# Session A2:
# Tests and prediction of student succes

Poster 02

## Screening for flawed multiple choice items before test administration or not? A generalizability study

### Program text
Awareness of construction flaws in multiple choice items is considered important, but we found systematic screening for flaws before test administration to be lacking in dependability.


### Abstract
Construction errors in multiple choice items are prevalent and constitute threats to test validity. However, very little research on the usefulness of systematic item screening by local review committees before test administration seem to exist. With modern validity theory as our theoretical framework, we suggest that some fundamental validity assumptions for a review committee's qualitative screening for technical item flaws are that 1) the sample of reviewers is large enough to control for reviewer bias, 2) we may extrapolate item quality from the results of their flaw detection. The aim of this study was to examine these validity assumption as well as feasibility aspects of review committee screening for item flaws in a Danish context. Five independent reviewers screened 180 multiple choice items (2 medical school exam papers) for 19 internationally recognized technical item flaws. The reliability of item reviewers' independent judgments of the presence of item flaws was examined with a generalizability study design. Results indicated that a review committee screening approach required many reviewers (n=31) for higher levels of dependability ($\phi$=0.90). In contrast, post-exam quantitative screening seemed to be a more feasible and efficient way of improving the overall ability of the tests to discriminate between examinees. The validity of human judgments of item flaws is important not just for sufficiently sound quality assurance procedures and validity of exams in local contexts, but also for global research on the effects of item flaws. While awareness of the many possible technical item flaws to avoid may be very useful in the training of item writers, our results seem to question their use in systematic judgements of larger numbers of items by a review committee in a local educational context.



### Authors
Lotte O'Neill, SDU; Sara Mathilde Radl Mortensen, AU; Cita Nørgård, SDU; Anne Lindebo Holm Øvrehus, OUH; Ulla Glenert Friis, SDU


### Literature

Haladyna TM, Downing AM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment, Appl Meas Educ. 2002;15(3):309-333.

Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. 3rd ed. Philadelphia: National Board of Medical Examiners; 2002.

Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ. 2005;10(2):133-143.

Kane MT. Validation. In: Brennan RL, editor. Educational Measurement. Westport: ACE/Praeger; 2006. p. 17-64.

Norcini J, Grosso L. The generalizability of ratings of item relevance. Appl Meas Educ. 1998;11(4):301-309.