996). Threats to the Valid Use of Assessments. Assessment in Education: Principles, Policy & Practice, 3:3, 265-286.

Danielson, C (2007). Enhancing Professional Practice: A Framework for Teaching 2nd. Association for Supervision & Curriculum Development. Kindle Edition.

Danielson C & Dwyer C (1995). How Praxis III Supports Beginning Teachers. Educational Leadership, March 1995, 66-67

Delandshere G & Petrosky A (1998). Assessment of Complex Performances: Limitations of Key Measurement Assumptions. Educational Researcher, 27(2): 14-24, 1998.

Denegar C & Ball D (1993). Assessing Reliability and Precision of Measurement: An Introduction to lntraclass Correlation and Standard Error of Measurement. Journal of Sport Rehabilitation, 1993, 35-42.

Denner P, Norman A & Lin S (2008). Fairness and Consequential Validity of Teacher Work Samples. Journal of Personnel Evaluation in Education, Volume 21(3): pp 235–254.

Dwyer C (1994). Development of the knowledge base for the Praxis III: Classroom Performance Assessment criteria. Princeton, NJ: Educational Testing Service.

Dwyer C (1998). Psychometrics of Praxis III: Classroom Performance Assessments. Journal of Personnel Evaluation in Education 12(2): 163–187, 1998.

Engelhard G (1994). Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. Journal of Educational Measurement, 31(2): 93-112, 1994.

Haertel E (1999). Validity Arguments for High-Stakes Testing: In Search of the Evidence. Educational Measurement: Issues and Practice, Winter 1999, 5-9.

Guilford, J (1954). Psychometric methods. New York: McGraw-Hill, 1954.

Goe, L. (2007). The link between teacher quality and student outcomes: A research synthesis. Washington, DC: National Comprehensive Center for Teacher Quality.

Goe L, Bell C & Little O (2008). Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. Washington, D.C.: National Comprehensive Center for Teacher Quality, 2008.

Gordon R, Kane T & Staiger D (2006). Identifying Effective Teachers Using Performance on the Job. Brookings Institution Hamilton Project, 2006.

Güerere C (2013). Value-Added and Observational Measures Used in the Teacher Evaluation Process: A Validation Study. Dissertation, University of South Florida, 2013.

Hazi H (2014). The Marketing of Teacher Evaluation: The Seductive Claims of Instruments. The Washington Educational Research Association Journal, 6(1): 20-9, 2014.

Hibpshman T (2005). The Logic and History of Kentucky's New And Experienced Teacher Standards. Frankfort, Kentucky: Education Professional Standards Board. Unpublished manuscript, 2005.

Hibpshman T (2013). KTIP, PGES, and All That. Frankfort, Kentucky: Education Professional Standards Board, 2013. Powerpoint presentation.

Hill H, Ball D , Blunk M, Goffney I & Rowan B (2007). Validating the Ecological Assumption: The Relationship of Measure Scores to Classroom Teaching and Student Learning. Measurement: Interdisciplinary Research and Perspectives, 5(2-3): 107-118, 2007.

Hill H, Charalambous C, Blazar D, McGinn D, Kraft M, Beisiegel M, Humez A, Litke E & Lynch K (2012). Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation. Educational Assessment, 17:88–106, 2012.

Hill H, Charalambous C & Kraft M (2012). When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. Educational Researcher, 41(2): pp. 56–64, 2012.

Ho A & Kane T (2013). The Reliability of Classroom Observations by School Personnel. Bill and Melinda Gates Foundation, Met Project Research Paper, 2013.

Holtzapple E (2003). Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System. Journal of Personnel Evaluation in Education 17(3): 207–219, 2003.

Hood, L (2015). PreK-3 Danielson FFT Research Study. Paper presentation at the 2015 Gateways to Higher Education Forum.

Hood L, Kasperski D, Hunt E, DeStefano L, Rodriguez S, Garcia G & Kirchoff A (2015). Studying the Danielson Framework for Teaching in PreK-3rd Grade Classrooms: A White Paper of the Research and Preliminary Findings. Urbana, Illinois: Center for the Study of Educational Policy, 2015.

Ingersoll, R (2002). The Teacher Shortage: A Case of Wrong Diagnosis and Wrong Prescription. NASSP Bulletin, 86(631): 16-31, 2002.

Jaeger R (1993). Live vs. Memorex: Psychometric and Practical Issues in the Collection of Data on Teachers' Performances in the Classroom. Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993). Eric document 360325.

Johnson E & Semmelroth C (2015). Validating an Observation Protocol to Measure Special Education Teacher Effectiveness. Journal of Special Education and Early Childhood Studies, Fall 2015, pp. 98-119.

Kane M (1992). An Argument-Based Approach to Validity. Psychological Bulletin, 112(3): 527-535, 1992.

Kane M (2001). Concerns in Validity Theory. Journal of Educational Measurement, 38(4), Measurement Update for the 21st Century: 319-342, 2001.

Kane M (2002a). Inferences about Variance Components and Reliability-Generalizability Coefficients in the Absence of Random Sampling. Journal of Educational Measurement, 39(2): pp. 165-181, 2002.

Kane M (2002b). Validating High-Stakes Testing Programs. Educational Measurement: Issues and Practice, Spring 2002, 31-41.

Kane M (2004). Certification Testing as an Illustration of Argument-Based Validation. Measurement, 2(3): 135-170, 2004.

Kane M (2008). Errors of Measurement, Theory, and Public Policy. The 12th Annual William H. Angoff Memorial Lecture presented at Educational Testing Service, Princeton, New Jersey, 2008.

Kane M (2011). The Errors of Our Ways. Journal of Educational Measurement, 48(1): pp. 12–30, 2011.

Kane M, Crooks T & Cohen A (1999). Validating Measures of Performance. Educational Measurement: Issues and Practice, 5-17, Summer 1999.

Kane T, Taylor E, Tyler J & Wooten A (2010). Identifying Effective Classroom Practices Using Student Achievement. Cambridge, Massachusetts: National Bureau of Economic Research, NBER Working Paper 15803, 2010.

Kentucky Advisory Council for Internships (KACI) (2013a). Minutes, August 29, 2013.

Kentucky Advisory Council for Internships (KACI) (2013b). Minutes, November 18, 2013.

Kentucky Advisory Council for Internships (KACI) (2014). Minutes, January 30, 2014.

Kentucky Department of Education (KDE) (2014). Framework for Teaching. Frankfort, Kentucky: Author, 2014.

Kentucky Education Professional Standards Board (EPSB) (2015a). Kentucky Teacher Internship Program Handbook. Frankfort, Kentucky: Author, 2015.

Kentucky Education Professional Standards Board (EPSB) (2015b). KTIP Pilot Training. Frankfort, Kentucky: Powerpoint presentation.

Kentucky Education Professional Standards Board (EPSB) (2015c). KTIP Timeline. Frankfort, Kentucky: Internal planning document.

Kentucky Education Professional Standards Board (EPSB) (2015d). KTIP Initial Training Evaluation Report. Frankfort, Kentucky: Internal reporting document.

Kentucky Education Professional Standards Board (EPSB) (2016). Kentucky Teacher Internship Program (KTIP) Committee Training. Frankfort, Kentucky: Author, 2016, Powerpoint presentation.

Kentucky Education Professional Standards Board (EPSB) (ND). Intern Performance Record. Frankfort, Kentucky: Author, no date.

Kessler V & Howe C (2012). Understanding Educator Evaluations in Michigan: Results from Year 1 of Implementation. Lansing, Michigan: Michigan Department of Education, 2012. Downloaded from https://www.michigan.gov/documents/mde/Educator_Effectiveness_Ratings_Policy_Brief_403184_7.pdf on December 28, 2016.

Lane S, Parke C & Stone C (1998). A Framework for Evaluating the Consequences of Programs. Educational Measurement Issues and Practice, Summer 1998, 24-28.

Lane S & Stone C (2002). Strategies for Examining the Consequences of Assessment and Accountability Programs. Educational Measurement Issues and Practice, Spring 2002, 23-30.

Lash A, Tran L & Huang M (2016). Examining the validity of ratings from a classroom observation instrument for use in a district's teacher evaluation system. Regional Education Laboratory at Westfield, 2016.

Lazarev V & Newman D (2013). Title: How Non-Linearity and Grade-level Differences Complicate the Validation of Observation Protocols. Paper presentation at the SREE Fall Conference, 2013.

Liguori B (2015). Lessons learned from Teacher PGES Implementation. Frankfort, Kentucky: video, Kentucky Department of Education.

Lissitz R & Samuelsen K (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. Educational Researcher, 36(8): 437-448, 2007.

Mehrens W (1997). The Consequences of Consequential Validity. Educational Measurement: Issues and Practice, Summer 1997, 16-18.

Messick S (1994). THE Interplay of Evidence and Consequences in the Validation of Performance Assessments. Educational Researcher, 23(2), pp. 13-2, 1994.

Milanowski A (2004). Relationships Among Dimension Scores of Standards-Based Teacher Evaluation Systems, and the Stability of Evaluation Score – Student Achievement Relationships Over Time. Madison, Wisconsin: Consortium for Policy Research in Education, Wisconsin Center for Education Research, University of Wisconsin-Madison, CPRE-UW Working Paper Series TC-04-02, 2004.

Milanowski A (2011). Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching. Paper presented at the American Education Research Association annual meeting on April 10, 2011 in New Orleans, LA.

MIslevy R (2004). Can There Be Reliability without "Reliability?" Journal of Educational and Behavioral Statistics, 29(2), Value-Added Assessment Special Issue (Summer, 2004), pp. 241-244.

Moss P (1998). The Role of Consequences in validity Theory. Educational Measurement: Issues and Practice, Summer 1998, 6-12.

Moss P (2007). Comments on Lissitz and Samuelsen. Educational Researcher, 36(8): pp. 470–476, 2007.

Myford C, Villegas A, Reynolds A, Camp R, Danielson C, Jones J, Knapp J, Lehman P, Mandinach E, Morris L, Sims-Gunzanhauser A & Sjostrom B (1993). Formative Studies of Praxis III: Classroom Performance Assessments – An Overview. Princeton, New Jersey: Educational Testing Service, 1993.

Naizer G (1992). Basic Concepts in Generalizability Theory: A More Powerful Approach to Evaluating Reliability. Paper presented at the Annual Meeting of the Southwest Research Association, Houston, Texas, January-February 1992.

Noell G, Brownell M, Buzick H & Jones N (2014). Using Educator Effectiveness Measures to Improve Educator Preparation Programs and Student Outcomes. CEEDAR Center, 2014.

Nunnally J & Bernstein I (1994). Psychometric Theory. McGraw-Hill, 1994.

Roegman R, Goodwin A, Reed R & Scott-McLaughlin R (2016). Unpacking the data: an analysis of the use of Danielson's (2007) Framework for Professional Practice in a teaching residency program. Educational Assessment Evaluation and Accountability 28:111–137, 2016.

Ryan K (2002). Assessment Validation in the Context of High-Stakes Assessment. Educational Measurement: Issues and Practice, Spring 2002, 7-15.

Scriven M (1987). Validity in Personnel Evaluation. Journal of Personnel Evaluation in Education 1: 9-23, 1987.

Shrout P & Fleiss J (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin 86(2): 420-428, 1979.

Sireci S & Green P (2000). Legal and Psychometric Criteria for Evaluating Teacher Certification Tests. Educational Measurement: Issues and Practice, Spring 2000, 22-31.

Stata.com (ND). anova — Analysis of variance and covariance. STATACORP user's manual. Accessed January 3, 2017 from http://www.stata.com/manuals13/ranova.pdf.

Stata.com (ND). factor — Factor analysis. STATACORP user's manual. Accessed January 4, 2017 from http://www.stata.com/manuals13/mvfactor.pdf.

Stata.com (ND). xtreg — Fixed-, between-, and random-effects and population-averaged linear models. STATACORP user's manual. Accessed January 3, 2017 from http://www.stata.com/manuals13/xtxtreg.pdf.

Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9(4). Retrieved September 30, 2016 from http://PAREonline.net/getvn.asp?v=9&n=4.

Tobias, S. (2009). An eclectic appraisal of the success and failure of constructivist instruction. In S. Tobias & T Duffy (Eds.) Constructivist instruction: Success or failure? (pp. 335-350) New York: Routledge.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. Review of Educational Research, 73, 89–122.

Webb N & Shavelson R (2005). Generalizability Theory: Overview. In Everitt B & Howell D (eds.), Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, 2005.

Webb N, Shavelson R & Haertel E (2016). Reliability Coefficients and Generalizability Theory. In Rao C & Sinharay S (eds.) Handbook of Statistics, Vol. 26, pp. 81-124. Elsevier, 2016.

Zarębski T (2009). Toulmin's Model of Argument and the "Logic" of Scientific Discovery. Studies in Logic, Grammar, and Rhetoric, 16(29): 267-283, 2009.