Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 17, Number 13, October 2012 ISSN 1531-7714

Evaluating the Appropriateness and Use of Domain Critical Errors

Chad W. Buckendahl & Susan L. Davis-Becker
Alpine Testing Solutions

The consequences associated with the uses and interpretations of scores for many credentialing testing programs have important implications for a range of stakeholders. Within licensure settings specifically, results from examination programs are often one of the final steps in the process of assessing whether individuals will be allowed to enter practice. This article focuses on the concept of domain critical errors and suggests a framework for considering their use in practice. Domain critical errors are defined here as knowledge, skills, abilities, or judgments that are essential to the definition of minimum qualifications in a testing program's pass–fail decision-making process. Using domain critical errors has psychometric and policy implications, particularly for licensure programs that are mandatory for entry-level practice. Because these errors greatly influence pass–fail decisions, the measurement community faces an ongoing challenge to promote defensible practices while concurrently providing assessment literacy development about the appropriate design and use of testing methods like domain critical errors.

The consequences associated with the uses and interpretations of scores for many credentialing testing programs have important implications for a range of stakeholders. This assertion particularly holds true for licensure testing programs in which the primary responsibility is public protection (Mehrens, 1995; Shimberg, 1982). Within licensure settings, results from examination programs are often one of the final steps in the process of assessing whether individuals will be allowed to practice medicine, fly and land an airplane, invest an individual's life savings, or test drinking water for public consumption. In each of these cases, one can identify numerous stakeholders who rely on the licensure examination process to identify and protect the public from candidates who do not have the minimum skills necessary to enter practice.

Test developers for licensure testing programs are therefore responsible for creating examinations that represent the important components of the job domain and supporting decisions about whether examinees meet the minimum qualifications defined by the subject matter experts and policymakers that represent the profession. Because of its role in public protection, some licensure testing programs have incorporated conjunctive decision rules that impact standard setting policy and ultimately the pass–fail status of candidates. In these instances, programs have implemented additional criteria that make a candidate's ability to pass the examination contingent upon demonstrating minimum competency on a small set of items or sometimes a single item, task, or scoring element within an examination. Under such a model, the stakes associated with a particular item (or set of items) are equal to that of the entire exam. The purpose of this article is to evaluate the appropriateness and use of these types of items and tasks in licensure testing programs. Using a multi-state clinical skills licensure examination program in dentistry to illustrate the concept, we discuss the conditions under which these items may or may not be appropriate for

testing programs that make important decisions about individual examinees.

Psychometrically, making candidate-level pass–fail decisions on the basis of a small number of items, tasks, or scoring elements is perilous given the challenge of defending the practice relative to professional standards. Specifically, guidance on reliability of scores and decisions is included in the section on reliability and errors of measurement within the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). However, within some licensure and certification testing programs, this situation is regularly observed when policy intersects with psychometric practice. In particular, this occurs because certain areas of the domain have been judged as critical by subject matter experts and the sponsoring agency during the development process, often at the point of domain specification (e.g., practice analysis, blueprint development). The result of failing to demonstrate minimum expectations in one of these areas signals a "domain critical" error and results in failure of a section of the examination or the full examination.

In this article, we define "domain critical error" as a lack of a knowledge, skill, ability, or judgment that warrants substantive weight in the definition of minimum qualification at the point of a pass–fail decision about a candidate's competency in the domain. Although we generically refer to these measurement opportunities as "items," they may be observed as individual items, tasks, or scoring elements. Further, these errors may occur at a logical point in a sequence of processes or at a point where a number of tasks have been performed to produce a product. We note that the use of these items is not widely observed in practice; however, they do appear in some expected—and some not-so-expected—programs.

The concept underlying domain critical errors is not new. The concept is analogous to a Guttman scale (1944) in which a sequence of items produces response patterns where examinees respond correctly up to a given point and then respond incorrectly beyond that point. The assumption of such a scale is that knowledge, skills, or abilities above that point would be answered incorrectly. In practice, examinees' responses to cognitive ability measures are not perfectly correlated, so the observation of these ideal patterns is unlikely. A Guttman-like scale response pattern could be the result of the application of domain critical errors to a program that is evaluating procedural characteristics of a domain. However, a domain critical error may be defined as an element of a product of a complex performance task; therefore, the sequential nature of the Guttman scale would not always hold.

Domain critical errors have also been conceptualized, labeled, and discussed in other forms in professional literature and practice. For example, Fortune and Cromack (1995) describes "go no-go" items from a clinical skills dental licensure exam that used analytical scoring practices. Childs, Dunn, van Barneveld, and Jaciw (2007) and Childs, Dunn, van Barneveld, Jaciw, and McIlroy (2003) reference "killer items" in the context of a medical licensure examination in which candidates were asked to demonstrate patient management skills. With respect to performance tests, Judd (2009) suggests the term "gating items" for a range of performances that presumably would be related to activities for which a desired sequence or protocol was not followed (i.e., a gate that precludes examinees from continuing the examination). Semantics aside, these items are meant to evaluate what has been defined as a critically important job-related ability. For items, tasks, or scoring elements that have been designated as "domain critical," candidates who cannot perform them will fail the exam. As suggested by one of the reviewers, this approach takes the concept of conjunctive decision-making to an extreme level. In addition, it creates challenges for testing programs to support valid interpretation and use of scores from a psychometric perspective because evidence of domain representation, reliability, and standard setting are weakened relative to the potentially full set of information.

Other literature that is relevant to a foundational understanding of earlier work in the concept of domain critical error items focused primarily on selected response tests. For example, Webster, Goldfarb, Norcini, Shea, and Murray

(1987) evaluated instances when dangerous options were selected by candidates on one of the tests designed by the American Board of Internal Medicine. From this study, they concluded that candidates who made these domain critical errors also performed poorly on other parts of the exam. The authors in this study may have suggested a different scoring option or decision rule if results had shown no relationship or a negative relationship.

Similarly, Floreck, Guernsey, Clyman, and Clauser (2002) evaluated domain critical errors on Step 3 of the U.S. Medical Licensure Exam that involved scenario-based items that target candidates' clinical judgment abilities. In contrast to other research that involved fully compensatory scoring, the testing program described by Floreck et al. increased the weights associated with domain critical errors and assigned negative points to these items to reflect their increased criticality within the domain. Manning (2000) (in the context of air traffic controllers) and Woychesin (2002) (in the context of airline pilots) also describe situations in which domain critical errors are included as part of the scoring criteria. Given the stakes associated with the outcomes of such exams and the need to develop and implement psychometrically appropriate and defensible practices, further investigation into the appropriateness of such individual items (or subset of items) is necessary.

Although still uncommon in a majority of credentialing testing programs, advocates for the use of domain critical errors in performance testing (e.g., Judd, 2009) combined with a lack of clear guidance in the professional literature have raised the need to address this topic within the broader measurement community. Stakeholders contributing to the exam development process representing content and policy perspectives often have compelling, but subjective reasons for requesting test items that would automatically fail examinees if answered incorrectly. Content experts generally take strong ownership of their domains and can be challenged to consider the knowledge, skills, and abilities of the target population of examinees (e.g., entry-level, minimally qualified). This ownership may result in overemphasis on very small parts of the domain. In addition, policymakers, particularly in licensure settings, take their charge of public protection very seriously. Consequently, they may reflect this responsibility by prioritizing a greater tolerance for Type II errors (i.e., candidates who are qualified, but did not meet the required performance expectations) over Type I errors (i.e., candidates who are not qualified, but did meet the required performance expectations). This heightened sense of public safety can be illustrated through the use of domain critical errors in practice given that most of the available research has been related to healthcare and air travel.

In this article we evaluate the appropriateness and use of domain critical error items in a clinical skills examination program in which scores are used to inform licensure decisions for dentists in a consortium of state licensing boards. Specifically, we evaluate the domain representation of the content, definition of domain critical errors, classification of domain critical errors, standard setting procedures, pass–fail rates, and score profiles of candidates who failed one or more exams for one organization's clinical skills examinations in dentistry. Data were based on test administrations from the 2009 calendar year testing cycle and included tests of amalgam restoration, composite restoration, endodontics, and fixed prosthodontics. We structured our evaluation with input from existing literature and current uses of these item types in practice. Our discussion concludes with recommendations for factors that testing programs should consider and be prepared to defend prior to including domain critical errors on an examination. We recognize that like us, most readers will not have content knowledge of some of the topics noted herein. However, we use the information from this program to illustrate the domain-specific nature of the task. In the final section we provide guidance for generalizing the framework to other programs.

Overview of Clinical Skills Tests in Dentistry

The dental testing program used as an exemplar throughout this article is supported by one of five regional U.S. state consortia that have formed to share the clinical skills test development,

maintenance, and administration responsibilities. The clinical examinations developed by these consortia are one requirement of most states in receiving a license to practice dentistry. Although the focal examination program also includes separate tests that measure candidates' clinical judgments (e.g., diagnosis, assessment, treatment planning), the domain critical errors at the heart of this article are limited to the clinical skills portion of the examination program (e.g., instrument use, handpiece manipulation, domain-relevant materials, manual dexterity).

Domain Representation

For this organization, there are four clinical skills dental examinations. Two of these examinations are within the domain of operative procedures (e.g., removing tooth decay, preparing a filling, placing a filling), one examination is in endodontics (e.g., performing a root canal), and one examination is in fixed prosthodontics (e.g., placing crowns). The two operative procedures are performed on patients who have been pre-identified by the candidates as meeting specific eligibility criteria. There are two steps in each of these procedures: preparation and restoration. In the preparation step, the patient's tooth is prepared to receive the respective restoration material by using a high-speed handpiece (i.e., drill) to remove defective tooth structure and any remaining decay. After the preparation step is completed, the candidate's work is submitted to a blind scoring process conducted by trained examiners. In the second step, the candidate is required to restore the tooth to its natural contour. The examiners then score the restoration step.

The endodontics and fixed prosthodontics examinations are conducted on simulated patients using a dentoform (i.e., artificial set of teeth) within a typodont (i.e., artificial hinged jaw with gums) that is mounted in a manikin (e.g., artificial head, lips, cheeks). For the endodontics examination, tasks are assigned on one anterior and one posterior tooth to represent the range of locations and skills performed in practice. On the fixed prosthodontics examination, candidates are required to perform two procedures. Although the endodontics and fixed prosthodontics examinations are performed on simulated patients, the scoring procedures are similar to the operative procedures. Specifically, candidates are given a certain amount of time to perform the procedures, and then the dentoforms are submitted for scoring by the examiners. The organization's decision to measure some procedures with patients versus simulated patients is based on a combination of psychometric and policy factors (e.g., fidelity, job-relatedness, invasiveness of the procedure).

Scoring

Across all four clinical examinations, each procedure is divided into sub-tasks that represent scoring criteria that are analytically scored dichotomously as 1 (minimally competent performance or higher) or 0 (less than minimally competent performance). Because additional factors such as the task, examiners, and location can influence the estimate of error, three examiners who have been trained in the examination procedures independently score the candidate's performance. Although psychometrically it is more efficient to have two examiners score the performance and then adjudicate as needed, the logistics of the clinical skills exams for patient-based procedures are easier when three examiners provide judgments. Scoring decisions can be calculated after the patients have left the scoring area.

For each scored sub-task within the procedure, a decision rule is then applied to evaluate whether or not an "error" is present. Specifically, an error is counted against a candidate only if it was observed by two or more examiners. If a given error is observed by only one examiner, it is not counted against the candidate. This decision rule is applied across sub-tasks within a procedure, and the candidate's sub-task scores are summed to estimate their total score on the procedure. If the total score meets the minimum passing score, the candidate passes the respective examination. These decision rules are applied to all clinical skills procedures. However, not all errors are weighted equally.

Most errors are characterized as minor. A candidate can make any of these errors and still pass the exam as long as his or her total score still meets

the overall passing score. However, some errors are considered to be domain critical; committing any such error will result in an automatic failure decision for the candidate on that respective examination. As used within these examinations, domain critical errors represent egregious performance within the assigned procedure (e.g., remaining decay, unsupported enamel) or within the environment but beyond the scope of the procedure (e.g., major infection control violation, preparing the wrong tooth) that signify a skill level that could significantly threaten—in the judgment of the testing organization—the health or well-being of the public if the candidate were deemed eligible for a license. As a consequence, a candidate can fail an examination on the basis of a single scoring element if a domain critical error is independently observed by two or more examiners. Passing each of the examinations separately is required for licensure eligibility.

Framework for Considering Domain Critical Errors

In this section, we highlight the steps in the test development and validation process that warrant targeted attention by practitioners on this topic. We propose using the framework illustrated in Figure 1

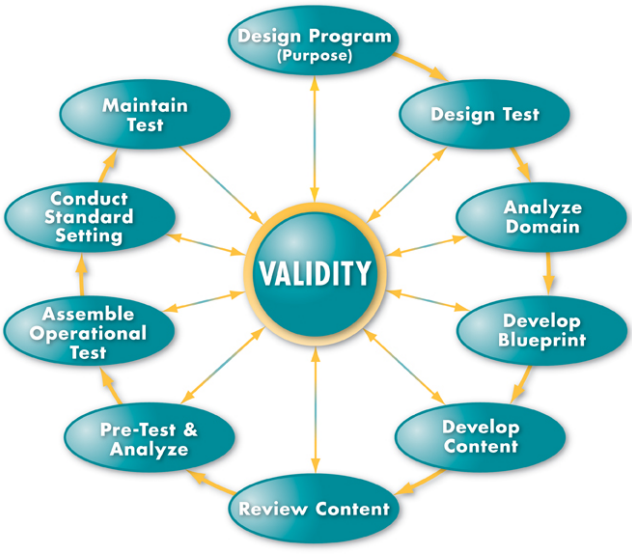

Figure 1. Test Development Process

to systematically evaluate the conditions under which domain critical errors may be considered and how practitioners can help program sponsors understand when they are appropriate and when they are not. We note five stages of test development that can be considered prior to implementing domain critical errors: program design, domain analysis, content development, reliability, and standard setting. For each of these areas, we describe the information that programs would evaluate and then apply the concept to the dental examinations described above to illustrate how a program could use this framework.

Program Design

In the program design stage, the intended uses and interpretations of test scores are defined. For a testing program considering the use of domain critical errors, this is the first opportunity to consider the appropriateness of having such elements within the exam. A primary consideration is the stakes associated with the decisions made from the scores and the potential consequences for stakeholders and candidates. In mandatory licensure testing programs in which the primary interest is public protection, the intended use is to keep incompetent candidates from entering the field where they may do harm. In this situation, test developers (e.g., policymakers, subject matter experts) may have an increased awareness of the potential risk of unqualified examinees. This type of program contrasts with a voluntary certification testing program in which successful performance on the exam may not have similar consequences for stakeholders. At this stage in the development process, policymakers should reflect on potential areas of the job domain that are more critically important than others and anticipate potential discussion points that may emerge from the domain specification step of the process. At this point, we recommend test developers facilitate subject matter experts in thinking about what acts would not be acceptable in practice (i.e., what the subject matter experts have seen in practice through disciplinary actions, public complaints to licensing boards, licensure enforcement, or through legislatively defined practice).

In the context of clinical skills examinations in dentistry, the charge of public protection is an important part of any initial discussion because state licensure boards are often political appointees. A challenge at this stage is to communicate the importance of following acceptable test development and validation practices that may conflict with stakeholders' personal standards, an intuitive understanding of their role, and differing interpretations of their charge. Another challenge at the program design and renewal stage for this specific program is the established legacy of relying on conjunctive decisions and using domain critical errors for many years prior to the consideration of defensible psychometric practices. Fortunately, for licensure programs, the general policy definition of the minimally qualified candidate is similar (i.e., protection of the public) given the purpose of these programs. Conversely, certification programs can be much more diverse in their design depending on the domain, consequences associated with pass–fail decisions, and the underlying purpose of the testing program.

Domain Analysis

For any testing program, there is a stage of domain specification that considers the content, cognitive demand, environment, and performance demand of entry-level expectations in the field. Within licensure programs, this activity often occurs through a systematic job analysis in which it is important to link the knowledge, skills, abilities, and judgments to the job-related elements of the domain. Because job analysis activities often involve surveys of practitioners in the field of interest, the results can provide evidence to inform discussions of test design and content weighting including the criticality of specific content elements within the domain. If the use of domain critical errors is anticipated for a program, specific questions can be included in the job analysis questionnaire to gather this information from a broader sample of practitioners in the field. The evidence from the job analysis should then confirm the criticality of the skills or abilities that the program is considering domain critical.

Within the illustrative dental licensing program, job analysis is conducted approximately every 5 years to evaluate the content representation of the domain to inform potential changes in the examinations. Another use of these results is to evaluate prior decisions about whether certain tasks or skills should continue to be domain critical errors on future examinations. One important topic when considering the criticality of errors is the concept of a "correctable error." Although content experts may want any error on the examination to fail a candidate given high expectations for new professionals, it is difficult to defend identifying a particular error as domain critical when it would be correctable in a real-world situation (e.g., with additional time, resources, or more job-related knowledge/skills).

To extend the example, suppose one of the tasks on an operative procedure in the dental licensure examination requires a candidate to prepare a tooth for restoration; however, the characteristics of the preparation were less than the ideal internal structure. Should failure to demonstrate this skill automatically fail a candidate on the examination if the situation occurs regularly in practice? Using the correctable error model, this error would not rise to the level of domain critical because there are often multiple approaches for recovering the preparation at this point in the process. However, if a candidate were to "prepare" the wrong tooth and unnecessarily remove sound tooth structure in doing so, this situation cannot be corrected. In an analogous situation, the same could be said about a candidate who prescribes a lethal dose of a medication to a simulated patient in a written scenario of patient management or to a nuclear engineer whose actions in managing a reactor cause a core meltdown. These situations may occur as the result of a combination of errors or one egregious decision or action. However, in each instance, the error is not correctable, and the results could have lasting impact.

Table 1 lists the domain critical errors that were identified during the domain analysis stage of this dental examination program. The table displays two types of critical errors: those that were observed within the assigned performance task and those that

were procedural or outside the scope of the assigned task, but within the job-related scope of practice. This table illustrates how the decision process for defining domain critical errors occurred for each examination and reinforces the need for domain expertise in determining these characteristics.

practice. Second, one must consider the potential damage that could be realized by the public if similar performance was observed in practice. Again, this is where the concept of the "correctable error" should be considered in evaluating domain critical items. For constructed response or performance items, the

Table 1. Summary of domain critical errors for each examination

	Task-specific domain critical errors	Procedural domain critical errors
Amalgam	Caries (decay) Remaining	Major Tissue Damage
	Restorative Materials Remaining	Anesthetized Prior to Approval
	Iatrogenic Exposure	
	Unrecognized Exposure	
	Major Tissue Damage	
	Excessively Open Proximal Contact	
	Open or Short Margins	
Composite	Caries (decay) Remaining	Major Tissue Damage
	Restorative Materials Remaining	Anesthetized Prior to Approval
	Iatrogenic Exposure	
	Unrecognized Exposure	
	Major Tissue Damage	
	Excessively Open Proximal Contact	
	Open or Short Margins	
	Sealant Detected	
Endodontics	Undiagnostic Radiographs	
	Underfill or Overfill of gutta percha	
	Improper Seal Apical 1/3	
	Perforation	
	Excessive Access Opening	
	Inability to Locate Canal Openings	
	Failure to Remove Roof of Pulp Chamber	
Fixed Prosthodontics	Excessive Over or Under Reduction	Removal of Tooth from Typodont
	Major Tissue Damage	Major Tissue Damage

Item Development

In developing items that will be used to represent domain critical errors, there are a number of considerations. First, there must be a clear connection between the content and cognitive demand of the item and the results of the job analysis. If challenged, a licensure program's first line of defense is to provide evidence that the knowledge, skills, abilities, or judgments represented on its examination are important to entry-level

related expectation is to develop the scoring guide or rubric to provide guidance for calibrating raters and defining the level of tolerance for these performances.

Within the illustrative dental examinations described in this article, the results of the program's job analysis directly informed the choice of procedures that are represented within these domains. The organization used a combination of frequency and criticality data to prioritize which

procedures and tasks were most important to entry-level practice, which procedures were most frequently observed in practice, and which skills could be reasonably measured within a job-related environment on the examination.

The specific procedures that were identified were then broken down into tasks and sub-tasks that comprised the procedure with the corresponding scoring guide that identifies the criteria for acceptable performance, minor errors, and the domain critical errors illustrated in Table 1. Exam developers used three sources to build the scoring criteria for a given procedure. First, they considered the organization's definition of minimum competency as interpreted by the subject matter experts on the committee. Second, faculty members in their respective domains from dental training programs provided feedback regarding how procedures were currently represented in the curriculum and taught in the simulation labs and clinics. Third, the committee drew on professional literature from textbooks and research to inform the current state of practice. Finally, psychometric input regarding limiting domain critical errors to those that cannot be corrected was also part of the discussion.

Similar to readers of this article, the authors have a layman's understanding of the domain. As a result, we are relying on the subject matter experts' recommendations regarding the elements that were identified as domain critical errors that may reasonably preclude an individual from entering practice due to the potential risk to the public. However, from our layman's perspective, it also appears that some elements that currently result in an automatic failure of a candidate may not rise to the level of an uncorrectable error. For example, within the endodontics examination, if an undiagnostic radiograph (i.e., x-ray) was produced in practice, a practitioner would likely simply take another radiograph. This may or may not be reasonable within the profession given guidelines regarding acceptable radiation exposure for patients within a given timeframe. Also, within both the amalgam and composite (i.e., operative) examinations, if remaining material was observed in practice, would a practitioner be able to correct the error without permanent damage to the patient's oral health? These are the types of questions that examination committees can ask of themselves when they consider the use of domain critical errors.

Reliability

Another source of validity evidence is information about the errors associated with the scores and decisions for examinations that include domain critical errors. A number of factors can influence estimates of errors for these item types (e.g., representation of the domain, number of items, inter-correlations among items). For selected response items, testing programs often rely on internal consistency, single administration decision consistency reliability information from a classical test theory perspective, or conditional standard error of measurement at the cut score estimates when applying an item response theory model. Professional standards provide guidance on levels of tolerance for making decisions. However, because the reliability of a single, dichotomously scored item is very low, the use of domain critical errors in the literature appears to be more likely to occur in performance testing settings.

For constructed response items or tasks that are inclusive of performance tests, additional sources of error include the number of tasks, number of raters, weightings associated with given scoring elements, and number of occasions. Applying generalizability theory (e.g., Shavelson & Webb, 1991) or the many-facets Rasch model (e.g., Bond & Fox, 2001) are both strategies for evaluating errors that consider multiple factors. For programs that may not have the technical sophistication to apply these methods, a commonly applied method in practice is to evaluate only errors associated with rater judgments. Given the scoring criteria, weighting applied to the domain critical errors, and the subjective nature of the scoring task, evidence of rater errors likely represents the greatest potential threat to validity evidence within the range of intended uses. Independent confirmation of domain critical errors, which are ultimately decisions, is a necessary element of the scoring criteria. Evaluating raters' performance is then an important component of internal quality control for the program to ensure

that examiners continue to provide consistent judgments as expected.

When applied to the performance examinations in our continuing illustration, the dental program has established a decision rule that requires two of three examiners to agree independently on the presence of an error before the error is counted against the candidate's performance. If this agreement occurs on any of the scoring elements that represent one of the domain critical errors, the candidate fails the respective examination. This rating and decision strategy provides some evidence of reliability to support the decision; however, it excludes some of the other factors noted above as potentially contributing to confidence in the decision. These judgments can then be summed across candidates within a given examination to evaluate the overall agreement among the examiners, each individual examiner's agreement with their colleagues, and the confidence that the program can have with its decisions because the raw scores are less meaningful when interpreting performance.

Standard Setting

When considering and implementing domain critical errors, there is an inherent interaction about the decisions among policy, program design, domain specification, item writing, and standard setting factors. Because the inclusion of these items in an examination is often determined at the point of the program design or domain specification, the standard setting decision rules are developed over multiple steps in the test development process rather than occurring solely as a distinct event following pilot or operational administration. Methodologically, this approach has elements of the Dominant Profile Judgment Method described by Plake, Hambleton, and Jaeger (1997) and Judgmental Policy Capturing (Jaeger, 1995). Both of these methods involve subject matter experts evaluating profiles of candidates' performance on the examination. These methods heavily weigh policy factors in the standard setting process and consider decision rules that may be compensatory, conjunctive, or some combination of these rules. Another element of the policy considerations at this point is to revisit the program's tolerance for Type I (false positives) and Type II (false negatives) errors in the decision-making process that were discussed at the outset in the program design. In the case of a licensure program, there may be a lower tolerance for Type I errors given the potential consequences for the public of an incompetent candidate entering practice.

With respect to these dental examinations, differential decision rules were applied based on the potential harm to the public that may be caused by less than acceptable performance by the candidate. As shown in Table 1, the performance demonstrations that were judged to represent domain critical errors within each examination are a function of the criticality and frequency of the respective domain. In addition, within the amalgam, composite, and fixed prosthodontics examinations, candidates can also fail due to an accumulation of minor errors. However, all errors within the endodontics examination were judged to be domain critical and result in candidate failure if any error is observed by two or more examiners. To avoid unrealistic expectations, the definition of domain criticality for these examinations is anchored at the point of minimum competency to remain consistent with the purpose of the examination. To illustrate the impact of these domain-specific policies, Table 2 lists the overall pass–fail rates and then breaks down the proportion of failures that were attributable to the range of observed minor and domain critical errors to evaluate how candidates failed a given examination.

We can see in Table 2 that the overall pass rates for each examination are high. Because the clinical examinations are the third step in the dental licensure testing process, these pass rates are not unexpected; the first two layers of written examinations filter out other candidates in the staged process. For these programs, approximately 500 candidates take each examination in a given calendar year. It is interesting to note that although the pass rates are fairly similar, the patterns of how candidates fail vary across the examinations. For

Table 2. Summary of candidate pass–fail status by number and breakdown of errors for failing candidates of each examination for an illustrative program year

	Amalgam	Composite	Endo-dontics*	Fixed Pros
Pass rate	97%	99%	99%	92%
Overall failure rate	3%	1%	1%	8%
Fail - Minor errors	19%	0%	-----	8%
Fail - Minor error plus 1+ DCE	31%	43%	----	84%
Fail - 1 DCE	25%	43%	100%	8%
Fail - 2+ DCE	25%	14%	0%	0%

*Note that for the endodontics examination, there are no minor errors.

example, within the endodontics examination, there are no defined minor errors; candidates who failed this examination were determined based solely on the confirmed observation of a domain critical error. However, within the amalgam examination, there was more of an equal distribution of reasons for failure across the possible categories. Because both the amalgam and composite are considered operative dentistry skills, we hypothesized that the failure patterns would be similar between these two examinations given the similar skill sets being demonstrated. However, not only were the pass rates slightly different; candidates who failed the composite examination did so due to a domain critical error. Finally, the fixed prosthodontics examination yielded the lowest pass rate among these four examinations with failure explained predominantly by the combination of minor errors and one domain critical error.

Conclusions

Although domain critical errors have been used in practice for many years, the research literature on the appropriateness and use of such errors is sparse with respect to guidance on whether they should be used—and if so, under what conditions. In this article, we described the range of potential uses for domain critical errors from a collection of items, specific tasks, or scoring elements. Further, we discussed how these errors could be defined and observed within a logical sequence of activities or as heavily weighted components in scoring the product from a complex performance task. In addition, we provided a framework for how to consider the use of domain critical errors and illustrated the application of the framework to a dental licensure program.

Using domain critical errors creates psychometric, policy, and legal implications, particularly for licensure programs that are mandatory for entry-level practice. The use of these items within the voluntary arena of certification testing was also briefly discussed to the extent that the core elements to evaluate are analogous across these programs. Psychometrically, it is important that the validity evidence for the program can support the intended uses and interpretations of the scores for the defined purpose. For fields in which there is a high risk to the public in granting a credential to an unqualified applicant (e.g., healthcare, aviation, architecture), the purpose to protect the public is taken as a serious charge by policymakers who oversee these testing programs. Thus, the potential consequences for the public, policymakers' interpretations of controlling legislation (e.g., State Practice Acts), and the ability to communicate expectations to a lay audience may supersede psychometrics as the primary concern for these programs.

Although psychometrics may sometimes be relegated in these situations to a more technical consideration, practitioners can still anchor the validity of the decision with an argument about the criticality of the job-relatedness of the item, task, or scoring element. We also recommend that practitioners apply a criterion of whether or not the potential domain critical error is correctable. This strategy forces subject matter experts to align their judgments to operational practice and to determine whether or not the item, task, or scoring element should be included in the examination as a sample of what is reasonably encountered by entry-level practitioners in a given field.

The consideration of whether to include domain critical errors on performance tests that incorporate human subjects raises a related ethical question when potential lasting damage can occur during the examination, if the assumption of an

uncorrectable error is applied. When faced with this situation, practitioners evaluating the examination process would need to build a validity argument around whether the use of human subjects can be defended as an important component of job-relatedness. This would likely have the most applicability in healthcare examination situations that might use candidate-recruited patients or standardized patients. Although not a concern in simulated performance examinations (e.g., flight simulators, nuclear reactor management simulation), allowing unlicensed candidates to demonstrate minimum competency on human subjects can result in the types of uncorrectable errors that licensure testing programs seek to prevent in the broader public. Considering the potential impact of a single, unqualified practitioner allowed to practice independently, this may be an acceptable policy trade-off.

Given the lack of guidance in the literature, future opportunities exist for study of this topic in operational settings. For example, how would a testing program defend a legal challenge from a candidate who failed on the basis of a domain critical error? How would a validity argument be constructed differently for selected response items versus a performance task? What additional data analyses could be conducted to provide empirical evidence to support scores and decisions? Clearly, there is much work to do in this area. From a licensure program's perspective and consistent with the *Standards for Educational and Psychological Testing* (AERA et al., 1999), evidence of reliability is often focused on the question of decision consistency when evaluating alternative measurement practices. However, the obvious ongoing challenge for the measurement community is to promote defensible practices while concurrently providing assessment literacy development about the appropriate design and use of testing methods like domain critical errors.

References

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

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Childs, R. A., Dunn, J. L., van Barneveld, C., & Jaciw, A. P. (2007). Does it matter if you "kill" the patient or order too many tests? Scoring alternatives for a test of clinical reasoning skills. *International Journal of Testing, 7*(2), 127-139.

Childs, R. A., Dunn, J. L., van Barneveld, C., Jaciw, A. P., & McIlroy, J. H. (2003). Differential weighting of errors on a test of clinical reasoning skills. *Academic Medicine, 78* (Oct. suppl.), S62-S64.

Floreck, L., Guernsey, M., Clyman, S., & Clauser, B. (2002). Examinee performance on computer-based case simulations as part of the USMLE Step 3 Examination: Are examinees ordering dangerous actions? *Academic Medicine, 77* (Oct. suppl.), S77-S79.

Fortune, J. C., & Cromack, T. R. (1995). Developing and using clinical examinations. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 149-165). Lincoln, NE: Buros Institute of Mental Measurements.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*, 15-40.

Judd, W. (2009). Gating items: Definitions, significance, and need for further study. *Practical Assessment, Research, & Evaluation, 14*(9). Accessed October 9, 2012, from http://pareonline.net/getvn.asp?v=14&n=9

Manning, C. A. (2000). *Measuring air traffic controller performance in a high-fidelity simulation*. Report No. DOT-FAA-AM-00-2. Washington, DC: Government Printing Office.

Mehrens, W. (1995). Legal and professional bases for licensure testing. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 33-58). Lincoln, NE: Buros Institute of Mental Measurements.

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement, 57*, 400-411.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Shimberg, B. (1982). *Occupational licensing: A public perspective*. Princeton, NJ: Educational Testing Service.

Webster, G., Goldfarb, S., Norcini, J., Shea, J., & Murray, L. (1987). Performance of a dangerous answer subtest within a subspecialty certifying examination. *Medical Education, 21*, 426-431.

Woycheshin, D. (2002). Validation of the Canadian Automated Pilot Selection System (CAPSS) against primary flying training results. *Canadian Journal of Behavioural Science, 34*, 84-91.

Citation:

Buckendahl, C. W. & Davis-Becker, S. L. (2012). Evaluating the appropriateness and use of domain critical errors. *Practical Assessment, Research & Evaluation*, *17*(13). Available online: http://pareonline.net/getvn.asp?v=17&n=13

Note

The authors would like to acknowledge the helpful feedback from the editors and two anonymous reviewers on earlier versions of this article.

Corresponding Author:

Chad W. Buckendahl
2467 Cordoba Bluff Ct.
Las Vegas, NV 89135
Phone/Fax: (702) 586-7386
Email: chad.buckendahl [at] alpinetesting.com or
drcbuck [at] gmail.com