UNIVERSITÄT
BERN

Faculty of medicine
Institute of Medical Education
Assessment and Evaluation Unit

Standardized exam analysis as performed by the IML (Institute of Medical Education, CH-Bern)

After computerized scanning of all answer sheets at IML, and basic „counting“ of all candidates’ answers to each question, the evaluation (statistical analysis) of each multiple choice exam consists of the following standard 3 steps:

Institute of Medical Education
Assessment and Evaluation Unit
Konsumstrasse 13
CH-3010 Bern
info@iml.unibe.ch
www.iml.unibe.ch

Step 1: item (i.e. exam question) analysis

For every item, the statistical values for it's *difficulty p* and it's *discrimination index* r_{iT} (corrected item-total correlation, "selectivity") are calculated. Moreover, the statistical result pattern for each item, and possible *group differences,* are determined.

The difficulty p of an item = number of candidates who answered the item correctly divided by the maximum possible correct answers (i.e. total number of candidates)

- p therefore is a value between 0 and 1
- an „easy" item has a high p, a „difficult" item has a low p
- p-values between 0.4 and 0.9 are considered adequate

The discrimination index r_{iT} of an item measures the power of an item to differentiate candidates with a good result overall from candidates with bad results.

- r_{iT} is calculated as a correlation between the item-answers of all candidates and the total points (minus the item to be analyzed)
- r_{iT} therefore is a value between -1 and +1
- for a trustworthy exam performance, items with a positive item/total correlation are called for, if possible ≥ 0.20, at least ≥ 0.10
- items with a discrimination index = 0 do not discriminate between good or bad candidates, items with negative item/total correlation are counterproductive

The analysis of non-random „attractive" wrong answers points out problematic items, or answer positions, as well as wrong answer keys. Those wrong answers are characterized by $p > 0.4$ and $r > 0.1$.

Group differences are analyzed after all multi-lingual exams for example, or exams held at different locations, or if applicable and desired according to other group characteristics (age, gender, training....).

Items with exceptional difficulty or discrimination index or answer patterns, or group differences, are commented on by exam experts at IML, and discussed by content experts in the exam committee. If the experts come to the conclusion that item performance deficits are indeed due to content validity problems, an exclusion from the overall exam evaluation can be considered. At least, such an item will be reviewed before being reused in a future exam.

Step 2: exam recalculation / pass mark

The relevant decisions by the exam administrators/experts are implemented.

The exam characteristics (distribution, mean, reliability etc.) are being checked for plausibility a second time. The overall reliability as a measurement of internal consistency - Cronbach's alpha – is calculated. Ideally, a value of above .80 should be reached in a high stake exam.

In regularly administered exams, the goal is to keep the requirements for passing rather constant, on the basis of re-used items. One exam will be analyzed according to the so-called Rasch model.

The very first exam therefore guides the future passing parameters. Keeping the probabilities for solving an item constant will facilitate to set comparable passing requirements for every exam to follow.

Nonetheless, the expert exam committee decides upon the definitive passing score after each exam.

Step 3: definitive results

After the exam, a (usually linear) grading will be applied to classify candidate's scores according to a grade system.

Definitive results can now be depicted as a frequency distribution diagram covering the grades, countries of origin, exam locations etc.

The pass/fail letters with grades will be generated for the candidates. The following information will be given:

1. the pass/fail score limit for the exam at hand
2. the score reached by the individual, „pass" or „fail", and grade
3. if applicable, the partial results (i.e. in a subspecialty) in comparison to the pertaining overall scores.

Literature:

Bortz, J. (2005). Statistik. Springer.

Cronbach, L.J. (1951). Coefficient alpha and the initial structure of tests. Psychometrika, 16(3), 297 - 334.

Dunn, T. J., Baguley, T. and Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology. doi: 10.1111/bjop.12046

Kelley, T., Ebel, R., & Lincare, J.M. (2002). Item discrimination indices. Rasch Measurement Transactions, 16(3), 883 - 884.

Lienert, G.A. & Raatz, U. (1998). Testaufbau und Testanalyse. Beltz, PVU.

Linn, R.L. & Gronlund, N.E. (2000). Measurement and Assessment in Teaching. Englewood Cliffs, NJ, Prentice-Hall.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Kopenhagen, The Danish Institute for Educational Research.

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. Journal of Personality Assessment, 80, 99 - 103.