Computerized standardsetting using the The Data-Driven Direct Consensus (3DC)

Efficient and innovative standardsetting

Jesse Koops, Remco Feskens & Frans Kleintjes,

Cito, national institute for educational measurement, The Netherlands

Keywords: Angoff, bookmark, direct consensus, empirical data, standard setting

Introduction

In educational measurement we distinguish between content standards and performance standards. Content standards are typically devised by policy makers. They specify what knowledge and skills can and should be expected of students at certain points in their education. This is often specified in a curriculum.

Performance standards define the relation between the content standards and the performance on a specific test. The performance standard can thus be seen as an operationalization of the content standard at a specific test.

The main question during standard setting is therefore often: How would a student who is exactly on the border of mastery of the content standards, perform on this test? Various procedures for establishing performance standards have been proposed in the literature. Among the best-known examples are the Angoff procedure, the Bookmark procedure and the Direct Consensus procedure. These procedures have their strengths and weaknesses. Some procedures make it possible to establish performance standards relatively efficiently and quickly, but lack empirical rigor. Other procedures do include empirical data, but are time consuming and not very intuitive.

Data-Driven Direct Consensus (3DC) procedure

The strengths of the aforementioned standard setting procedures were brought together in a new one: the Data-Driven Direct Consensus (3DC) procedure. The 3DC procedure divides the full test into a number of clusters and uses empirical data to relate the
scores of the clusters to the scores of the full test. The relationships between the clusters and the full test are presented to the panellists on a specially designed assessment form. Panellists are asked to use the assessment form to indicate the score that students would be expected to achieve in each cluster if they were exactly on the borderline of the selected mastery level. Because of the design of the assessment form, the assessment is easily allowed to be based on both content information and empirical data.

In setting a performance standard, it is not possible to completely avoid the subjectivity of human judgement. The goal of a specific standard setting procedure like the 3DC is to ensure a defensible and valid standard.

The Data Driven Direct Consensus procedure (3DC) is a very flexible method for standard setting. As the name implies, it borrows considerably from the Direct Consensus procedure but it adds psychometrical rigour and guidance for the panellists by providing extra information about the relative difficulty of the clusters in relation to each other and to the test score.

The 3DC method allows for a variety of dichotomously or polytomously scored item types and any prediction method can be used to construct the form. A complete description of the 3DC method of standard setting can be found in Keuning, J. Straat, J.H. & R.C.W. Feskens (2017).

This contribution focusses on applying the 3DC method in practice.

*Workflow*
In a typical workflow, a psychometrician constructs an ability scale based on assessment data. Items are then grouped in clusters, preferably in a logical manner based on their content. The ability scale and the clusters are used to build a form using random draws from a general partial credit model (an R script to construct the form data is provided). The form data is saved and later used by the 3DC application during a standard setting session.

For this session, a small network of laptops is set up, often using the panelists own laptops. The only requirement of participants’ laptops is that a recent version of Google Chrome is installed. Each panelist can provide his or her indicators by clicking on the form. Black circles are used for round 1. After a central discussion panelists can - if so desired - correct their indicators in round 2 which will then be shown in red.

The application also allows participants to view and browse through the items in each cluster. Items in QTI or HTML format are natively supported - including multimedia - and there are workarounds to view items in other formats.

Below the progress in standard setting a specific item is shown. In the left hand lower corner the group of items used in the standard setting is presented by their identification number. The item with the red label is displayed for evaluation, more centrally in the box below. All items can be viewed by the rater, who subsequently can indicate the boundary score on the corresponding line in the top of the box.
Impact evaluation

During a session the group leader can review all selected performance indicators of all the panelists. Basic statistics and useful graphs are also provided in real time. For example, the below graph shows the impact of the mean performance standards in the target population.
The application, the R script to generate the form data and a manual can be downloaded from the Cito website and can be used free of charge.

Reference:

Manual and download.
The software package, full manual and the R-script can be downloaded from the Cito website. www.cito.com/our-expertise/implementation/3dc