VARI-EPP: Valid and Reliable Instruments for Educator Preparation Programs -

A Statewide Model

# Objectives

**The purpose of the study (research questions and / or study objectives) should be clearly and succinctly stated. In experimental designs, objectives will be stated as hypotheses to be tested.**

The purpose of this study is to examine the validity and reliability of three unit-wide evaluation instruments used during student teaching. These instruments include a student teaching pedagogy evaluation, a dispositions evaluation, and subject-specific student teaching assessments (i.e., the observation evaluation for preservice mathematics teachers that was created to meet the requirements of the National Council of Teachers of Mathematics, which is the Specialized Professional Association for this content area). Through the studies described in this proposal, we hope to answer the following questions:

* Are the student teaching evaluation instruments valid? Specifically, what are the results of construct, content, and concurrent validity studies?
* Are the student teaching evaluation instruments reliable? Specifically, what are the results of test – retest, internal consistency, and inter-rater agreement studies?

# Background and Rationale

**Summarize and synthesize the available research (including published data) to provide justification for the study. Evaluate prior research for relevance to the research question under study. When the proposed research is the first of its type to involve human participants, the results of relevant animal studies must be included. Discuss the anticipated results and potential pitfalls. Describe the significance of the research including potential benefit for individual subjects or society at large. Discuss how public health and social welfare might be enhanced.**

As educator preparation providers (EPP’s) prepare for the transition from the 2010 National Council for Accreditation of Teacher Education (NCATE) standards to the 2013 Council for Accreditation of Educator Preparation (CAEP) standards, many EPP’s are struggling with the new, rigorous standards for assessment. The new CAEP standards are in effect and EPP’s have little time to create a strategic plan to address the rigor prior to the reaccreditation reviews and self-study that begins in Spring 2016. Specifically, many EPP’s are struggling with the development of valid and reliable assessments given that true reliability and validity studies take a wealth of time, resources and require the expertise of content experts and psychometricians. Virtually all EPP’s use a performance-based evaluation tool during clinical experiences, which presents challenges due to the subjectivity of such measurements; therefore, validity and reliability studies on such instruments are warranted and large, research institutions such as The Ohio State University are well positioned to do such studies.

There are similarities between the 2010 NCATE and 2013 CAEP standards. The CAEP standards continue the previous emphasis on candidates’ performance and continuous quality improvement, but have a greater emphasis on program impact related to P-12 student learning which requires EPP’s to revamp assessments to be in compliance (Tomei, 2014). Some major differences that relate to the current proposal include (a) a focus on initial classroom preparation, (b) re-envisioned professional dispositions, (c) greater emphasis on diversity, (d) stronger clinical partnerships, (e) greater involvement of stakeholders, (f) increased emphasis on impact, (g) expectation of external benchmarking, (h) assurance that accountability metrics are valid, reliable, and fair, and (i) heightened expectations for the quality of evidence (Tomei, 2014). One drastic departure from the NCATE accreditation process is the requirement of valid and reliable assessments to demonstrate candidate quality and that various stakeholders must contribute to the validity of the assessments. Education Program Providers must consider the principles for measures used in the CAEP accreditation process, which include (a) validity and reliability, (b) relevance, (c) verifiability, (d) representativeness, (e) cumulativeness, (f) fairness, (g) stakeholder interest, (h) benchmarks, (i) vulnerability to manipulation, and (j) actionability (Ewell, 2013). These illustrate the need for the current proposal.

Validity and reliability are two of the most important criteria for evaluation and assessment tools. This is especially true if the assessment is used as a measurement instrument in research. Generally speaking, validity is defined as the degree to which a test or measuring instrument actually measures what it purports to measure or how well a test or a meaning instrument fulfills its function (Anastasi & Urbina, 1997). In educational research, there are four types of validity that are important to address including face, content, construct, and criterion-related (concurrent and predictive) validity (Oluwatayo, 2012). For the purposes of the current project, we will focus on **construct, concurrent and content validity. Predictive validity studies will be conducted as part of future studies**. Reliability refers to dependability, consistency, reproducibility or replicability over time, over instruments and over groups of respondents (Bowling, 2009). There are three principle types of reliability in educational research: stability, equivalence and internal consistency (Oluwatayo, 2012). For the purposes of this project, we collect data for **test-retest reliability, internal consistency, and inter-rate agreement**. The studies described in this proposal are the first steps for longer, more rigorous validity and reliability studies.

# Procedures

## Research Design

**The research design should be identified and should be appropriate to answer the research question(s) under study. Describe the type of research proposed (e.g. experimental, correlational, survey, qualitative) and specific study design that will be used (e.g. pre-test /post / test control group design, cross-sectional design; prospective longitudinal cohort design; phase III double-blind randomized control group design).**

This is a correlational study designed to investigate the validity and reliability of three evaluation instruments used to rate student teachers’ performance including the 1) pedagogy evaluation, 2) dispositions evaluation, and 3) a subject specific pedagogy evaluation (i.e., mathematics).

## Sample

**Describe the sampling approach. For experimental designs, include justification for sample size determination. Identify the procedures that will be used to recruit, screen, and follow study volunteers. Specifically define the study sample (number of subjects to be enrolled, characteristics of subjects to be included in and excluded from the research).**

For the majority of the analyses (construct validity, concurrent validity, test-retest reliability, and internal consistency) we will use all of the data from all of the forms for student teaching during the 2014-2015 academic year. We anticipate to have approximately 400 OSU students who will be completing student teaching next year and 500 students from other IHEs in Ohio. The Office of Educator Preparation is responsible for collecting, managing, and/or accessing all of the data for students in preservice teacher preparation programs at OSU. This is required as part of the accreditation process; therefore, our office already has access to all OSU preservice teachers’ ratings on the student teaching evaluation forms through Tk20 (our data management system). We are not requesting access to new sources of data for OSU students, we are asking permission to do a secondary analysis of data that will exist regardless of the research study. The other IHE’s who will participate in this study will also submit the ratings from the instruments of all of their students.

For the content validity portion of the study, we will invite approximately 5 content experts from various educator preparation providers around the state to participate in the study. These experts will review the developed instrument to evaluate content validity using the rubrics in the appendices.

For the inter-rater reliability portion of the study, we will select a random sample of 10% of the total sample to collect inter-rater agreement data. Basic training for how to complete the forms will be provided to all graduate assistants, staff, and faculty members who supervise student teachers. These supervisors will then have an opportunity to volunteer to do inter-rater reliability. Supervisors and student teachers who will be part of the inter-rater reliability sample must consent to be part of this study.

## Measurement / Instrumentation

**Identify the variables of interest and study endpoints (where applicable). Justify measurement techniques selected. Provide validity and reliability data for selected measures.**

The instruments that will be evaluated as part of this study are included in the Appendices A. Pedagogy, B. Dispositions, and C. Content Area Assessments). A rubric will be used to evaluate the instruments as part of the content validity portion of the study and the rubrics are included in Appendix D. Specifically, the rubric will be used by the content experts to evaluate each item for (1) representativeness of the content domain; (2) clarity of the item; (3) comprehensiveness of the measure.

## Detailed study procedures

**Methods for study data collection and for avoiding / minimizing subject risks should be included. Include a timeline for subject evaluations and the duration of subject participation in the project. Identify the plans the proposed safeguards for subject confidentiality (plans for coding data and for securing written and electronic subject records). Indicate how long personal information will be stored once the study is completed.**

**Methods will vary with the research approach used (qualitative, quantitative). The selected methods should be sufficiently described to justify the use of the approach for answering the defined research question. Methods should also be described in adequate detail so that IRB members may assess the potential study risks and benefits.**

**Validity Studies:**

Several types of validity and reliability analyses will be conducted within this study. Procedures for each of these portions of the study are described in the following sections. The first part of the study includes an analysis to investigate the **content validity** of the evaluation instruments. For this, we will recruit five experts in teacher preparation across the state to evaluate the instruments using the rubrics included in Appendix D. These experts are serving on a Valid And Reliable Instruments for Educator Preparation (VARI-EPP) Evaluation Committee that The Ohio State University has convened and currently chairs. The content experts will be given the evaluation instruments and corresponding rubrics and they will be asked to rate each item of the instruments with the following criteria: (1) representativeness of the content domain; (2) clarity of the item; and (3) comprehensiveness of the measure. Rubrics will be reviewed for percentage of exact agreement and adjacent agreement for each rubric item. Also, a content validity ratio will be calculated using the following formula: CVR = [(E - (N/ 2)) / (N/ 2)] where N stands for the total number of experts and E stands for the number who rated the object as essential.

For **construct validity**, the ratings that are given to student teachers on the evaluation instruments will be used to conduct a factor analysis. The factor analysis will provide information to answer the following questions: 1) Is the instrument measuring multiple constructs? And 2) Are the items on the instrument correlated amongst themselves? This information can be used to help the developers eliminate poorly written items on the instrument.

For **concurrent validity**, we will use a multiple regression analysis to see if there are correlations between students’ performance on the student teaching evaluation instruments and other performance and knowledge assessments that could be potentially related to the evaluation instruments under investigation. First, we will conduct an analysis to see if there is a correlation between students’ edTPA scores and ratings on the student teaching instrument. The edTPA is a subject area-specific performance-based assessment that is designed to measure the five dimensions of teaching (Planning, Instruction, Assessment, analysis of teaching effectiveness, and academic language development). The edTPA is currently required at OSU as a program completion requirement. Next, we will conduct an analysis to see if there is a correlation between ratings on the student teaching evaluation instruments and students’ licensure test scores. The analysis will consider if a test is taken multiple times before passing. The Ohio Assessment of Educators (OAE) tests are required to obtain a license from the Ohio Department of Education. The OAE tests include content specific tests and pedagogy tests. In general, all initial teacher licensure candidates must take one OAE pedagogy test and one OAE content test. Finally, we will conduct an analysis to determine if there is a correlation between students’ final grade for student teaching and ratings on the student teaching evaluation instruments. The Office of Educator Preparation collects and analyzes all of this data for OSU students currently as part of accreditation requirements; therefore, we are not asking permission to access data that we do not already currently have access at OSU. The other IHE’s have access to similar data for their own students; therefore, they are not asking to access anything new.

**Reliability Studies:**

For **test-retest reliability**, a statistical analysis will be conducted to examine if a correlation exists between the ratings students receive on the evaluation instruments at midterm (test) and the final (retest). Each student teacher receives ratings on these forms as part of a midterm and final exam requirements for student teaching.

For the **internal consistency reliability**, an analysis will be conducted to determine if the observer consistently rates items within the evaluation instruments. A correlational design using Cronbach’s Alpha test and/or Kuder-Richardson test will be used to measure internal consistency.

Finally, for the **inter-rater reliability** portion of the study, we will train all supervisors to use the student teaching evaluation instruments. Next, we will recruit supervisors who are willing to participate in the inter-rater reliability portion of the study. From the potential participants, we will select a representative sample of 10% of the total. Within this representative sample, we will ensure that various preparation programs are represented. These supervisors will complete student teaching evaluation instruments for inter-rater reliability purposes. Specifically, the supervisors will agree to observe an additional student teacher which includes observing the student teacher a total of 3 times and complete the student teaching evaluation instruments at the end of the semester. After the forms have been completed, we will calculate the percentage of exact agreement and adjacent agreement for each item on the evaluation instruments.

All potential participants will be informed of the purposes of the study and must sign a consent form to be included within the study. The participants will include experts for content validity, supervisors for inter-rater reliability, and student teachers who will be observed by inter-rater reliability supervisors. Participants will be informed that they can leave the study at any time with no penalty or consequences. We will provide two copies of the informed consent form, one for our records and one for the participants to keep. We do not need to obtain consent for the other analyses because we are requesting to do a secondary analysis of data we already collect, manage, and analyze. The duration of the study will vary per component. For content validity, the participants will dedicate a maximum of 8-12 hours over the duration of 2-3 months. The inter-rater reliability supervisors and student teachers who are being observed will be part of the study for the duration of one semester. The inter-rater reliability supervisor (secondary supervisor) will only need to commit an additional 8-12 hours of work for this study. The primary supervisors and the student teachers will not be required to commit any additional time for the purposes of this study because we are collecting reliability data; therefore, the expectations and time commit for them will remain the same. The secondary supervisor will receive a stipend to pay for gas and as compensation for the additional time spent.

The Ohio State Office of Educator Preparation is currently responsible for collecting, managing, and analyzing results on the student teaching evaluation instruments, edTPA, licensure tests, and student teaching grades for OSU students. All IHEs in Ohio routinely collect and use these data for accreditation purposes and to determine if candidates are eligible to be recommended for a teaching license by the state. In order to minimize risks and maintain confidentiality, after the database has been compiled, we will remove all student identifiers (such as names, OSU ID numbers, Tk20 identifiers, program, etc.) and assign a new identifier code for the analyses. Upon completion of data analyses, data will be reported in aggregate form. For students from other IHEs, data will arrive on our campus for analysis without student identifiers. Codes will be developed just for this project and will include an identification of the institution and a code for each student. The code identifiers will be housed at the originating institution and not provided to Ohio State. All reported data will be aggregated.

For the content validity portion of the study, expert ratings will be kept anonymous. The rubrics will not include the name of the expert who completed it and they will be asked to turn in the completed rubrics into a box at a statewide meeting. Researchers will leave the room when they submit their completed rubrics. Rubrics will be coded with a number for data analysis purposes. Finally, data obtained for the inter-rater reliability portion of the study will also be coded and identifiers will be removed to protect the identity of the supervisors, supervisor observers, and the student teachers who are being observed. All written products and records will be stored in a locked cabinet in the Office of Educator Preparation and only the primary researchers will have access to the raw data. All electronic records will be saved to the Assistant Dean’s drive in a folder that only the research team can access. We will work with OTEL to create this secure location and to ensure that access is only given when the appropriate steps have been taken. The products and records will be stored in these secure locations for five years.

The potential risks involved for participants include discomfort due to the additional time requirement and breach of confidentiality. We will address these potential risks by informing the participants they can leave the study at any time and by de-identifying the data and coding the data for analysis. Another potential risk is that the student teacher may be uncomfortable with having two supervisors observe him or her.. We will minimize this risk by obtaining consent from the student teacher and informing them that they can choose to leave the study at any time with no consequences. The supervisors are also in a position of authority over the student teachers; therefore, it is important to note that only the primary supervisor is responsible for the formal evaluation of the student teacher and the observer supervisor data will not be shared with anyone except the research team. If any of the participants feel uncomfortable, they can choose to leave the study and resume with the typical supervision model.

One anticipated benefit for the content validity participants includes the ability for the experts to have their voices heard on student teaching evaluation instruments that may be used beyond OSU. The evaluation instruments in this project are being developed for the new Council for Accreditation of Educator Preparation (CAEP) requirements. Our team plans to share the evaluation instruments with other educator preparation programs after we establish acceptable validity and reliability. A potential benefit to the student teachers in the inter-rater reliability portion of the study is that they will receive feedback and support from an additional supervisor, which can ultimately increase their success during student teaching. A potential benefit for the supervisors of this study is that they can gain a deeper understanding of the supervision process through additional practice opportunities.

## Internal and External Validity

**Threats to internal / external validity should be considered. Describe measures that have been taken to avoid study bias.**

Given that we are conducting correlational research and not experimental research, there are no threats to internal validity that need to be addressed. This is essentially a business as usual study examining student teaching assessments, and there is no additional intervention. We will be using the data for all student teachers for several portions of this study; therefore, our sample is the population, which minimizes threats to external validity for those portions of the study. For the content validity portion of the study, we have minimized threats to external validity by recruiting content experts from a statewide committee. This committee is representative of the educator preparation programs throughout the state (small vs. large, private vs. public, urban vs. rural). In order to minimize threats to external validity for the inter-rater reliability portion of the study, we will collect inter-rater reliability data for 10-20% of the sample, which is standard in these types of studies. Additionally, we will select supervisors from various program areas in order to increase representativeness within the supervisor group selected for inter-rater reliability and minimize threats to external validity

## Data Analysis

**Specify the analytic techniques the researcher will use to answer the study questions. Indicate the statistical procedures (e.g. specific descriptive or inferential tests) that will be used and why the procedures are appropriate. For qualitative data, specify the proposed analytic approaches.**

The analyses will consist of basic descriptive statistics, correlations, factor analyses, inter-rater agreement, content validity ratios, and internal consistency measures of reliability (coefficient alpha and Kuder-Richardson). These are the gold standard methods for assessing reliability and validity

# Bibliography

Include a reference list of literature cited to support the protocol statement.

**Appendices**

Appendix A: Student Teaching Evaluation Instrument: Pedagogy

Appendix B: Student Teaching Evaluation Instrument: Dispositions

Appendix C: Addendum to Student Teaching Evaluation Instrument: Mathematics

Appendix D: Addendum to Student Teaching Evaluation Instrument: Foreign Language (year 2 in study)

Appendix E: Addendum to Student Teaching Evaluation Instrument: Science (year 2 in study)

Appendix F: VARI-EP Validity of Instrument Rubric