Journal of Education and Learning; Vol. 1, No. 1; 2012
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

Reliability Evidence for Examination Cut Scores within a Medical School

Milton Severo[1,2], Rita Gaio[3], Daniel Moura[4], Rui Fontes[5], Teresa Rodrigues[2], Adelino Ferreira Leite Moreira[6], Isaura Tavares[7], Luis Delgado[8], & Maria Amélia Ferreira Tavares[1,9]

[1] Center for Medical Education, Faculty of Medicine of the University of Porto, Porto, Portugal

[2] Department of Clinical Epidemiology, Predictive Medicine and Public Health, University of Porto Medical School, Porto, Portugal

[3] Department of Pure Mathematics, University of Porto Science School, Porto, Portugal

[4] Institute of Pharmacology and Therapeutics Faculty of Medicine of the University of Porto, Porto, Portugal

[5] Department of Biochemistry (U38-FCT), Faculty of Medicine of the University of Porto, Porto, Portugal

[6] Department of Physiology, Faculty of Medicine of the University of Porto, Porto, Portugal

[7] Institute of Histology and Embryology, Faculty of Medicine of the University of Porto, Porto, Portugal

[8] Department of Immunology, Faculty of Medicine of the University of Porto, Porto, Portugal

[9] Institute of Anatomy, Faculty of Medicine of the University of Porto, Porto, Portugal

Correspondence: Milton Severo, Center for Medical Education, Faculty of Medicine of the University of Porto, Porto, 4200-319, Portugal. Tel: 351-225-513-611. E-mail: milton@med.up.pt

Received: November 22, 2011 Accepted: December 13, 2011 Published: June 1, 2012

doi:10.5539/jel.v1n1p77 URL: http://dx.doi.org/10.5539/jel.v1n1p77

Abstract

Establishing credible cut scores for performance-type examinations in health professions education can be challenging. The authors aimed to compare the pass-fail cut-score reliability with the maximum reliability cut-score from multiple-choice tests (MCTs) designed on different undergraduate disciplines. Using the cross-sectional evaluation of 1370 tests from six disciplines from Porto medical school, Portugal, in 2010, the pass-fail cut-score reliability was obtained from the one-parameter logistic model of item response theory model. The test information curve achieved maximum reliability for ability levels ranging from -1.40 to -0.01 standard deviations below the average. The pass-fail cut score for estimated ability ranged from -1.36 to 0.25. These results showed that all MCTs had a pass and fail threshold of competence, and that was appropriate for the maximum information obtainable from the examination to occur at the pass and fail level; nevertheless, the maximum information was not achieved in the pass and fail level.

Keywords: assessment, item response theory, medical education, undergraduate, pass and fail cut-score

1. Introduction

A central goal of all educational organizations involved in certification or licensure activities is to ensure that the competent are truly competent (De Champlain, 2004). When this is measured via examinations, it is crucial to ensure that the objectives defined for the curricular units are being met and assessed. With respect to the tests, each educational organization can use the process of benchmarking for the implementation and development of the quality of its courses. All assessments in medical education, in particular, require evidence of validity to be meaningful (Downing, 2003).

Establishing credible, defensible, and acceptable passing or cut-off scores for performance-type examinations in health professions education can be challenging (Norcini & Shea, 1997). Although there are several methods to establish such passing scores, setting standards for *local* performance examinations can be a time-consuming task (Yudkowsky, Downing, & Wirth, 2008).

Regarding the appropriateness of cut scores, Kane specifies four types of evidence that must be included in a structured validity argument: (1) scoring, (2) generalization, (3) extrapolation, and (4) decision (Brennan, 2006). The scoring component requires evidence that the test data were collected under appropriate conditions and were

scored accurately. Generalization focuses on internal structure/reliability or the stability of scores across replications of the assessment procedure. Extrapolation requires evidence of a relationship between test scores and the real-world behaviour or performance of interest. The decision component calls for evidence that decisions based on the established cut score are appropriate (Margolis, Clauser, Winward, & Dillon, 2010).

The Standards for Educational and Psychological Testing (American Psychological, 1999) noted that the internal structure relates to the statistical or psychometric characteristics of the test items and scores and many of the required statistical analysis are often carried out as routine quality-control procedures (Downing, 2003). Several statistical models are typically used to evaluate specific internal structure outcomes, such as the difficulty/discrimination of examination items, the testing-taking ability of examinees or reliability, and the reproducibility of the scores on the assessment. If the scores are not reliable and reproducible it is impossible to interpret the meaning of those scores, and therefore the pass and fail decision will lack validity, possibly passing students who should fail and failing candidates who should pass. Each educational organization should therefore assure the reliability of its test cut scores.

The approach typically utilised in classical test theory to estimate the reliability of test scores in written examinations employs the concept of internal consistency (Downing, 2004), usually estimated by the Cronbach alpha (Cronbach, 1951), while in item response theory to estimate the reliability of the ability in written examinations employs the concept of test information curve (Raju, Price, Oshima, & Nering, 2007).

The present study aimed to compare the pass-fail cut-score reliability with the maximum reliability cut-score from multiple-choice tests (MCTs), designed on different undergraduate disciplines from the same medical school - Faculty of Medicine of the University of Porto (FMUP).

2. Methods

FMUP has a protocol of automatic scanning, scoring and quality evaluation of multiple choice tests since 2006. This program started with the Pharmacology discipline and was then extended to Biochemistry, Physiology, Histology, Epidemiology, Immunology and Clinical Anatomy (Severo & Tavares, 2010).

This medical school offers a 6-year undergraduate medical curriculum, consisting of 3 years that are mainly theory oriented followed by 3 other years with high clinical focus. From a total of 16 curricular units from the 1st semester of the first 3 years, 6 (37.5%) disciplines participated in the protocol of automatic scanning, scoring and evaluation of the quality of the multiple choices tests. The first period of examination occurred in January 2010.

A test was said to have maximum quality in the response pattern if there was evidence of data integrity such that all sources of error associated with the test administration are controlled or eliminated to the maximum extent possible. In order to ensure maximum quality in the response pattern: two persons were in charge of the scanning process; the students were given the opportunity to check if the scanning and scoring were correct; the test key was delivered in digital format to minimize possible errors in its validation; for each test, a list with the names and IDs of the eligible students was delivered before the test in digital format, to allow for cross-validation with the students ID at the time of the test. From a possible total of 1755 individual tests, 1370 (78%) were completed in the first period of examination, for the six disciplines mentioned above.

2.1 Statistical Analyses

Classical test theory (CTT) analyses included item p-values (proportion of individuals in the sample with the correct answer for each item) and bi-serial correlation coefficients between each item and the final examination score excluding the item being tested. Exploratory factor analysis was used to evaluate homogeneity (i.e., to confirm there was a single continuous latent variable) and Cronbach's alpha (Cronbach, 1951) was used to measure the reliability of the test.

Item response theory (IRT) was used to assess each item quality. The relationship between the probability of endorsing item i correctly π_I and the latent ability of an examinee, z, can be described by a function called an Item Characteristic Curve (ICC), denoted by $\pi_i(z)$. These ICCs are characterised by two parameters, denoted by difficulty and discrimination. The difficulty parameter represents the ability value at which the probability of correctly answering the item is 50%. The discrimination parameter represents the slope at the respective difficulty parameter and thus indicates how well an item discriminates individuals with ability near the difficulty parameter.

The one-parameter logistic (1-PL) item response model was used to estimate the difficulty and discrimination parameters (due to sample size, the discrimination parameter was assumed to be the same for all items in each test).

Another feature of the IRT models is the test information curve (TIC), which indicates the precision (the inverse of the error variance – SEM square) of a test along the continuous underlying variable. The test information curve can be used to identify the point at which the test offers maximum reliability.

To estimate the reliability at the pass and fail cut score, we derived a table of one-to-one correspondence between the values of the scores and the values of the latent ability, and thus determined the pass and fail ability and the respective reliability.

The 1-PL model requires unidimensionality of the construct being measured and local independence of the test items (conditioned by the construct). The eigenvalues from a tetrachoric correlation matrix of the observed dataset were computed to support the unidimensionality (exploratory factor analysis), and the item-fit statistics and pairwise two-way margins residuals were used to confirm the local independence.

Statistical analyses were performed using the R Project for Statistical Computing software, version 2.8.1 (R Foundation, Vienna, Austria).

3. Results

3.1 Taxonomy of Multiple-choice Tests

The proportion of tests attendance ranged from 66% to 86%. From the 6 disciplines that were assessed by multiple-choice tests, 2 chose items from type A while the other 4 selected items of mixed type (Case & Swanson, 2001). Three disciplines used penalization in case of an incorrect answer. The number of items per test ranged from 50 to 100 (table 1).

Table 1. Description of the test items

	Year	ECTS	Item Type	Number Items	Number Students	Number attended the test n (%)	Penalization
A	1st	8	A	62	315	231 (73)	No
B	2nd	8	A, T/F and Space	88	285	244 (86)	Yes/No*
C	2nd	6	A	60	280	184 (66)	Yes
D	3rd	3,5	A and K	50	288	213 (74)	No
E	3rd	6	A and R	100	297	250 (84)	No
F	3rd	3,5	A and R/B	50	290	248 (86)	Yes

*20 True/False items have penalization.

3.2 Classical Test Theory

Unanswered examination items were treated as incorrect. The median (25th to 75th percentile) p-value ranged from 40 (31-57) to 65 (51-79) percent. The median (25th to 75th percentile) bi-serial coefficient excluding the item being tested varied from 0.33 (0.21-0.42) to 0.55 (0.40-0.65). The minimum and maximum estimated tests internal consistency were 0.78 and 0.94, respectively. The percentage of items that after elimination would increase the reliability ranged from 12% to 19% (table 2).

Exploratory factor analysis conducted for each test strongly suggested a unique factor; the first eigenvalue was always greater than 2.5 times the second eigenvalue (table 2).

Table 2. Item and test parameters from classical test theory

	p-value	Bi-serial Correlation	Alpha (%)*	EFA† Eigenvalues	
	Med ($1Q^{st}$; $3^{rd}Q$)	Med ($1Q^{st}$; $3^{rd}Q$)		1	2
A	50 (38;63)	0.38 (0.24;0.46)	0.86 (19)	12.036	2.826
B	55 (42;69)	0.29 (0.20;0.34)	0.82 (17)	9.786	3.928
C	72 (53;87)	0.36 (0.27;0.45)	0.82 (15)	11.455	4.819
D	40 (31;57)	0.33 (0.21;0.42)	0.78 (14)	8.061	3.045
E	65 (51;79)	0.55 (0.40;0.65)	0.94 (16)	30.190	4.593
F	72 (58;82)	0.40 (0.31;0.52)	0.84 (12)	11.020	3.942

* Percentage of items that after elimination would increase the reliability

†Exploratory Factor Analysis

3.3 Item Response Theory

The median difficulty parameter ranged from -1.46 (-2.95; -0.19) to 0.73 (-0.49; 1.40). The 1-PL IRT model revealed a wide range in item discrimination parameters, varying from 0.50 to 1.05 (table 3). These values correspond to factor loadings of 0.45 and 0.72, respectively. The percentage of items with a poor fit ranged from 12% to 24%.

Information functions were computed for each of the six tests, and are displayed in Figure 1. Tests B, D and F showed reasonably smooth TICs, while tests A, C and E exhibited TICs with a peak at lower levels of ability (Figure 1). The maximum discrimination ranged from 3.8 to 21.8 and this maximum information was reached at ability values varying from -1.9 to 0.6, respectively (table 3).

The pass and fail ability level ranged from -1.36 to 0.25.

A comparison between the optimum cut score ability and the pass and fail cut score ability, for each test, revealed that test A (-0.01 vs. -0.14), test B (-0.45 vs. -0.56) and test F (-1.23 vs. -1.36) showed similar ability levels, test C (-1.40 vs. -0.56) and test E (-0.71 vs. -0.37) showed an optimum cut-point lower than the pass and fail cut score ability, while test D (0.60 vs. 0.25) showed an optimum cut-point higher that the pass and fail cut score ability (table 3).

Table 3. Item and test parameters from item response theory

	Difficulty Parameter Med ($1Q^{st}$;$3^{rd}Q$)	Discrimination Parameter (se)	Factor Loading	Item-fit statistics* N (%)	Max Inf. (optimum cut score)	Inf. (pass and fail cut score ability)
A	-0.01 (-0.85;0.78)	0.70 (0.04)	0.57	11 (18)	6.5 (-0.01)	6.5 (-0.14)
B	-0.38 (-1.76;0.66)	0.50 (0.03)	0.45	11 (12)	4.7 (-0.45)	4.7 (-0.56)
C	-1.46 (-2.95;-0.19)	0.71 (0.05)	0.58	8 (13)	5.2 (-1.40)	5.1 (-0.56)
D	0.73 (-0.49;1.40)	0.62 (0.04)	0.53	6 (12)	3.8 (0.60)	3.8 (0.25)
E	-0.65 (-1.42;0.01)	1.05 (0.04)	0.72	24 (24)	21.8 (-0.71)	21.5 (-0.37)
F	-1.30 (-2.04;-0.49)	0.83 (0.04)	0.64	9 (18)	6.9 (-1.23)	6.8 (-1.36)

*Number of items with an item-fit with p-value less than 0.05

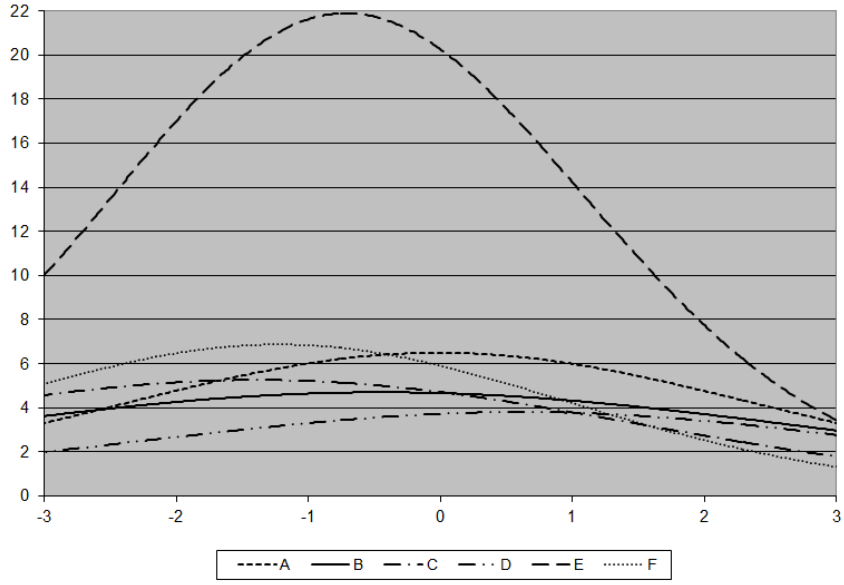

Figure 1. Test information curve for each discipline, which shows reliability according to different levels of ability

4. Discussion

Most academic staff designs assessment tests relying on their empirical knowledge obtained from years of experience; the present study showed that the evaluated tests themselves and their correspondent items were of good quality.

Studies have used Cronbach alpha greater than 0.8 as an adequate measure of internal consistency (Kehoe, 1995); yet, this measure is dependent on the number of items, the dimensionality of the scale and the inter-correlations between the items (i.e., items discrimination) (Cortina, 1993). In our case, all tests but one revealed a higher reliability than the usual criteria of 0.8.

De Champlain (De Champlain, 2010) defined that TIC standard depends on the intended use of test scores. If a test is a selection examination, it is important to measure a broad range of abilities with a similar level of precision or reliability out of fairness to candidates. In practice, the TIC should be high and reasonably smooth over the relevant ability range (-3,3) (Partchev, 2004). If test is a licensure examination, reliability or information needs to be maximised at the cut score value, because this is where decision accuracy needs to be at its highest point. The main objective of all evaluated examinations was to assess whether or not examinees had met an adequate standard of performance (De Champlain, 2004), not necessarily to demonstrate advanced mastery of the topic. Consequently, all tests were constructed with a pass and fail threshold of competence, and it was appropriate for the maximum information obtainable from the examination to occur at the pass and fail level. In the present study, three (50%) of the six considered tests showed reasonably smooth TICs over the relevant ability range, whereas the other tests (50%) presented TICs with a highest point (that is, with a highest absolute value for the second derivative, as a function of the ability). Also, when we compared the pre-specified pass and fail cut score ability with the optimum cut score, for each test, we observed great discrepancies among the tests, half of them showing an optimum-cut score discrepant from the pre-specified pass and fail cut score.

The major limitation of the present study is its small sample size which limited our statistical analysis to the use of the 1-PL model. Whereas the minimum number of examinees required to properly fit a 1-PL model is approximately 200 (Downing, 2003), a proper 2-PL model (including the possibility of a different discrimination parameter for each item) would require a much larger sample size. An inadequate sample size would be expected to yield unstable item parameters and higher standard errors. This was reflected in the item-fit (approximately 15% of the items showed a poor fit with the 1-PL model). However the main reason identified for the poor fit was the low or high discrimination of the item compared with the remaining. Yet we have confidence in our results because the Bayesian Information Criteria suggested that equal discrimination parameter across items was the best solution for all models except for one.

In conclusion, the evaluated multiple-choice written tests from different disciplines within the same school, which were designed on an empirical basis, showed good internal structure. All multiple choice tests were designed on an empirical basis and had a natural pass and fail threshold of competence; It would have been appropriate to have the maximum information obtainable from the examination to occur at the pass and fail level, however, the maximum information was not achieved at that level, and the reliability/information was reasonably smooth over the relevant ability range and not maximised at the cut score value as it should have been.

Calibration of the item bank can improve the reliability at the pass/fail cut score ability on empirical based tests. The results from this study were shared with the individual disciplines whose examinations were assessed and will serve as guidelines to prepare future examinations. IRT/CTT can be used to provide information about the evaluation process in general to the teaching staff, and information on how to identify, revise or discard problematic questions. IRT/CTT can also be useful tools to teachers that need to compile items in a multiple choice examination: item parameters (difficulty and discrimination) will allow the teacher to establish an item bank that can be used in the future to build and calibrate examinations. In the long run, it is expected an improvement in the course and program outcomes, that can be then reflected on the faculty status and on the enhance of accreditation qualifications.

Acknowledgements

The authors thank the collaboration of the Biochemistry, Physiology, Basic Histology and Embryology, Epidemiology, Pharmacology and Basic Immunology teaching teams.

References

American Psychological, A. (1999). *Standards for educational and psychological testing*. National Council on Measurement in Education.

Brennan, R. L. (2006). *Educational measurement*. Praeger Pub Text.

Case, S. M., & Swanson, D. B. (2001). Constructing written test questions for the basic and clinical sciences. *Philadelphia: National Board of Medical Examiners*, 19–29.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology, 78*(1), 98-104. http://dx.doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-333. http://dx.doi.org/10.1007/BF02310555

De Champlain, A. F. (2004). Ensuring that the competent are truly competent: an overview of common methods and procedures used to set standards on high-stakes examinations. *J Vet Med Educ, 31*(1), 61-65. http://dx.doi.org/10.3138/jvme.31.1.61

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Med Educ, 44*(1), 109-117. http://dx.doi.org/10.1111/j.1365-2923.2009.03425.x

Downing, S. M. (2003). Validity: on meaningful interpretation of assessment data. *Med Educ, 37*(9), 830-837. http://dx.doi.org/10.1046/j.1365-2923.2003.01594.x

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Med Educ, 38*(9), 1006-1012. http://dx.doi.org/10.1111/j.1365-2929.2004.01932.x

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Practical assessment, research & evaluation, 4*(10), 19-36.

Margolis, M. J., Clauser, B. E., Winward, M., & Dillon, G. F. (2010). Validity evidence for USMLE examination cut scores: results of a large-scale survey. *Acad Med, 85*(10 Suppl), S93-97. http://dx.doi.org/10.1097/ACM.0b013e3181ed4028

Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in education, 10*(1), 39-59. http://dx.doi.org/10.1207/s15324818ame1001_3

Partchev, I. (2004). A visual guide to item response theory. *Friedrich Schiller Universität Jena*.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement, 31*(3), 169. http://dx.doi.org/10.1177/0146621606291569

Severo, M., & Tavares, M. A. (2010). Meta-evaluation in clinical anatomy: a practical application of item

response theory in multiple choice examinations. *Anat Sci Educ, 3*(1), 17-24.

Yudkowsky, R., Downing, S. M., & Wirth, S. (2008). Simpler standards for local performance examinations: the Yes/No Angoff and whole-test Ebel. *Teach Learn Med, 20*(3), 212-217. http://dx.doi.org/10.1080/10401330802199450