**Validity and Reliability of Assessments**

**The EPP’s Review and Plan**

The EPP recognizes the need to establish validity and reliability on all assessments used in the initial and advanced programs. These assessments include:

* Key assessments used by the EPP throughout the Pathway to Teaching
* Clinical evaluations
* Performance assessments
* Course embedded assessments

In addition to these assessments, other EPP surveys/documents are reviewed.

To begin the examination of these assessments, the EPP conducted a detailed audit of the key assessments as they align to the CAEP standards and InTASC standards. The purpose of this audit was to provide evidence that the EPP has measures that “monitor candidate progress, completer achievements, and operational effectiveness aligned with all CAEP standards.” Much of this was accomplished through the development of the EPP’s Quality Assurance System. The information in this document presents a review of the key assessments and the EPP’s process and plan for establishing the validity and reliability of key assessments.

**Review of Assessments**

The EPP started by identifying assessments that were used at each of the transition points on the Pathway to Teaching. Originally, the EPP had four transition points that were used to monitor candidate progress and completer achievements: admission to the teacher education program, admission to student teaching, completion of the program, and follow-up surveys. With the transition to the yearlong residency model for elementary and secondary majors in the undergraduate program, the EPP added a fifth transition point which allowed for closer monitoring of candidates after admission to the program and before admission to student teaching. (See Attachment A: Key Assessment Matrix)

Next, the EPP discussed the assessments that were required by the WVDE in Policy 5100, Approval of Educator Preparation Programs. The WVDE requires that candidates successfully complete the Praxis Core assessments in reading, writing and mathematics prior to admission to the teacher education program. Additionally, the WVDE requires that candidates have successfully completed the Praxis PLT and Praxis II subject assessments within their content areas prior to student teaching. Praxis tests are accepted by CAEP as valid and reliable assessments of candidate knowledge.

The WVDE also requires that during student teaching, candidates are evaluated through a performance assessment based upon the WV Professional Teaching Standards. A review of the EPP’s student teacher final evaluation found that this performance assessment did not meet CAEP sufficiency for EPP created assessments. During fall 2018 the West Virginia Evaluation Rubric for Teachers (WVERT) a proprietary instrument developed by the WVDE was piloted for student teachers. The WVERT was used during spring 2019; however the EPP had also decided to pilot the North Dakota Student Teacher Observation Tool (STOT) with a group of student teachers during spring 2019. The STOT is a proprietary instrument of the North Dakota AACTE. The EPP plans to continue using both the WVERT and the STOT during fall 2019 and spring 2020. The West Virginia Teacher Performance Assessment (WVTPA) is also completed by student teachers. The WVTPA is considered proprietary according to the WVDE.

During fall 2017, the EPP explored means for assessing candidate dispositions. Up to this point, the GRIT scale had been used to determine candidate dispositions, but the EPP did not feel that this met CAEP requirements. The EPP developed a disposition assessment following the use of a Q-sort of components necessary for inclusion in the disposition assessment. Once the components had been selected, the list was shared with a group of 24 experts (EPP faculty and P-12 teachers and administrators) who selected the components that were most important according to Lawshe’s Content Validity Ratio. Following this, the rubric was prepared and presented to the expert group for approval. The EPP created Disposition Assessment was implemented during the spring 2018 and fall 2018 semesters. Believing that the EPP created Disposition Assessment could be improved, the EPP explored the use of Watermark’s Educator Disposition Assessment (EDA), a proprietary instrument, during fall 2018. Faculty completed calibration training in January 2019 and the EDA was used during this semester for candidates applying for admission to the teacher education program.

The process for developing the EPP’s Disposition Assessment also served the purpose for revising the requirements for the Teacher Candidate Portfolio (TCP) required for admission to the teacher education program. The TCP was designed to reflect dispositions as identified as essential. Candidates provide artifacts documenting their belief that all students can learn, their ability to be teacher leaders, their openness to change and new ideas, their use of reflection for professional growth, and their positive response to feedback for professional growth. The TCP was piloted during the spring 2018 semester, and has continued to be used as a way to assess candidate dispositions. During spring 2019, the EPP engaged in a study of inter-rater reliability of the TCP and plans to continue this during the 2019-20 academic year.

Field experience evaluation tools were reviewed and, based upon response from mentor teachers who completed these evaluations, these evaluations were revised completely. During spring 2019, the EPP and the Concord-Teacher Educator Council (C-TEC) collaborated to create new field experience evaluation tools for Level I, II, III, and IV field experiences. Using the ten InTASC Standards, the experts determined which InTASC Standards were most appropriate for the requirements of each field experience level. The group also examined the STOT, and the rubric used in the STOT was approved to be used corresponding to the InTASC Standard assessed at each level. The new field experience evaluations will be piloted during fall 2019.

The EPP uses a variety of surveys for monitoring completers. Two EPP created surveys seek feedback from completers and supervisors. Three surveys from the Network for Excellence in Teaching (NExT) – the Exit survey, the Transition to Teaching Survey, and the Supervisor Survey – provide extended data about the achievements of completers and the impact they are having in the classroom. To strengthen this, the EPP developed a plan for conducting case studies of completers that involves a three year follow-up. The beginning phase of this case study was piloted during the 2017-18 academic year.

By the completion of the spring 2019 semester, all key assessments had been reviewed. During this process that began during fall 2017, key assessments had been revised, updated, replaced and/or planned. The goal of this process has been to make these key assessments measure what is intended to be measured, and to be sure that each assessment allows the EPP to make appropriate inferences regarding a candidates’ knowledge, skills, and dispositions. Each assessment and corresponding rubric has been examined in an effort to define in very specific terms what should be addressed & assessed. The EPP has attempted to align key assessments with the program’s mission and goals, InTASC standards, and CAEP standards. The process of reviewing assessments is not complete. The EPP has developed a Quality Assurance System that will allow all data to be collected, aggregated and disaggregated, disseminated, and analyzed for improving the teacher education program and for monitoring candidate performance. The process will involve a pilot semester of the assessment. At the end of the pilot, the expert group will review the assessment and make any necessary changes. The instrument will then fall into the 3-year cycle of EPP assessments and be revisited when the cycle comes around again.

**Key Assessment Validity**

In a CAEP presentation on establishing content validity, Dr. Stevie Chepko defined content validity as “the extent to which a measure represents all facets of a given construct and the extent to which an indicator measures what it was designed to measure.” Content validity involves obtaining a level of agreement among “experts,” in this case, EPP faculty members, EPP clinical educators, and P-12 faculty and administrators, the use of Q-sort methodology, and Lawshe’s Content Validity Ratio. The EPP completed two validity studies. The results of these are presented.

**Examples of assessments that have been evaluated for validity:**

EPP Created Disposition Assessment and Teacher Candidate Portfolio

The Lawshe study was conducted with the EPP and the Teacher Advisory Council on September 20, 2017. Participants were given the following instructions: The following items have been identified as key factors that could be included in the assessment of initial candidates; dispositions. We are asking you to identify what are the most important elements that should be included in the Disposition Assessment and for components of the Teacher Candidate Portfolio for Admission to the Teacher Education Program. Based on your current understandings of the requirements for our candidates during their first course and field experience, and of the items on the list, please identify each individual item as either : Essential; Useful but not essential; or Not necessary. The following chart identifies the results of the study

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dispositions Validity Study Lawshe** | | | | | |
| **N=24 (17 Essentials needed)** | **Essential** | **Useful** | **Not Necessary** | **% Marked Essential** | **CVR .42**  **Needed** |
| The teacher candidate demonstrates a strong knowledge of content. | 4 | 20 | 0 | 17 | -0.67 |
| The teacher candidate designs instruction aligned with state standards. | 6 | 18 | 0 | 25 | -0.50 |
| The teacher candidate has a professional appearance. | 19 | 4 | 1 | 79 | 0.58 |
| The teacher candidate assesses student learning | 6 | 1 | 17 | 25 | -0.50 |
| The teacher candidate demonstrates the implementation of technology. | 6 | 10 | 8 | 25 | -0.50 |
| The teacher candidate demonstrates regular attendance and punctuality. | 24 | 0 | 0 | 100 | 1.00 |
| The teacher candidate is prepared for his/her daily responsibilities. | 24 | 0 | 0 | 100 | 1.00 |
| The teacher candidate demonstrates ethical behavior regarding fairness and confidentiality. | 23 | 1 | 0 | 96 | 0.92 |
| The teacher candidate demonstrates leadership qualities | 17 | 4 | 3 | 71 | 0.42 |
| The teacher candidate knows and supports the school mission. | 5 | 2 | 17 | 21 | -0.58 |
| The teacher candidate has a positive response to feedback. | 23 | 1 | 0 | 96 | 0.92 |
| The teacher candidate demonstrates good and effective communication skills. | 21 | 3 | 0 | 88 | 0.75 |
| The teacher candidate demonstrates an understanding of learner growth and development. | 6 | 14 | 4 | 25 | -0.50 |
| The teacher candidate reflects upon his/her performance. | 18 | 6 | 0 | 75 | -0.75 |
| The teacher candidate implements a variety of instructional strategies. | 3 | 17 | 4 | 13 | 0.83 |
| The teacher candidate values and respects diversity | 22 | 2 | 0 | 92 | 0.42 |
| The teacher candidate demonstrates an understanding of current research in education. | 2 | 6 | 16 | 8 | -0.83 |
| The teacher candidate collaborates and is engaged in learning with others. | 17 | 7 | 0 | 71 | 0.42 |
| The teacher candidate applies pedagogical knowledge to teaching. | 3 | 18 | 3 | 13 | -0.75 |
| The teacher candidate believes that all students can learn. | 22 | 2 | 0 | 92 | 0.83 |

Field Experience Evaluations

The Lawshe study was conducted with EPP faculty and Teacher/Principal Advisory Council members on March 1, 2019. The participants were divided into four groups, one each for Level 1, Level 2, Level 3, and Level 4 field experiences. Each group was provided a copy of InTASC standards, WV Professional Teaching Standards, and the STOT, along with information regarding the requirements for each field experience. Groups were asked to identify which InTASC standard was Essential; Useful but not essential; or Not necessary. The following chart identifies the results of the study.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **InTASC Validity Study for Field Experience Evaluations - Lawshe**  **EPP Faculty and Concord Teacher Advisory Council** | | | | | | |
| **N=49 (For a panel of 40, 26 Essentials are needed or a CVR of** | **Level 1 Standards Identified as Essential**  **(100%)** | **Level II Standards Identified as Essential**  **(100%)** | **Level III Standards Identified as Essential**  **(100%)** | **Level IV Standards Identified as Essential**  **(100%)** | **CVR** |
| **Standard #1: Learner Development**. The teacher understands how children learn and develop, recognizing that patterns of learning and development vary individually within and across the cognitive, linguistic, social, emotional, and physical areas, and designs and implements developmentally appropriate and challenging learning experiences. | **X** | **X** | **X** | **X** | **1** |
| **Standard #2: Learning Differences**. The teacher uses understanding of individual differences and diverse communities to ensure inclusive learning environments that allow each learner to meet high standards. | **X** | **X** | **X** | **X** | **1** |
| **Standard #3: Learning Environments**. The teacher works with learners to create environments that support individual and collaborative learning and that encourage positive social interaction, active engagement in learning, and self-motivation. | **X** | **X** | **X** | **X** | **1** |
| **Standard #4: Content Knowledge**. The teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) he or she teaches and creates learning experiences that make these aspects of the discipline accessible and meaningful for learners to assure mastery of the content. |  | **X** | **X** | **X** | **1** |
| **Standard #5: Applications of Content**. The teacher understands how to connect concepts and use differing perspectives to engage learners in critical/creative thinking and collaborative problem solving related to authentic local and global issues. |  |  | **X** | **X** | **1** |
| **Standard #6: Assessment**. The teacher understands and uses multiple methods of assessment to engage learners in their own growth, to monitor learner progress, and to guide the teachers’ and learner’s decision making. |  | **X** | **X** | **X** | **1** |
| **Standard #7: Planning for Instruction**. The teacher plans instruction that supports every student in meeting rigorous learning goals by drawing upon knowledge of content areas, cross-disciplinary skills, and pedagogy, as well as knowledge of learners and the community context. |  | **X** | **X** | **X** | **1** |
| **Standard #8: Instructional Strategies.** The teacher understands and uses a variety of instructional strategies to encourage learners to develop deep understanding of content areas and their connections, and to build skills to apply knowledge in meaningful ways. |  |  | **X** | **X** | **1** |
| **Standard #9: Professional Learning and Ethical Practice.** The teacher engages in ongoing professional learning and uses evidence to continually evaluate his/her practice, particularly the effects of his/her choices and actions on others (learners, families, and other professionals, and the learning community), and adapts practice to meet the needs of each learner. | **X** | **X** | **X** | **X** | **1** |
| **Standard #10: Leadership and Collaboration.** The teacher seeks appropriate leadership roles and opportunities to take responsibility for student learning, to collaborate with learners, families, colleagues, other professionals, and community members to ensure learner growth, and to advance the profession. |  |  |  | **X** | **1** |

**Key Assessment Reliability**

Reliability refers to the ability of an assessment to measure candidate characteristics or knowledge consistently, For the EPP, reliability involves at least two actions. First, EPP created assessments can be made more reliable and have less measurement error by making sure that they have been written with clear directions so candidates know exactly what is expected of them from the assessment. Second, Faculty must be sure to have appropriate questions that clearly measure the content taught and are not confusing students with the wording or biased to any subgroup of students. It would also be desirable that faculty have gathered appropriate feedback from colleagues and/or candidates who have read through the assessment or taken the assessment to provide feedback to the assessment’s clarity.

Inter-rater Reliability (IRR) is the degree to which two or more raters obtain the same results when using the same instrument/criteria for evaluation. This is the primary method the EPP will use to measure the reliability of its assessments as it addresses the consistency of the assessment implementation methods. To achieve IRR, the EPP has selected to use to use inter-rater agreement (IAA), which is expressed as a percentage of exact agreement among reviewers, and the goal is to reach at least the 70% agreement threshold (Jonsson & Svingby, 2007). IAA is a simple calculation determined by counting the number of exact agreements on rating variables on any part or whole of an assessment and then dividing this by the number of total variables (Among 10 ratings, reviewer 1 and reviewer 2 agreed on 8 of them, this would then give you 80% agreement among the two reviewers).

The following is the basic set of guidelines for IRR training sessions used to promote objective and consistent evaluation of candidate performance. (These steps are taken from: Graham, M., Milanowski, A., & Miller, J. (2012). Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Retrieved from http://www.tifcommunity.org)

1. Provide a process overview to give the reviewers the big picture.

Participants are introduced to reliability, what it means to have a reliable measure and why it is important. Associated assignments are reviewed. In addition, participants discuss how rubric results are used to provide feedback to students and instructors. For students, these ratings serve as indicators of performance. For instructors, the results from the training sessions are used as part of the continuous improvement process to actively track the reliability of their ratings.

2. Explain the rating dimensions (standards of performance & rubrics). This is to help the raters become (more) familiar with the rubrics. Each dimension on the rubric is defined, and what each rating represents and means is explained. Participants discuss how they interpret the ratings; it is emphasized that if each participant has distinct understandings and interpretations, it becomes difficult to evaluate how consistent ratings are between participants.

3. Help raters put aside biases. Common biases are identified, and strategies to support objective ratings are discussed.

4. Explain common rater errors to avoid. We discuss common rater errors such as leniency rating, confirmation bias, similarity ratings, and halo, and participants discuss strategies to work through them.

5. Describe the process for decision-making. Raters develop objective strategies for making ratings, including note taking, sticking to the rating scale, and rating on each dimension separately.

6. Participants practice connecting evidence to performance dimensions, with feedback provided. Raters pull out evidence from a provided example. The raters discuss their findings and receive feedback.

7. Participants practice interpreting the rubrics. For this step, raters work through the language presented in the rubric rating scales. Words will be explored to better understand how different raters interpret the words selected in the ratings. Prototype examples are shared, which raters use for practice. Discussion follows this step.

8. Rater training will be concluded with independent ratings on common examples, which are documented for reliability via adjacent agreement.

**Examples of assessments that have been evaluated for reliability:**

Teacher Candidate Portfolio Rubric

The EPP has worked with reliability on the Teacher Candidate Portfolio which was piloted spring 2018. The EPP has plans to continue work on key assessment reliability as they are piloted.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **INTER-RATER RELIABILITY - PERCENTAGE AGREEMENT** | | | | | | |
|  | **RATER 1** | **RATER 2** | **RATER 3** | **RATER 4** | **RATER 5** | **INTER-RATER RELIABILITY** |
| **BELIEF THAT ALL STUDENTS CAN LEARN** | 2 | 2 | 2 | 2 | 2 | 100.00% |
| **LEADERSHIP** | 3 | 3 | 2 | 3 | 3 | 60.00% |
| **OPEN TO CHANGE & NEW IDEAS** | 2 | 2 | 2 | 2 | 2 | 100.00% |
| **REFLECTION: DESCRIPTION OF EXPERIENCE** | 2 | 2 | 2 | 2 | 2 | 100.00% |
| **REFLECTION: CORRELATION TO STANDARD** | 2 | 2 | 3 | 2 | 2 | 60.00% |
| **REFLECTION: RELATIONSHIP TO FUTURE TEACHING** | 2 | 2 | 2 | 2 | 2 | 100.00% |
| **RESPONSE TO FEEDBACK** | 3 | 2 | 3 | 3 | 2 | 40.00% |
| **VOLUNTEER HOURS** | 3 | 3 | 3 | 2 | 3 | 70.00% |
| **TOTAL SCORE** | 19 | 18 | 19 | 18 | 18 |  |
| **AVG SCORE** | 2.375 | 2.25 | 2.375 | 2.25 | 2.25 | **78.75%** |

West Virginia Teacher Performance Assessment (WVTPA):

Since the development of the WVTPA, each IHE that participates is required to provide training to reviewers in assessing the WVTPA. For each version of the WVTPA that has been issued (we are now at 4.0), the EPP has conducted inter-rater reliability training. A sample WVTPA from a previous semester is used, and each EPP faculty participant scores the WVTPA using the current rubric. Prior to the review, the group discussed the tasks, the prompts for each task, and the rubric indicators for each prompt. Following the individual ratings, participants shared scores and discussed why the score was chosen. Plickers were used to tabulate the responses and the inter-rater reliability was established.

**A Process for Validity and Reliability**

The EPP has developed a Quality Assurance System that will allow all data to be collected, aggregated and disaggregated, disseminated, and analyzed for improving the teacher education program and for monitoring candidate performance. The process will involve a pilot semester of the assessment. At the end of the pilot, the expert group will review the assessment and make any necessary changes. The instrument will then fall into the 3-year cycle of EPP assessments and be revisited when the cycle comes around again.

2019-20 Assessment Review Cycle to Establish Validity and Reliability:

Key Assessments

Educator Disposition Assessment (EDA)

* Collect Data : Select three candidates from EDUC 210/EDSP 303, two candidates from the elementary block, and two candidates from the secondary block and have the EDA completed by two course professors for each candidate.
* Complete Inter-Rater Reliability: EDUC 210 and EDSP 303 course professor ratings will be reviewed for inter-rater agreement at a minimum of 80%.
* Present findings to the EPP and stakeholders.

Student Teacher Observation Tool (STOT)

* Training: Mentor teachers and university supervisors will be trained in the use of the STOT.
* Collect Data : Select ten student teachers. The mentor teacher and the university supervisor will both observe a lesson and evaluate the student teacher using the STOT.
* Complete Inter-Rater Reliability: The mentor teacher and university supervisor lesson observation ratings on the STOT will be reviewed for inter-rater agreement at a minimum of 80%.
* Present findings to the EPP and stakeholders.

TPACK Technology Integration Observation Tool (TIOT)

* Training: Mentor teachers and university supervisors will be trained in the use of the TIOT.
* Collect Data : Select ten student teachers. The mentor teacher and the university supervisor will both observe a technology lesson and evaluate the student teacher using the TIOT.
* Complete Inter-Rater Reliability: The mentor teacher and university supervisor lesson observation ratings on the TOIT will be reviewed for inter-rater agreement at a minimum of 80%.
* Present findings to the EPP and stakeholders.

WV Evaluation Rubric for Teachers (WVERT)

* Training: Mentor teachers and university supervisors will be trained in the use of the WVERT.
* Collect Data : Select ten student teachers. The mentor teacher and the university supervisor will both observe a lesson and evaluate the student teacher using the WVERT.
* Complete Inter-Rater Reliability: The mentor teacher and university supervisor lesson observation ratings on the WVERT will be reviewed for inter-rater agreement at a minimum of 80%.
* Present findings to the EPP and stakeholders.