Measuring a Moving Target: Validating Deployed Training Courses

Timothy R. Brock, PhD, CPT, ID(S&L+)
General Dynamics Information Technology
Orlando, FL
Timothy.Brock@gdit.com

Denise R. Stevens, EdD
General Dynamics Information Technology
Orlando, FL
Denise.Stevens@gdit.com

ABSTRACT

The Veteran Benefits Administration implemented a new requirement to validate the effectiveness of new or revised e-learning courseware after deploying it to the field using the total population as samples of the target audience. In the past, the validation effort occurred in a controlled environment using a small sample of the population prior to fielding the course. The methodologies used with the small sample population were the U.S. Army's Sequential Validation or Fixed Validation. Because of the push to deploy the required entry-level and refresher/recurring training quicker (as well as cheaper without sacrificing quality), course effectiveness validation is now conducted post-deployment. This mandate poses several challenges, one of which is how to determine whether a training course is effective when it is deployed to the field and completed by government employees expected to simultaneously meet their fast-paced daily production requirements in a high-stress work environment. This paper reports how an argument-based approach is being assessed as an alternative courseware validation process that provides practical evidence to allow reasoned, data-driven interpretations and conclusions regarding the effectiveness of a deployed course. The approach uses both qualitative and quantitative data to establish reasoned arguments to make the evidence-based interpretations of the data. This paper discusses how this argument-based framework for measuring, analyzing, and reporting validation results is evolving to make reasoned determinations about the effectiveness of deployed e-learning products conducted in uncontrolled work environments.

ABOUT THE AUTHORS

Dr. Timothy R. Brock is a Principal Human Performance Technologist with General Dynamics Information Technology. Dr. Brock is also the CEO of The Institute 4 Worthy Performance, an Associate of the ROI Institute, and a Practice Leader with The Institute for Performance Improvement. Before retiring from Lockheed Martin in 2012, he led their Global Training and Logistics Science of Learning and Performance Improvement team. During his Air Force career, Dr. Brock was responsible for the weapon system curriculum and high fidelity Missile Procedures Trainer simulation scenarios for all five of the Air Force's ICBM initial qualification courses. Dr. Brock is a Certified Performance Technologist (CPT) through the International Society for Performance Improvement (ISPI) and a Certified Instructional Designer with a specialization in Simulation and Labs (ID(S&L+)) through The Institute for Performance Improvement. He co-authored a chapter titled "Simulation Operations, Curriculum Integration, and Performance Improvement" in the book *Healthcare Simulation: A Guide for Simulation Specialists*. He holds a Doctorate degree from Capella University in Education with a specialization in Training and Performance Improvement. Dr. Brock is an Adjunct Professor at Franklin University and is a member of the Advisory Board for Full Sail University's Masters of Science degree in Instructional Design and Technology.

Dr. Denise R. Stevens is the Chief Learning Officer for General Dynamics Information Technology's Training and Simulation Sector. She has over twenty-six years of experience in the application of all aspects of the Instructional Systems Design process and Human Performance Technology in applied research and development for Government and education. Dr. Stevens has extensive experience in the design of Training and Performance Support Systems and Job Performance Measures. Dr. Stevens has been involved in the large-scaled training design and development efforts resulting in over eight national awards. She has been involved with conducting cognitive and behavioral job-task analysis, learning objective development, instructional and performance-centered design for various training platforms, such as web-based, classroom-based, or blended deliveries, conducting test reliability and validity procedures, conducting individual and small group trails and sequential validation procedures. Dr. Stevens also participated in pioneering the first extensive foreign language training program with voice recognition capability. She has published work on cultural and linguistic diversity in American schools as well as various Government publications on cost and training effectiveness analysis. Dr. Stevens has over ten years of experience as an Adjunct Professor of Foreign Languages and is currently an Adjunct Professor at the Department of Instructional Design and Technology Master's Program at the University of Central Florida.

Measuring a Moving Target: Validating Deployed Training Courses

Timothy R. Brock, PhD, CPT, ID(S&L+)
General Dynamics Information Technology
Orlando, FL
Timothy.Brock@gdit.com

Denise R. Stevens, EdD
General Dynamics Information Technology
Orlando, FL
Denise.Stevens@gdit.com

AN INTRODUCTION TO THE CHALLENGE

A program management truism is "you can have a product fast, cheap, and good—pick two." This time-price-quality equation is encountering increased resistance from customers who not only say they now want all three, but also want deliverables faster, cheaper, and just as good. As a result, customer mandates require us to explore and adapt different and foreign principles and practices beyond our current, comfortable best practices. Those who create training solutions have devised alternatives to accommodate customer demands for the accelerated deployment of quality products for less money. A common accommodation is to validate course effectiveness during the first external offering to the target population rather than internally to a small group that represents the target population. Results from either approach with a limited target population are typically minor development edits or design adjustments after the first three phases of the ADDIE process (Analyze, Design, Develop, Implement, and Evaluate) are accomplished (U.S. Department of Defense, 2001).

What do you do when a customer assumes the "ADD" phases will produce an 80% good-enough e-learning solution to meet their ever-increasing demand for faster deployments and mandate skipping the course effectiveness validation phase prior to enterprise-wide rollout? The only viable alternatives are to (1) refuse to do the work, (2) try to negotiate returning to a pre-deployment validation solution, or (3) find a reasonable alternative validation methodology based on a literature review to learn what other practice-proven options are used by other learning and assessment professionals.

Even more challenging, in most instances, customer-provided post-deployment validation participants are not dedicated full-time to the validation study and complete the training at their workstations while simultaneously trying to meet their daily productivity requirements. To consider a course valid, the initial standard required 80% of the target population who fail the pretest to then pass the first or second variant of the posttest (with a score of 80% or higher). Now, new directives require a higher robust testing threshold where training and testing expects learners to meet an extremely high performance standard (100% passing criterion). These directives are coupled with new testing strategies specifically designed to meet these new performance standard requirements. Not enough time is provided to conduct prototyping research or target population analysis to determine feasibility to credibly assess capabilities. Further, there is no agreement on what passing standard that a course must meet to determine that it is "instructionally sound and psychometrically pure" when conducting a validation study after deployment. How does one assure customers that a course is effective when less than 50% of the target population typically fails to pass the posttest courses on their first or second attempt? Is the root cause a deficient training course or do other workplace or motivation barriers prevent them from testing to standard? Maybe both are true. This is especially challenging when these government employees are expected to perform professionally, accurately, and rapidly on the job at a 97% proficiency level while at the same time are expected to score 100% on a course posttest that can take over an hour to pass, not counting the pretest and instructional unit they must pass after failing the pretest.

This increases the challenge of interpreting test scores as an indicator of a course's effectiveness after assessing a test taker's performance completing an assessed task, within a specified situation, and under certain conditions using a pretest and posttest. Test score results are also used to make a judgment about the test taker's level of mastery or other level of achievement as well as "some standing on a trait, or some probability of succeeding" in a given job, education program, or other activity (Kane, 2013, p. 1). According to Messick, a trait is "a relatively stable characteristic of a person…which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances" (1989, p. 15). A fundamental assumption is the test results reflect the test taker's best effort and is an accurate representation of their capabilities to perform a task to specified standards in defined conditions. If the learning and test taking conditions and standards change, the assumptions and

interpretations of test scores become suspect and can lead to questionable causal interpretations and claims about the test taker's traits and capabilities as well as the meanings of their test scores.

LEGACY APPROACH

Not all job functions are the same and not all job-tasks fit neatly into "hard-skill" or "soft-skill" categories. A vast grey area exists where some jobs are neither; this is the case for most of the jobs within the VBA, where most jobs are procedural in nature and cognitively demanding. While all processes rely on ensuring the application of laws and regulations, this can be subjective at times because one can interpret the laws and regulations slightly differently. Processing Veterans' caseloads is a very complex process. Employees must have the same level of dedication with each case file to analyze hundreds of pieces of evidence related to a Veteran's life and medical issues in order to make the best determination possible that will benefit the Veteran. Employees who are hired for these jobs have to be trained beginning the first day of employment on the multitude of processes that are unique to VBA. Over 18 years ago, VBA had the fortitude to insist the training offered employees be the best, providing measurable results and keeping backlogs and quality errors to a minimum. For this reason, VBA adopted a performance-based philosophy where training events mirrored the job at its highest fidelity. With the Training and Performance Support System program, which is still in use today, VBA employees practice job-tasks using simulated casefiles before taking assessments that are designed as job-performance measures. VBA applies a systematic development process to create training and assessments.

In the past, as part of this process, each course underwent several formative evaluation procedures prior to deployment, including external Subject-Matter Expert (SME) reviews, course test validity and reliability studies with a panel of SMEs, courseware small-group tryouts, and formal validation procedures prior to deployment.

Course effectiveness validation procedures were formal and used a sequential validation methodology because this method required a smaller sample size than the fixed method. The sequential validation procedure was conducted in accordance with TRADOC JA 350-70-10.6e (2004). Validation depends on using different plotting charts depending on specified numbers of passing posttest scores. For example, if the validation pool consists of 30 people taking the test, 26 must pass to have a 90% level of confidence in the effectiveness of the course. If 20 people take the test, 18 must pass. The level of criticality was usually determined during the test reliability effort when the SME panel rated each terminal learning objective (TLO) based on four different 4-point scales. This systematic development process took a long time because it required SME access to a sample of the target audience. Because the need to expedite the process and make the training available as soon as possible to the field, it was necessary to reevaluate the process. One major decision was to move the validation study to after a course has been deployed.

Today, VBA has an established Competency-based Training System where for each job position, a clear training path has been delineated from novice to journey-level. This system allows for the constant re-addressing of training needs and new instructional and testing design to ensure that each employee is trained on the latest changes in laws and regulations. In addition, VBA's systematic development process is more agile in nature and is able to produce training at a much faster pace.

Transition

This new process introduced changes that challenged the legacy approach. To understand the context and implications of this change, it is important to first understand what did not change.

The courseware development process did not change. First, a courseware development team develops the course content and tests with SME and/or Advanced Performer (AP) support to meet course objectives (which are developed and approved prior to development). Then the development team conducts vigorous course content review and test validity and reliability studies with a panel of SMEs/APs. Formative evaluation such as small-group tryouts are also conducted; therefore, VBA considers the courseware development process and the requirements established by their Standard Operating Procedures (SOP) as a satisfactory risk-mitigation strategy for not conducting an additional course validation study prior to deployment. The assumption is that the organization has confidence it is deploying effective courses prior to conducting any post-deployment course effectiveness validation studies.

In addition, the course effectiveness validation pass rate threshold remained unchanged. In contrast to the stated assumption about the effectiveness of the course prior to deployment, to consider the course effective, it is still expected that 80% of those who fail the pretest and complete the training course will pass the posttest within two attempts. In addition, it is still expected that a variation of the U.S. Army's Sequential Validation Methodology be used to determine whether a course is effective. This methodology establishes the effectiveness of a course based on the pass rate of 9 to 30 participants.

What changed were the definition of a test variant, establishing a validation pass rate, and the passing cutoff scores. The next three sections explain these changes.

Test Variant Definition

With previous test strategies, a test variant was easy to define. A course test included two variants that used a parallel form of cases. A test would include several cases with associated questions that the student had to "pass" individually to pass the entire test. The two tests were static in that, if the students were to fail both variants, they would be sent to a training coordinator for one-on-one remediation prior to attempting the same tests again. The course met the pass rate standard or it did not. If not, content traceability with cases and test item analysis occurred to verify the reliability and validity studies conducted with SMEs and/or APs before deployment. In addition, analyses of validation survey Likert Ratings and open-ended question responses were made to make course improvements that increased the pass rate until it met the validation pass rate standard.

For the new testing strategy, a test is comprised of a pool of cases where a student must pass a certain number of cases. The number of cases varies depending on the number of tasks covered in the course. Cases can also have a different number of questions depending on the complexity of the tasks. For example, for one package of courses for a veteran service representative position, the number of cases required to pass a test ranges from two to four and the number of questions for each case range from two to eight. To pass a case, students must answer all questions correctly. When they miss one question, they fail the case and are given another case. The question is, how many cases are they allowed to process to pass the required number of cases for each test?

Validation Pass Rate

When there were two variants of the same test, it was simple to determine a credible course pass rate. The validation pass rate standard was typically 80%, depending on the criticality of the task. This meant 80% of the participants taking the posttest had to do so during their first or second attempt using different variants of the same test.

When a new testing strategy was introduced that involved a randomized test bank of cases with outcome-based responses, they established the cutoff score for every case at 100%. The rationale was the test takers were expected to work error-free on their jobs, and the tests needed to reflect the same standard since they were currently doing the job. However, the policy that the participants can take as many cases as necessary to pass the minimum number remained unchanged. The unresolved dilemma is how many opportunities should a participant have to pass the minimum number of cases with a 100% score to meet a defined post-deployment validation criterion?

Different attempts at defining what was formerly called a variant have failed. Should the test taker get one extra case for each case missed? For example, if they must pass four cases, can they take up to eight cases to pass four to consider these two attempts for establishing a validation pass rate? Decision makers did not want to give them these many opportunities because they could fail four cases before passing four cases. In other words, they would pass the validation threshold with a score of 50%. However, even using this approach, courses rarely achieve the 80% pass rate for validation since there is no limit to the number of cases they can take to pass to receive credit for the course. This issue remains unresolved.

Passing Cutoff Score

Finally, the new testing standard is 100% for each case. In the past, passing scores represented typical cutoff scores. This new perfection threshold has created consternation for test maker and test taker alike. Cases require more of a principle-based approach to processing a case because judgment is required in varied situations due to each Veteran's situation being unique. Practices unique to different regional offices at different locations are applied in these workplaces to process claims. There is no room for error when creating the realistic, relevant cases, the outcome-based questions (to assess the test takers proficiency to process the case), or the rationale for each correct

and incorrect response (to satisfy the interpretations made by test takers reading the cases and selecting/providing the correct answers to the case test questions). Does the test accurately reflect the test takers capabilities or inhibit an accurate measure?

The default response to each of these issues continues to focus on fixing the courses if 80% of the test takers do not pass the course within a limited number of tests/cases. Our response is to fix programming discrepancies that are identified (typically corrected during pilots). Also considered are comments from the few students who voice dissatisfaction with the content and the new case-based approach/100% scoring standard. We also conduct detailed item analysis studies looking for faulty test questions or answer trends. We typically deem courses effective in that the content covered each learning objective and the cases and the associated test questions were directly linked to the course objectives through the course content. In addition, the course objectives, content, and test questions/responses were approved by SMEs/APs provided by the VBA. How does one establish credible evidence to prove or disprove this conclusion?

The Discovery

We became aware of Kane's argument-based approach while one of the authors attended the world's largest healthcare simulation conference and learned how the medical community uses it to validate the effectiveness of medical education and training courses, with and without simulation training devices. This framework helped reprogram our thinking about how we were approaching this course validation challenge. As a result, we adopted this approach and began to include evidence not considered in the past to interpret the meaning of the test results produced by the new testing strategy using this practice-proven, research-based validation methodology. This paper will first describe the elements of Kane's argument-based approach to provide context and then show how we applied it to determine if it is a credible means to validate the effectiveness of this customer's deployed e-learning courseware.

Kane's Argument-Based Approach to Validation

Test validity has traditionally been about the accurate and appropriate interpretations assigned to test scores and how those test scores are used (Sireci, 1998). It is based on evidence and the degree this evidence supports the interpretation of test scores (AERA, APA, & NCME, 1999). Therefore, validity is not about the properties of the assessment instrument, it is about the interpretations and uses of the interpretations of the test results (Downing, 2003). Since the emphasis is on the interpretation of evidence derived from test scores, Kane proposed the argument-based approach to establish a framework for collecting and communicating convincing validity evidence based on interpretive arguments. This includes the inferences and assumptions made—even the most questionable assumptions (1992). The argument-based approach provides a framework to evaluate the plausibility of the claims about and use of test results (Kane, 2013).

This argument-based approach to validation introduces a practice-driven, research-based methodology to interpret quantitative and qualitative arguments based on appropriate evidence to support plausible interpretations of test-score results (Kane, 1992). It "reflects the general practices of construct validity without requiring formal theories" (Kane, 2013, p. 9). It includes four phases, where an individual "(a) decides on the statements or decisions to be based on test scores, (b) specifies the inferences and assumptions leading from the test scores to these statements and decisions, (c) identifies potential competing interpretations, and (d) seeks evidence supporting inferences and assumptions in the proposed interpretative argument and refuting potential counterarguments" (Kane, 1992, p. 527). The evidence does not result in a decision about valid or invalid evidence since available evidence is typically not complete or trustworthy. Rather, validity is about the degree of plausibility of the evidence to support interpretations of the test score results since it is not possible to credibly prove any conclusions (Kane, 1992). Stated another way, "assessments are not valid or invalid; rather, the scores or outcomes of assessments have more or less evidence to support (or refute) a specific interpretation (such as passing or failing a course)" (Downing, 2003, p. 830).

Criterion for Evaluating Practical Arguments

Since this approach is argument-based, Kane (1992) advocates three general criteria for evaluating practical arguments (see Figure 1) based on research of validity (House, 1980) and reasoning (Toulmin, Reike, & Janik, 1979). The first criterion is to have a clearly stated argument. The argument provides the foundation for deriving conclusions and their underlying inferences and assumptions. The greater the detail, the greater the clarity. Without this, the following two criterions are meaningless. The second criterion is to state an argument that is coherent so assumptions can flow to logical and reasonable conclusions. Inferences are required along the logic chain that is supported by acceptable research and evidence-based best practices. The third criterion is plausibility of the assumptions used in the argument.

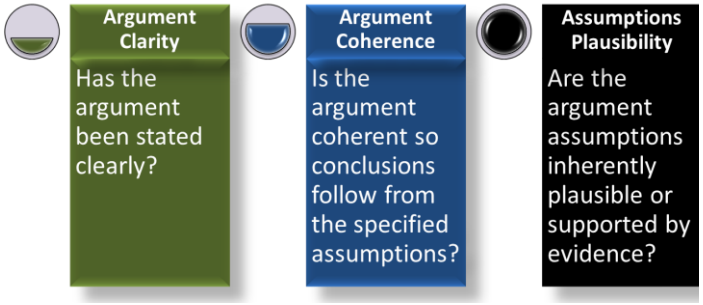

Figure 1. Criterion for Evaluating Practical Arguments

Does the evidence support the argument's assumptions? The plausible arguments may be self-evident, rely on empirical data (such as multiple failed falsification attempts), or emerge after careful documentation and scrutiny of procedures (Schuwirth & van der Vleutin, 2012). Therefore, the test of an assumption's plausibility is evidence that supports or contradicts it. A clear argument is required to identify the assumptions made when interpreting the evidence so counterarguments can emerge to test the most questionable assumptions.

Evaluating Interpretive Arguments

Kane's (1992) approach describes different inference categories to support the logic chain of interpretive or practical arguments (Figure 2). These categories include the associated assumptions about inferences and the possible evidence to support each one and their associated assumptions deemed plausible a priori or by evidence. Further, to strengthen the credibility of the interpretations, it is helpful to develop parallel lines of evidence, when possible, and to consider plausible counterarguments. Below is a brief explanation for three of the six categories relevant to this paper.

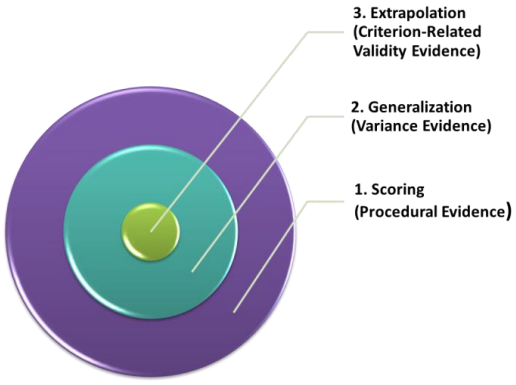

Figure 2. Logic Chain Inference Categories

1. **Scoring** – This is the most basic inference for interpreting test score results. It assumes the methods used to assign a score are consistent with the measurement procedure (Newton & Shaw, 2014). Therefore, the support evidence is procedural. If a standardized test is used, it assumes the procedures are followed exactly. In addition, the environment where the data is collected is also a consideration that affects test scores. If conditions are extreme, such as the temperature of the room, noise, distractions (internal or external), interruptions, etc., they may establish a plausible counter interpretation for low test scores. Procedural evidence is limited for establishing the plausibility of an interpretive argument, but it can be central to refuting the argument. Interpretive arguments are undermined if the procedures are not followed or are inadequate. For our purposes, scoring makes inferences regarding the evidence that the test was properly administered and that the scores were accurately derived and recorded. The focus is on the test setting and the confidence in the claim the score represents.

2. **Generalization** – This inference assumes that certain types of variances are not relevant to the interpretation of scores. Examples are the time and place of the test, to type of test used, and how the test was scored. This inference is supported by what Kane classifies as invariance laws where the conditions involved in the measurement are allowed to vary in certain dimensions without much impact on the outcomes. Reliability studies (Feldt & Brennan, 1989) or generalizability studies (Brennan, 1983) are used to support the assumptions

about invariance. Kane (1992) notes that "reliability is a necessary condition for validity because generalization is a key inference in interpretive arguments, but it is not a sufficient condition because generalization is not the only inference in the argument." Generalization assesses warrants and backing for assumptions that a test score represents a true score (Newton & Shaw, 2014). For our purpose, generalization considers the evidence about score reliability. This includes item/case sampling, test length, and score precision. The focus is on how reliable and reproducible the scores are considering different variables such as time of day/week generated, location, etc. that can influence test results.

3. **Extrapolation** – This is the argument used the most. Test score interpretations establish an indication about testing and nontesting behaviors. It assumes that a relationship between these variables exist and is understood. Empirical evidence can indicate a relationship between the test performance and nontest behavior. The inference uses criterion-related validity evidence to establish a direct link between test and nontest behavior. However, Newton and Shaw (2014) argue that this inferential process is susceptible to error. One example given is the artificial nature of the test situation itself. If students are poorly motivated, they are likely to underperform. Error is also introduced through test anxiety, content relevance, lack of consequences, and other variables that affect student performance during the testing (and learning) situation. For our purposes, extrapolation looks for evidence that the observations made about the test scores are relevant to the proficiency of the target audience measured by the test. The focus is on the state of the learners and how the test scores relate to real life performance.

These three inferences establish the primary framework for the measurement argument because they "explain why it is legitimate to interpret evidence from test performance in terms of the attribute that is supposedly being measured" (Newton & Shaw, 2014). From these three inferences, we interpret the data and make decisions. Two fundamental questions asked are, (1) where in this chain of inferences does evidence either support or refute the argument that the course is effective?, and (2) does the empirical evidence support a correlation inference that the learner who satisfies the 100% score standard on the posttest will perform at a 97% proficiency level on the job after the course?

Adaption of Kane's Approach to Meet the Course Effectiveness Validation Mandate

Our new approach consists of more clearly stating our arguments supported by plausible assumptions to derive credible conclusions. We have built our arguments around three of the Kirkpatrick (1994)/Phillips (1983) levels of evaluation:

- *Level 1 (Reaction/Reaction, Satisfaction, Planned Action)*
- *Level 2 (Learning)*
- *Level 3 (Behavior/Job Application/Implementation)*

For Level 1, our argument is that the participant reactions to and satisfaction with the course, and their planned action to use what was learned are indicators to the likelihood they were engaged to perform well on the posttest and predictors to their commitment to apply what they learned on the job. This opinion evidence uses qualitative data collected from end-of-course surveys, Likert Scale ratings, and other sources represented by Figure 3.

Figure 3. Level 1 Evidence

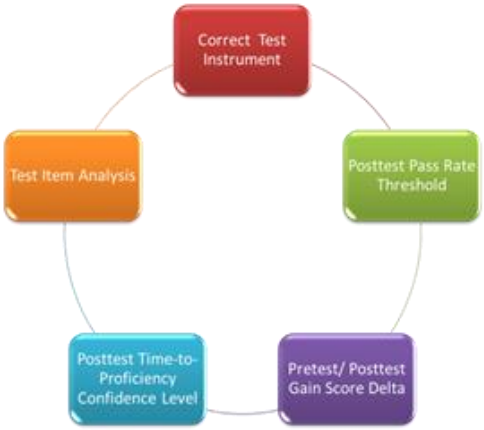

Figure 4. Level 2 Evidence

For Level 2, our arguments stem from five data sources reflected at Figure 4. The posttest pass rate threshold established by VBA remains 80%. We also conduct a test item analysis when enough students have taken the posttest to produce statistically sound results. That is difficult to do when you have at most 9 cases that are randomly seen by 9 to 30 students. We also learn about their opinions about the relevance and accuracy of the course content, from their perspectives, and how clearly the course unfolds to ensure they successfully learn the course content to pass the posttest. We also added a new data point called the posttest time-to-proficiency confidence level. Since we have not yet established a cutoff for how many cases an individual is allowed to fail before passing the

minimum number of cases to pass the course, we determine the number of cases required to reach the 80% pass rate. Then we assign a level of confidence regarding the time it took for the participants to demonstrate proficiency by passing the minimum number of cases. For example, if test takers are required to pass three cases, and 80% of those who took the posttest passed the first three cases, we state we have the highest confidence in the results. If it takes four cases to meet the 80% threshold, we have high confidence. The more cases it takes to meet the 80% threshold, the less confidence we have in the proficiencies of the target population.

None of these data sources differs from the data typically collected, so no additional costs were involved. What we did at Level 3 is different.

For Level 3, an industry standard for determining the effectiveness of a training course is the performance of course graduates on the job. If those who do well during training do well on the job, and those who do not perform at a lower level on the job, as predicted by the posttest results, the course is considered an effective and reliable predictor of workplace performance. It is not known at this time whether there is an infrastructure to measure the workplace performance of individuals that is suitable for making this determination. While we cannot determine how training performance affects workplace performance, we are seeking to include arguments about how the workplace affects training performance. This is because the training occurs at the workplace in an uncontrolled learning and testing environment. This argument is considered a Level -3 (negative 3) consideration since it is the reverse of the Level 3 results affected by the results of Levels 1 and 2. See Figure 5 for a visual that compares these two approaches to understanding the training/workplace relationship to determine Level 3 and Level -3 results.

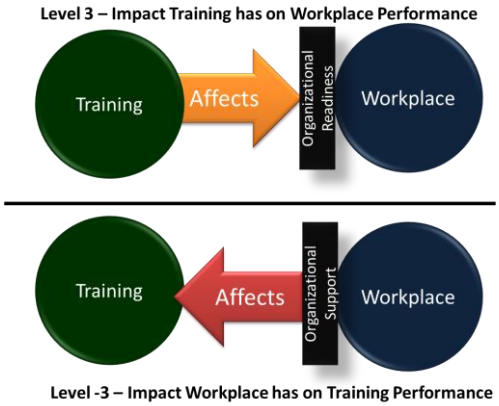

Figure 5. Level 3 and Level -3 Evidence

Example of Applying Kane's Criterion for Evaluating Practical Arguments

What follows are examples of arguments that are based on plausible assumptions to make a determination if the evidence supports or disproves our arguments. The reader should note how these arguments and assumptions establish a more rigorous evaluation framework to make evidence-driven judgments about the effectiveness of a course in the given situation.

Argument Claim 1: The course is instructionally sound and psychometrically pure.

This standard was established by VBA. Our assumptions for this argument are:

Assumption 1: The content for the course reflects the principles, practices, and procedures approved by the VBA per the SMEs and/or APs they provide to support the analysis, design, and development phases.

This assumption is evaluated by reviewing the role of the SMEs/APs while developing the course and their role during the content and test reliability and validation studies. Traceability documentation is used to link the tests to the course content to the course objectives.

Assumption 2: This development process (to include building a reliability/validity traceability matrix based on SME reviews) and the requirements established by the SOP are a satisfactory risk-mitigation strategy for not conducting a course validation study as a formative assessment prior to deployment.

This assumption is evaluated by reviewing how well the instructional design and development team adhered to the approved curriculum development process and SOP requirements. In addition, no credible evidence exists to contradict this assumption or to support it. This includes evaluating the methodology for establishing the passing score cutoff for each case and the methodology for establishing the 80% pass rate to consider the course effective.

Assumption 3: There are no sources of systematic error that will bias the interpretation of the test scores.

This assumption is evaluated by ensuring students understand the purpose of the training and its relevance to their job performance. It also evaluates the clarity of the learning and testing instructions to prevent misunderstandings and issues that might prevent the software from inputting their selected responses. In addition, we ensure Section 508 compliance. Additional evidence considered to evaluate this assumption is the motivation of the participants and the environmental conditions that either support or hinder their ability to demonstrate accurately their task performance capabilities on the assessment instruments. Another source is English-as-a-second-language comprehension barriers. The data for this argument comes from Level 1 comments, Likert Scale rankings, and traceability documentation.

Argument Claim 2: The course identifies low performers with low-test scores.

Assumption 1: An appropriate measure of success in the course is available.

Course test scores represent the evidence used for this assumption. What is difficult to determine is if the pretest was designed to allow high performers to test out. Unfortunately, anyone in the organization can take a course and self-identify if they represent the target population for the course on the end-of-course survey. This means we cannot identify the high performers from the target population. In addition, we are unable to determine the high performers from the workplace because there is currently no need to have or track that data. Time in position does not translate to high performance. In addition, a high performer could fail the pretest by mismarking a test response or the system incorrectly recording a response. The pretest currently stops if the test taker misses one question and is sent immediately to the instructional element of the course. We have requested permission from the customer to collect demographic information before beginning the pretest to help identify the number from the target population who pass the pretest to get a sense of the entering proficiency levels of the designated workforce.

Assumption 2: The instructional course is effective in improving task skills of low performers measured by the assessment instruments.

If the pretest accurately identifies low performers, then it is reasonable to assume an increase in the posttest score reflects an effective course. The pretest-posttest data seems sufficient evidence to prove an increase but does not help understand why the 80% pass rate threshold is not met within a specified number of cases. Improvement does occur, sometimes significantly, but unless the 80% threshold is met, the course is not considered effective. There is no data to identify the entering performance level of participants when they take the pretest to determine whether the course improved posttest scores or if they already possessed the skills and were required to complete the instructional material due to an erroneous failure on the pretest.

Argument Claim 3: The course identifies high performers with high-test scores.

Assumption 1: Participants with high skill level will not substantially improve those skills in the instructional course and therefore would not substantially improve their chances of success on the posttest.

This claim is difficult to prove because pretest scores seem unreliable to identify low-performers from high-performers because of the "miss one question, you fail" strategy. In addition, those who pass the pretest do not take the course or the posttest. Without a pilot study, it is not possible to make a credible interpretation to conclude the argument is true.

Additional Considerations Related to Arguments

From a scoring perspective (see Figure 2), we have two assumptions in conflict. One is the course is effective prior to deployment based of the validity and reliability studies conducted after the course is developed (prior to deployment). The conflicting assumption is the course is not effective after deployment until a certain number of participants satisfy an 80% pass rate threshold.

A fundamental question is, does the test measure what it intends to measure and does it do so reliably? From an argument-based validation approach, the answer seems to be yes because pre-deployment validity and reliability studies established this. However, if the test scores seem to indicate that learning did not occur during the course and the test scores reflect this lack of learning, is the course truly effective? TT&E expects courses that are

instructionally sound and psychometrically pure, which is determined during the pre-deployment validity and reliability studies. For this situation, a new term is warranted: psychometric realism—"the view that attributes characterizing the aim of psychological and educational measurement exists in the actual world *and* that the claims about their existence can be justified" (Hathcoat, 2013). The proponents of psychometric realism are realists about the real world and theory.

When looking at the arguments and assumptions previously presented, it seems the scoring category raises the most significant red flags preventing a credible conclusion. The internal pre-deployed validation study was conducted in a controlled environment where participants were able to concentrate on the course without internal or external noise to distract them. The noise from the workplace introduces a varied level of error depending on each participant's situation, thereby corrupting the resulting test scores and an opportunity to make meaningful judgments about them. No evidence exists that the poor test scores are a result of an ineffective course. Therefore, applying the scoring inference of the argument-based approach previously discussed (see Figure 2), we conclude that the test scores accurately reflected the performance of the participants—not necessarily their capabilities.

Table 1 provides additional evidence for this conclusion. One course completed two post-deployment validation studies—one in a semi-controlled environment and the other in a controlled environment at the same location. The outcomes for the same course were different. One lesson met the 80% pass rate during the first validation study and all three met it during the second study. In addition, when the same "validated" course was evaluated later against the same validation criterion, none of the lessons met the 80% standard.

Table 1. Validation Results Comparison in Three Environments

	Apr Val - Baltimore (Semi-Controlled) Apr 2014				Jul Val – Baltimore (Controlled) Jul 2014				Post-Val (Uncontrolled) 2015			
Lesson	Val Pop	Pass	Fail	Pass Rate	Val Pop	Pass	Fail	Pass Rate	Val Pop	Pass	Fail	Pass Rate
1	137	52	85	38%	174	146	28	84%	98	62	36	63%
2	78	54	24	69%	151	134	17	89%	80	47	33	59%
3	72	61	11	85%	171	151	20	88%	94	44	50	47%

The only variable that changed was the training/testing environment. The Validation Population (Val Pop) represented the same sampling population. Only when the students completed the course in an environment that allowed them to concentrate on the course content and test did all three lessons meet the 80% pass rate.

Conclusions

The focus of this paper is on our attempt to adapt Kane's Argument-Based Approach to Validation to meet revised VBA requirements to validate a training course's effectiveness after it has deployed rather than before. As a result of the application of this course validation approach, we have been able to see beyond test score results to gain a better perspective of what the data are and are not telling us. The three argument claims we developed established anchor assumptions that were either overlooked or missing from the course validation data required to establish plausible interpretations of test scores. For example, different perspectives to display and interpret the data included reasonable flexibility defining how many cases were required for 80% of the participants to pass the required number of cases. If the test data results indicate an 80% pass rate was not achieved that required participants to pass three cases in no more than six cases seen but does occur within seven cases, is that sufficient evidence to conclude the course was effective? This approach established a time-to-proficiency metric that allowed the key stakeholders to make a data-based argument to make a plausible interpretation of the data rather than submit to an immutable but arbitrarily assigned validation pass rate cutoff.

Math and statistics provide meaningful data. The argument-based approach to validation, however, is about reasoning, not math and statistics. Without clear and coherent arguments supported with plausible assumptions, math and statistics are limited in their ability to help us understand the meanings of test data results and their interpretations.

We are still exploring and maturing our understanding and application of this validation approach. We tend to agree with a study that evaluated the differences between the argument-based approach to validity and the approach described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). It concluded that the argument-based approach provided concepts and practices that were useful and new (Chapelle, Enright, & Jamieson, 2010). We discovered this was true as well. However, we remain unconvinced that this methodology is a credible substitute for validating the effectiveness of a course to meet the learning objectives it was designed to achieve, especially when you consider the comparative test score results presented at Table 1. We continue to engage with our customer to resolve this unique challenge.

REFERENCES

AERA, APA, & NCME. (1999). *Standards for Educational Psychological Testing.* Washington, DC: Author.

Brennan, R. L. (1983). *Elements of Generalizability Theory.* Iowa City, IA: American College Testing.

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

Downing, S. M. (2003). Validity: On the Meaningful Interpretation of Assessment Data. *Medical Education, 37*, 830-837.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn, *Educational Measurement, 3rd Edition* (pp. 105-146). New Yord: American Council n Education and Macmillan.

Hathcoat, J. D. (2013). Validty Semantics in Educational and Psychological Assessment. *Practical Assessment, Research & Evaluation, 18*(9), 1-14.

House, E. R. (1980). *Evaluating with Validity.* Beverly Hills, CA: Sage.

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

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kirkpatrick, D. L. (1994). *Evaluating Training Programs.* San Francisco, CA: Berrett-Koehler Publishers.

Messick, S. (1989). Validity. In R. L. Linn, *Education Measurement, 3rd Edition* (pp. 13-103). New York: American Council on Education and Mcmillan.

Newton, P., & Shaw, S. (2014). *Validity in Educational and Psychological Assessment.* Los Angeles, CA: SAGE Publications.

Phillips, J. J. (1983). *Handbook of Training Evaluation and Measurement Methods.* Houston, TX: Gulf Publishing.

Schuwirth, L. W., & van der Vleutin, C. P. (2012). Programmitic Assessment and Kane's Validity Perspective. *Medical Education, 46*, 38-48.

Sireci, S. G. (1998). The Construct of Content Validity. *Social Indicators Research, 45*(1-3), 83-117.

Toulmin, S., Reike, R., & Janik, A. (1979). *An Introduction to Reasoning.* New York: Macmillian.

U.S. Department of Defense. (2001). *MIL-HDBK-29612-A, Instructional Systems Development/Systems Approach to Training and Education (Part 2 of 5).* Washington, DC: Government Publishing Office.