An Empirical Evaluation of Using Domain Critical Errors in a Performance Assessment

Susan L. Davis-Becker

Chad W. Buckendahl

Alpine Testing Solutions

Paper presented at the 2011 Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

Inquiries regarding this paper should be addressed to:
Susan Davis-Becker
Susan.davisbecker@alpinetesting.com

Abstract

Credentialing programs are more and more considering the use of domain critical errors within the scoring approach of their testing program. These errors are identified as crucial to the content domain and a candidate who commits such an error should not be permitted to obtain a license or certification and be granted the associated rights and privileges. While the use of such errors may be appealing to policy makers or other stakeholders, practitioners caution test developers about the associated risks and the precautionary steps that should be taken before putting them in operational use (Buckendahl & Davis, 2010). In this paper we suggest two evaluative criteria that could be used to assess how well items containing domain critical errors are functioning within an operational examination and demonstrate types of analyses that can be used to address these criteria.

An Empirical Evaluation of Using Domain Critical Errors in a Performance Assessment

The stakes associated with the outcomes of many test programs are high for individuals and educational training programs. For example, examination programs are often the final step in the process for individuals to be allowed to practice medicine, fly 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 that are counting on the credentialing examination to serve as one means of protecting them from candidates who are not ready for such responsibilities. Test developers in the credentialing field are charged with the mission of determining whether examinees meet the minimum qualifications necessary for licensure or professional certification. Given the stakes associated with the outcome of such credentialing programs, some test developers have gone beyond the requirement to pass the examination (with a compensatory score) and have implemented additional, conjunctive criteria that make a candidate's ability to pass the examination contingent up demonstrating sufficient competency on one item or a small set of items within the examination.

In this paper, we use the term "domain critical" to refer to a component of knowledge or type of skill, ability, or judgment that is paramount in making decisions about a candidate's ability to enter a profession or job role. In practice, such critical content is then translated into specific exam items or tasks that an examinee must answer correctly to achieve the desired credential. In other related literature, these item types have been referred to "go no-go" items (Fortune, 1995), "killer" items (Childs, Dunn, & van Barneveld's 2003; 2007), or "gating" items (Judd, 2009). Semantics aside, making pass-fail decisions on the basis of a candidate's performance on a small number of items or tasks is perilous given the challenge of defending the practice psychometrically. In addition, there are those who advocate the use of domain critical errors in performance testing (e.g., Judd, 2009) but without providing sufficient guidance for practitioners on how to properly use and evaluate such items within such important programs.

The policymakers and content experts who contribute to the exam development process often have compelling, but subjective reasons for requesting test items that would automatically fail examinees. 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 and 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 be qualified, but did meet the required performance expectations). This heightened sense of public safety particularly as applied to licensure testing programs can be illustrated through the use of domain critical errors in practice.

Examples of such item types can be found on the U.S. Medical Licensure Exam (Floreck, Guernsey, Clyman, & Clauser, 2002) that involves 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. Other examples can be found in air traffic controllers examinations (e.g.,

Manning, 2000) and pilots (Woychesin, 2002). 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.

Buckendahl and Davis (2010) also described the use of domain critical errors in the context of performance tasks in clinical skills examinations for dentistry. They proposed an evaluation framework for testing programs that may be considering the use of these types of errors. Specifically, they outlined five areas of test development that are important to attend to when using domain critical errors: program design, domain analysis, content development, reliability, and standard setting. However, they did not provide empirical evidence to support the conceptual framework they recommended. This paper seeks to extend their work. In this paper, we define and describe criteria for psychometrically evaluating the use of domain critical items using empirical evidence. To illustrate the process and serve as a frame of reference for our discussion, we use the design and results from the same licensure testing program described in Buckendahl and Davis (2010).

Evaluation Questions

To demonstrate how practitioners could go about psychometrically evaluating the use of domain critical items or tasks, we will address two specific evaluation questions that will help provide needed validity evidence:

1. How is performance on domain critical items related to performance on related exam content?

As an important element of content validity evidence, one must examine the relationship between performance on the domain critical items and performance on the rest of the examination. For example, Webster et al. (1987) evaluated instances when dangerous options were selected by candidates on one of the tests designed by the American Board of Internal Medicine. However, this study focused on selected-response items. The current study will focus on applying similar concepts to performance tasks. Specifically, we will demonstrate how practitioners can support the relevance and importance of the content deemed domain critical through analysis of candidate performance.

2. What reliability evidence is available to support making decisions based on domain critical items?

Given the stakes associated with performance on such items, it is paramount that practitioners provide reliability evidence to support the consistency by which such performances are evaluated. In selected-response tests, scoring is typically handled electronically and consistency in scoring can be evaluated systematically. This process becomes more complex with subjectively-scored performance tasks. In this paper we will demonstrate a process for evaluating score and decision consistency in how such performances are judged.

Examination Program

The dental testing program discussed in this paper is sponsored by a regional agency consisting of member states. The clinical examinations developed by these consortia are one requirement of most states in receiving a license to practice dentistry. Although the clinical examination program also include separate tests that measure candidates' clinical judgments (e.g., diagnosis, assessment, treatment planning), the domain critical errors at the heart of this paper are limited to the clinical skills portion of the examination program (e.g., instrument use, handpiece manipulation, domain-relevant materials, manual dexterity).

There are four clinical skills dental examinations; two that are focused on operative procedures (i.e., removing tooth decay and restoring a tooth surface), one that is in endodontics (i.e., root canal), and the fourth includes fixed prosthodontics (i.e., crowns). Each examination requires the candidate to successfully and safely complete a procedure on a patient or simulated patient. Across all four examinations, each procedure is divided into sub-tasks that represent scoring criteria which are analytically scored as 1 (minimally competent performance or higher) or 0 (less than minimally competent performance). Given the subjective nature of the performance tasks and scoring criteria, three examiners who have been trained in the examination procedures independently score the candidate's performances and note any errors that are present. In the final scoring, an error is only counted against a candidate if it was observed by two or more examiners. This decision rule is applied across sub-tasks within a procedure to 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 and a candidate can make any of these errors (i.e., criteria errors) and still pass the exam as long as their total score still meets the overall passing score. However, some errors are considered to be domain critical and committing any of such errors will result in an automatic failure decision for the candidate. As used within these examinations, domain critical errors represent egregious mistakes within the assigned procedure (e.g., leaving decay in a tooth) or within the environment but beyond the scope of the procedure (e.g., not following protocol for infection control) which 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 granted a license.

For example, within the composite procedure, a candidate is required to prepare a tooth for a filling by removing the dental decay and any tooth structure that is unsound. There are two steps to this procedure that are evaluated: preparation (for the filling) and restoration (placement of the filling). Two errors that might be observed in the preparation are:

1) The decay has been removed and but the shape of the remaining tooth material is not completely ready to accept the filling. This will likely be corrected during the restoration.
2) In removing the decay, the candidate drilled too deep into the dentin to the point where the roots of the tooth are exposed. At this point a filling cannot be placed and a crown must be made forth the tooth to protect the exposed root.

The first error is a criteria error – while it represents a mistake in the process, it is one that can be corrected during the second part of the procedure with no harm to the patient. The second error is a domain critical error – this candidate has removed healthy tooth structure, caused additional damage to the tooth, and required the patient to have a different (more extensive) procedure. It is important to note that in the design of this examination, the sponsoring organization determined that all errors that could be committed within the Endodontics examination are domain critical.

Data

The data used in this study represent one year of operational administration of the four examinations. During this year, a total of 284 candidates participated in the examination program and completed 882 procedures which were judged by 41 examiners resulting in 3,732 examinations[1]. Across all 882 procedures, 2,439 criteria errors were confirmed (observed by 2+ examiners) and 280 domain critical errors were confirmed.

Results of Analysis

The results presented for this study are organized within the two evaluation questions.

1. How is performance on domain critical items related to performance on related exam content?

In the design of this exam, the criteria errors and domain critical errors are both defined as identifiable violations of the successful completion of the procedure. The difference between the two is the severity of the error and ability to successfully complete the procedure. Therefore, it is logical to explore the relationship between the domain critical errors and the criteria errors. To do so, two levels of analysis were conducted; one at the error level and the other at the decision level (e.g., pass/fail). The results in Table 1 detail the specifics of the error-level analysis. For the total dataset and each examination, we computed the average number of criteria errors and domain critical errors along with the correlation between the observances of each type of error across all candidates.

At the summary level, the most criteria errors are identified within the Amalgam and Fixed Prosthodontics procedures. Beyond the Endodontics examination (where all errors are considered domain critical), the most domain critical errors were identified within the Amalgam procedure. However, it is clear that more criteria errors were identified than domain critical across the three procedures that include criteria errors. This finding is good feedback to the training programs that these major errors are identified much less frequently. In terms of the correlation, the highest correlation is observed for the Amalgam procedure followed by the Fixed Prosthodontics procedure. A minimal correlation is observed for the Composite procedure.

[1] Although the process guidelines dictate that typically three examiners will review each work product, for those procedures where judgments are provided at two stages (Amalgam & Composite), different examiners may review the work product from first stage than review the work product from the second stage .

Table 1. Relationship between Confirmed Criteria and Domain Critical Errors by Procedure

	Average Number of Errors		Correlation between Number of Errors (DC & C)
	Criteria	**Dom Critical**	
Total*	0.79	0.05	0.38
Amalgam	0.97	0.06	0.55
Composite	0.55	0.04	0.06
Fixed Pros.	0.91	0.04	0.36
Endodontics	N/A	0.44	N/A

***** The Total values do not include Endodontics as all errors are considered domain critical

The second part of the analysis of the relationship between the domain critical errors and the other exam performance is focused at the decision (pass/fail) level. In Table 2 we identify the number of candidates who failed each procedure and the reason for failure (criteria errors, domain critical errors, or both). At an overall (total) level, the decision consistency is provided in terms of percentage. This value represents the proportion of all examinees who either passed or failed due to both an accumulation of criteria errors and committing a domain critical error. Overall, the decision consistency values are very high which may be due in part to the high pass rate observed for the Amalgam, Composite, and Fixed Prosthodontics procedures. Therefore, we provided an estimate of the decision consistency for the failing candidates. Similar to the error-level analysis, the consistency is highest for the Amalgam procedure.

Table 2. Pass/Fail Consistency between Criteria and Domain Critical Errors by Procedure

			Reason for Failure			Decision Consistency	
	Procedures	**Pass Rate**	**Criteria Only**	**Dom Critical Only**	**Both**	**Total**	**Failures (Both)**
Total*	711	96%	4	22	5	96%	16%
Amalgam	275	96%	0	7	4	97%	36%
Composite	276	96%	0	11	0	96%	0%
Fixed Pros.	160	94%	4	4	1	95%	11%
Endodontics	171	80%	N/A	35	0	N/A	N/A

* The Total values do not include Endodontics as there is no means by which to compare decision consistency.

The results of this analysis suggest that there is an overall relationship between the observance of criteria and domain critical errors. Using Cohen's (1977) criteria, the correlation for the overall program would be described as a medium effect. The same would be the case for the Fixed Prosthodontics procedure. In contrast, the relationship between criteria and domain critical errors in the Amalgam procedure would be a large effect whereas the finding for the Composite procedure would be considered a small effect as the correlation suggested almost no relationship. At the decision level, only 16% of the candidates who failed any part of the examination did so because of both criteria errors and domain critical errors. Higher consistency was observed for the Amalgam procedure whereas no consistency was observed for the Composite procedure as not a single candidate failed due to an accumulation of criteria errors.

2. What reliability evidence is available to support making decisions based on domain critical items?

Given the significant impact that the domain critical errors may have on the overall performance of a candidate on the exam, it is important to consider the reliability by which these errors are observed. In Table 3, we summarize the consistency observed at the error-level which is defined as the number of observed errors that were also confirmed. Overall, 24% of all errors identified were confirmed by one or more examiners. A higher consistency rate was observed for the domain critical errors than for the criteria errors and the highest consistency rates were observed within the Endodontics and Amalgam procedures.

Table 3. Confirmation of errors (consistency) by procedure

	Total Errors		**Criteria**		**Domain Critical**	
	N	**Confirmed**	**N**	**Confirmed**	**N**	**Confirmed**
Total	2719	662 (24%)	2439	551 (23%)	280	111 (40%)
Amalgam	1113	279 (25%)	1069	262 (25%)	44	17 (39%)
Composite	733	162 (22%)	687	151 (22%)	46	11 (24%)
Fixed Pros	704	144 (20%)	683	138 (20%)	21	6 (29%)
Endodontics	169	77 (45%)	N/A		169	77 (46%)

Given that the results of this exam are used in making licensure decisions about candidates (pass/fail), it is also important to consider consistency at the decision level. In Table 4, the consistency estimates are summarized at the procedure level indicating the frequency by which the examiner-level decision agreed with the overall decision. The four categories of decisions are: pass, fail due to an accumulation of criteria errors, fail due to one or more domain critical errors, and fail due to both criteria and domain critical errors. If anything, separating the fail condition by reason lowers the agreement consistency (which would be based on pass versus fail) but it is necessary for the purpose of this paper and evaluating the consistency within the use of domain critical errors. Some data were excluded from these analyses. Because some examiners may only see part of a candidate's Amalgam or Composite procedure (e.g., preparation or restoration) we excluded the decisions of those examiners who did not have the opportunity to observe the reason for failure (e.g., an examiner who only reviewed the preparation when the domain critical error occurred in the restoration). Out of 3,733 examinations this excluded 50 examinations.

The findings from this analysis suggest that there is a high level of agreement at the decision level. However, the estimates of consistency by reason for failure are somewhat lower. By procedure, the highest agreement by failure reason was observed for the Endodontics procedure.

Table 4. Consistency in Examiner Candidate-level Decisions by Procedure

	Pass		Fail – Criteria		Fail - Dom. Critical		Fail - Both	
	N	**Agree**	**N**	**Agree**	**N**	**Agree**	**N**	**Agree**
Total	3349	3223 (96%)	17	7 (41%)	316	177 (65%)	30	3 (10%)
Amalgam	1259	1237 (98%)			50	21 (59%)	22	2 (9%)
Composite	1225	1190 (97%)			85	29 (50%)		
Fixed Pros	460	435 (95%)	17	7 (41%)	14	7 (50%)	8	1 (13%)
Endodontics	413	370 (90%)			167	120 (72%)		

The results of the consistency analysis suggest that there is some consistency in the identification of domain critical errors for these examinations. Overall, the highest level of consistency was observed for the Endodontics procedure at both the item and decision levels (reason for failure) whereas the lowest levels were observed for the Fixed Prosthodontics examination.

Summary

The use of domain critical items is growing within credentialing programs. More practitioners are exploring the possibility of identifying such critical content and making successful performance mandatory for passing an examination. 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. In this paper we provide a process for psychometrically evaluating the use of such items and provide an example with a licensure examination program currently using domain critical items for performance tasks.

In this demonstration of evaluating domain critical errors, two areas were explored: relationship between domain critical errors and other measure of exam performance and the reliability of identification and use of domain critical errors. In examining the first aspect, we found an overall moderate relationship between the observance (and confirmation) of domain critical errors and criteria errors. When this was investigated at the decision level, less consistency was observed. For this program, we would suggest further investigation to better understand the expected relationship between candidates' committing criteria errors and domain critical errors and then determine if this specific relationship exists within the data.

In examining the second aspect (reliability), the analysis revealed highest consistency for the Endodontics procedure at both the item level and the decision level. The overall item-level consistency estimates were somewhat lower than expected. This could suggest that the rubric should be clarified and/or the need to improve rater training in identification and recording of errors.

References

Buckendahl, C. W. & Davis, S. L. (2010). Evaluating the appropriateness and use of domain critical errors. Paper presented at the annual meeting of the National Council of Measurement in Education. Denver, CO.

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.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York, NY: Academic Press.

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.

Judd, W. (2009). Gating items: Definitions, significance, and need for further study. Practical Assessment, Research, & Evaluation, 14(9). Available online: 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.

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.