Universität Zürich[UZH]
Institut für Bildungsevaluation
Assoziiertes Institut der Universität Zürich

Multi Stage Testing in Secondary Education in Northwestern Switzerland
September 2015

A Multistage Test (MST), called Check S2, was designed to determine the abilities of secondary education students in Northwestern Switzerland, with respect to Mathematics, German, French, English, and Science. The administration of the Check S2 allowed for the development of a large item pool that can be used for future test administration. Therefore, the present paper was aimed to provide an overview of the choices made with respect to assessment design and test construction.

Multi Stage Tests

With the recent advances in technology, Computer Based Tests (CBTs) have gained popularity. A distinction can be made between linear CBTs and adaptive CBTs. In a linear CBT, all of the students are given the same set of items and, in this sense, a linear CBT is similar to a linear paper based test, except for the administration mode (computer vs. paper). In an adaptive CBT, commonly called a Computerized Adaptive Test (CAT), the difficulty level of the items can be adapted to a student's ability during the administration of the test. Therefore, CATs have the potential for determining the student's ability more accurately, or reaching the same accuracy with a shorter test length. Since the difficulty level of the items is adapted to the student's ability level, a student's frustration with test items that are either too difficult or too easy is reduced.

The Multi Stage Test (MST), a specific type of CAT, has been gaining popularity in educational testing, as it combines the favorable properties of both linear and fully adaptive tests. An MST allows for an adaptation in the difficulty of a subset of items, commonly called a "module". Compared to a fully adaptive test, in which every single item is tailored to a student's ability, in an MST, whole modules varying in difficulty are tailored to the student's ability (see e.g. Yan, Lewis, & von Davier, 2014; van Groen, 2014). This gives test developers greater control over a test's content, and the quality of the test structure, when compared to a CAT. At the same time, MSTs maintain the favorable measurement properties of a CAT (Zenisky & Hambleton, 2014).

Check S2

A graphical representation of the MST design, as implemented in the Check S2 for Mathematics, is presented in Figure 1. This MST design is called a "double" 1-3-3-3 MST, which means that there is only one module in the first stage, but three modules in each subsequent stage. The term double indicates that every module in every stage consists of

two parallel versions (A and B) at the same difficulty level. The implementation of parallel versions allows the administration of twice as many items, compared to a "single" 1-3-3-3 MST. In addition, cheating (by copying from other students) during the administration of the test and the exposure of the items were reduced.

Modules belonging to the same stage differ in their difficulty levels. For example, module 7A/7B contains items which, on average, are more difficult when compared to module 6A/6B, with items that are more difficult than those in 5A/5B, and so on. This MST design was chosen as a large variation in the students' abilities was expected because of the three different school levels in secondary education.

Since no empirical information about the difficulty level of each item was available before the test administration, the difficulty of the items was determined by content experts. The items were categorized into four difficulty levels and distributed among different modules, according to their categorization.

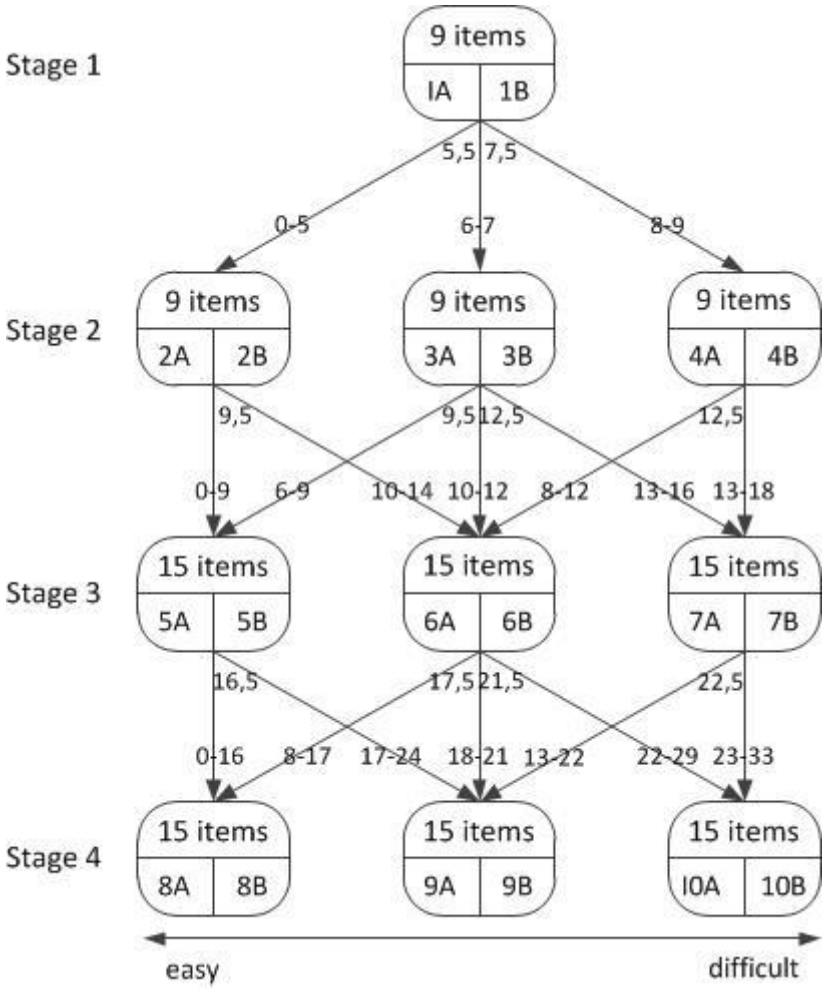

Figure 1: MST design for the Check S2 test administration for Mathematics.

Either module 1A or 1B was presented to each of the students at the beginning of the test. Depending on their raw scores after the first 9 items, they were given a module from stage 2. For example, a raw score between 0 and 5 resulted in the administration of module 2A or 2B, and a raw score of 8 or 9 resulted in the administration of module 4A

or 4B. The numbers directly below each module indicate the cut-off scores for routing to the next stage, and the numbers above the modules indicate the score range when entering the module. Based on the assumption that a significant number of students are assigned to a module appropriate for their ability after two stages, the modules in stage three and four contained more items than the modules in the first two stages.

It can be seen that not all of the routes between the two subsequent stages were made available. To illustrate, students in the easy module (2A/2B) were able to go to the easy or medium module in the next stage (5A/5B or 6A/6B), but not to the difficult module (7A/7B). Having four stages, as opposed to three, provided the opportunity for a transition from module 2A/2B to 6A/6B and to 10A/10B. Including this opportunity in an MST design is desirable, given that test anxiety (which negatively influences test scores) might be higher during the administration of the items in the first stage.

In the absence of empirical data, the routing rules were derived by assuming that the first module was comprised of about 2/3 of the items with low difficulty and 1/3 of those with medium difficulty. The routing rules for the higher modules were based on the score distribution from the maximum score of a module, and the assumption that the mean score of a module was about 2/3 of the maximum score.

Item Response Theory

One consequence of using an adaptive test design is that test scores of different students are based on different sets of items. It might seem illogical to compare a student's test scores when they were given different sets of items with different difficulties. This is why the primary focus in modeling the data from an adaptive assessment design was on the ability scores instead of the test scores. Compared to the test scores, the ability scores are a more fundamental concept. Students come to a test with ability levels or scores in relation to the construct being measured (Hambleton & Jones, 1993). Over time, their abilities may change, but at the time of a specific assessment, their ability levels are independent from the test.

Within the statistical framework, called item response theory (IRT), models have been formulated to determine how the performance on test items relates to the ability scores. Within the most commonly used IRT models, this relationship is modeled by means of item parameters and ability parameters. Both the item parameters and ability parameters are invariant, which means that the item parameters are independent of the ability distribution of the examinees, and that the ability parameters are independent of the form of the test given to the examinees. This is why the IRT framework is suitable for estimating the ability scores in an adaptive assessment design, thereby making meaningful comparisons between the students' test performances, with regard to being administered different sets of items.

Evaluation

After administration, the measurement quality of all of the tests was evaluated. For example, the Mathematics item pool consisted of 251 items, of which each student was presented 48, following the design shown in Figure 1. Items with undesirable

measurement characteristics were excluded from the test reports, including items that were too difficult (proportion correct below 10%) and items that showed misfits with regard to the IRT model. The remaining 237 items in the pool had an average proportion correct of 48%. It has been shown in the literature that a proportion of 50% results in optimal measurement precision; thus, 48% was a highly desirable result.

The reliability of MSTs (and fully adaptive tests) cannot be calculated directly, but must be investigated using a simulation study. The item scores for the Check S2 were simulated using the observed ability distribution, and the tests were simulated according to the specified design. Because of the exclusion of certain items, the tests consisted of, on average, 45 calibrated items, with a reliability of 0.90 (Macc, see Verstralen, 1997). In order to compare the performance of the MST to a linear test, linear tests consisting of 45 items were simulated, which were selected at random from the item pool. The reliability of the linear tests was 0.87. Using the Spearman-Brown formula, it is possible to calculate the number of items that are required to obtain a similar reliability, assuming that the items that were administered are representative of the remaining items in the item pool. Twelve additional items are required for a linear test to have a reliability equal to the MST of the Check S2 for Mathematics. Thus, by adapting the test to the student's ability, a higher reliability was achieved when compared to using a linear test.

References

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on educational measurement. *Issues and practice, 12*, 38-47. doi: 10.1111/j.1745-3992.1993.tb00543.x

van Groen, M. M. (2014). *Adaptive testing for making unidimensional and multidimensional classification decisions*. Doctoral thesis, University of Twente, Twente, The Netherlands.

Verstralen, H. H. F. M. (1997). *OPTAL: Inverse OPLAT and item and test characteristics in populations*. Arnhem: Cito Instituut voor Toetsontwikkeling.

Yan, D., Lewis, L., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier & L. Lewis (Eds.). *Computerized Multistage Testing Theory and Applications*. Taylor and Francis Group, LLC.

Zenisky, A. L., & Hambleton, R. K. (2014) Multistage tests designs: Moving research results into practice. In D. Yan, A.A. von Davier & L. Lewis (Eds.). *Computerized Multistage Testing Theory and Applications*. Taylor and Francis Group, LLC.