how to:

Wales Deanery Deoniaeth Cymru

Cardiff University Prifysgol Caerdydd

Medical Education @ Cardiff

Apply the Bookmark Standard Setting Method to a Machine Marked Test

Phil Matthews

The Need for Standard Setting

A test cut-score is usually intended to accurately reflect a particular level of knowledge or skill. A cut-score is simply the score that serves to classify those individuals whose score is below the cut-score into one level and those individuals whose score is at or above the cut-score into the next and higher level.

If cut-scores are not appropriately set, a valid and reliable, test can still produce a sound hierarchy and range of relevant ability amongst the test takers. However, in these circumstances, the point in the hierarchy of scores, at which test takers are deemed to have the requisite level of occupational competence, is less likely to be a reliable one. For this reason, if a test forms part of a professional selection or certification procedure, standard setting is a critical component of the development process.

Genesis and General Approach for the Bookmark Method

The Bookmark standard setting method was developed in 1996 by CTB/McGraw-Hill research scientists[1]. It relatively quickly became widely used in American secondary and higher education settings; and by 2002 was the method of choice in 31 state education boards in the United States[2].

In summary, this method firstly requires an analysis of actual test results from a live machine marked examination in order to accurately rank its questions in order of their easiness. Many proprietary software products are capable of quickly providing such an analysis. The effectiveness of the technique is then predicated on the ability of educators, expert in the content area being examined, to collectively and reliably identify cut off points for predetermined descriptors of candidate competency by considering questions in their ranked order. In undertaking their task, the experts quickly focus on ranked clusters of questions they consider to be critical in delineating each of the predetermined competency levels.

Overview of, and Comparison with, an Alternative Method

Widely used in the United Kingdom and developed in the 1950s, it requires individual judges to estimate the probability of candidates of a particular ability answering each exam question correctly. Collective judgement is then made by discussion and subsequent calculation of the means of expert judges 2nd scores.

How the Bookmark Method is Applied

This technique first requires that the test (or exam) in question is applied to either a representative sample or to a whole population for whom the test is intended. The test is then machine marked and the relative easiness of all its questions is calculated. Practitioners experienced in the field in question, and familiar with the range of abilities of the population being tested, are assembled to apply the method. These experienced practitioners, or "experts" as they are often called, are collectively charged with clarifying what the test scores actually mean. The practitioners apply the Bookmark method using question easiness ranking information, which is calculated from candidates' collective results. The experts can be asked to set anything from one to several threshold scores to delineate various bands of learning attainment which individual candidates from an appropriate peer group may demonstrate. These judgments are recorded by placing "bookmarks" in their ordered item booklets that "separate", for instance, the partially proficient student from the proficient student, the student whose performance is unsatisfactory from the partially proficient student, and the advanced student from the proficient student.

Key similarities and differences between the Bookmark and Angoff methods

Features shared by both techniques	Features of Bookmark not shared with Angoff
Requires visualisation of candidate types	Uses actual statistics from exam to rank questions
Utilises individual judgments, group discussion and feedback	Facilitator conveys impact of 1st round bookmarks to judges in 2nd estimation round
Employs median of final scores of small groups of judges	Requires in-depth judgements on some, but far fewer questions Is less complex & quicker Has lower standard deviations between experts' final scores

DEANERY UPDATES

Facilitating a Bookmark Standard Setting Session

- Lewis et al[3] recommend dividing the assembled group of experts into 3 to 4 sub-groups, each containing 5 to 7 experts. There is no fixed guideline and resource availability can be taken into account.
- Explain the Bookmarking process
- Discuss the levels of proficiency which need to be delineated
- Distribute the Test paper itself and Item booklet 1 (with the Test items ordered by easiness) to each expert to read, consider and then place their 1st bookmarks in their own Item booklet 1
- Distribute Item booklet 2 (which adds the impact of any bookmark levels on the number of candidates passing at that level)
- Instruct individuals to consider this new information in relation to their 1st bookmarks
- Individuals in their small groups discuss in a round the reasoning for & impact of each judge's 1st ratings
- Individuals then place their 2nd bookmarks (they may or may not choose to revise earlier judgements)
- Calculate the median bookmark in each small group, for each required level of proficiency
- In plenary, present the median rating for each required proficiency level from each small group to all
- For each stipulated level of proficiency, take the median score of all the small group ratings to set the final minimally required score.
- Remember to record each of these final scores both as an absolute and as a percentage for clarity.

An abbreviated example of an Item booklet 1 (as used by General Practice Training Wales)

Question Stem from an Extended Matching Option Test	**Easiness Ranking**	**Question number in Test**
A 45 year old male presents with weight loss and heat intolerance. Clinical examination demonstrates a fine tremor at rest and warm sweaty palms.	1	
▲		
A 35 year old woman with 5 months amenorrhoea presents with intermittent abdominal pain. On examination there is a firm mass arising from her pelvis.	82	117
▼		
An 18 year old woman was started on phenytoin for newly diagnosed epilepsy 2 days ago. She reports that this month's menstrual bleed has begun today; and that in 2 or 3 months time, she might want to start trying to conceive with her current boyfriend. Apart from any medical information you would like to ensure she is aware of, what management is most appropriate immediately?	150	

An abbreviated example of an Item booklet 2
(as used by General Practice Training Wales)

Question Stem from an Extended Matching Option Test	**Easiness Ranking**	**Question no. in Test**	**Percentage of candidates who answered this correctly**	**Percentage score candidate attains if she/he only gets this number of questions correct**	**Percentage of candidates who would pass at this bookmark score**
A 45 year old male presents with weight loss and heat intolerance. Clinical examination demonstrates a fine tremor at rest and warm sweaty palms.	1		98.625	0.67	100
▲					
A 35 year old woman with 5 months amenorrhoea presents with intermittent abdominal pain. On examination there is a firm mass arising from her pelvis.	83	55	59	55.33	66
▼					
An 18 year old woman was started on phenytoin for newly diagnosed epilepsy 2 days ago. She reports that this month's menstrual bleed has begun today; and that in 2 or 3 months time, she might want to start trying to conceive with her current boyfriend. Apart from any medical information you would like to ensure she is aware of, what management is most appropriate immediately?	150		6.5	100.00	0

Further Information

1. Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A Bookmark approach.* In D. R.Green (Chair), IRT-based standard setting procedures utilizing behavioural anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

2. Wisconsin Department of Public Instruction (2003). *Bookmark standard setting overview.* Retrieved May 10, 2003, from http://www.dpi.state.wi.us/oea/profdesc.html

3. Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The Bookmark standard setting procedure.* McGraw-Hill, Monterey, CA.

Phil Matthews is Sub-Dean, Deputy Director of GP Education and Head of GP Training, Postgraduate Medical and Dental Education Wales.

Interested in learning more about this and other educational topics? Why not professionalise your role with an academic qualification at PGCert, Dip or MSc in Medical Education via e-learning or attendance courses.

Contact: **medicaleducation@cardiff.ac.uk**

Series Editor: Dr Lesley Pugsley, Medical Education, School of Postgraduate Medical and Dental Education, Cardiff University.

Wales Deanery
Cardiff University, 9th Floor, Neuadd Meirionydd,
Heath Park, Cardiff CF14 4YS
Tel: +44 (0)29 2068 7451 Fax: +44 (0)29 2068 7455
E-mail: medicaleducation@cardiff.ac.uk

Wales Deanery Deoniaeth Cymru

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD

ISBN: 978-1-907019-36-4