Chapter 11

Computerized Classification Testing and Its Relationship to the Testing Goal

Maaïke M. van Groen

Abstract Assessment can serve different goals. If the aim of testing is to classify respondents into one of multiple levels instead of obtaining a precise estimate of the respondent's ability, computerized classification testing can be used. This type of testing requires algorithms for item selection and making the classification decision. The result of the test administration is provided in a report about the decision with sometimes additional feedback. The design of all these components of the test should be in line with the testing goal. Several goals have been defined for assessment which make a judgment about: pupils, the learning process, groups of students and schools, and the quality of education. The possibilities for use of computerized classification testing for different testing goals are investigated in the current paper.

Keywords: computerized classification testing, testing goals, test design

Introduction

Assessment can have different goals. In some testing situations, the aim is to classify respondents into one of multiple levels instead of making a precise estimate of the respondent's ability. This should be achieved by administering as few items as possible while maximizing the number of correct classifications. Computerized classification testing (CCT) is an approach that can be used for finding a balance between the number of items and the level of confidence in the correctness of the decision (Bartroff, Finkelman, & Lai, 2008). According to Thompson (2009), computerized classification tests assign an examinee into one of two or more mutually exclusive categories along the ability scale. In the current paper, this definition is further limited to tests based on item pools that have been scaled using modern psychometric methods.

One part of the procedure determines which items have to be selected. Another part of the CCT procedure determines whether testing can be stopped because enough confidence has been gained in making the decision or that an additional item has to be administered.

The classification method as well as the item selection method have to be in line with the testing goal and the report and feedback that have to be provided after testing has been

finished. This is illustrated in Figure 1. Based on the testing goal, a method for reporting the classification and the type of feedback can be determined. The testing goal also partly determines which classification method can be used and how it should be implemented. The goal influences the selection of the item as well. The way in which results can be reported and feedback can be provided is determined by the classification method as well as the item selection method. Ideally, these methods should be designed so the desired report and feedback can be provided afterwards. A test developer should always keep in mind that the goals, methods, and report should be synchronized with each other.

Figure 1 Testing goals and CCT components

Testing Goals and Computerized Classification Testing

In the previous section, attention was paid to testing goals. However, which goals testing can have was not described. The implications for classification and item selection methods and report and feedback have been mentioned only briefly. In this section, first, testing goals and computerized classification testing are described. The possibility of using computerized classification testing for specific testing goals is then described. In the last section, the implications of testing goals for designing computerized classification tests are investigated.

Testing can serve different goals. One taxonomy of testing goals is provided by Sanders (2011). He divides testing goals into:

- Assessment for making a judgment about students
- Assessment for making a judgment about the learning process
- Assessment for making a judgment about groups of students and schools
- Assessment for making a judgment about the quality of education.

A second distinction can be made regarding the importance of the consequences of testing because the importance of the test has a major influence on test design. Stobart (2008) defines a high-stakes test as having substantial consequences for some or all of the parties involved. A third distinction can be made regarding the type of assessment: assessment *for* learning or assessment *of* learning. Assessment for learning is used as a tool for supporting the learning of pupils by providing guidance for the instructional process. Assessment of learning includes all tests that measure knowledge after a period of instruction to assess whether the required knowledge level has been reached or not. Since the testing goal, the importance of testing, and the type of assessment are closely related to each other, they are described together.

Assessment for Making a Judgment about Pupils

Assessment for making a judgment about students can be subdivided into four subgoals:

- Selection
- Classification
- Placement
- Certification

Sanders (2011) explains that selection takes place if not all the students who want to enroll in a program or study can be admitted. Based on the selection decision, a fixed number of students are admitted to an educational program. A selection decision can be made based on a specially designed test, for example, the Law School Admission Test for admission to law school in the United States or based on a test with a more general goal. An example of the latter are the final examinations for secondary education in the Netherlands used for selecting students for admittance to medical school.

Based on the classification decision in a classification test, a different educational program is offered to the student that will lead to a different diploma (Sanders, 2011). The final test for Dutch primary education (Cito, 2012) is one of the instruments that can be used for deciding the level of secondary education a child will attend in addition to the teacher's advice and the parents' ideas. If placement is the testing goal, the student will be placed in a different educational program, but the final diploma will be the same. An example is an entrance driving test that is used for selecting students for a short driving course who will be able to

pass the driving examination after only a limited number of driving lessons. Those who are expected to need more lessons will be selected for a longer driving course. The last testing goal is certification. Certification tests are used in situations in which a final judgment has to be made regarding the student's level in order to receive a certificate or a diploma. Well-known examples of such tests are the final examinations in secondary education in the Netherlands. These testing goals have in common that they are a form of assessment *of* learning and that they can all be seen as a form of high-stakes testing. The goal is to make a summative judgment of the student's knowledge.

Assessment for Making a Judgment about the Learning Process

In assessments in which a judgment is made regarding the student's learning process, the goal is to obtain information that can be used in the instructional process (Sanders, 2011). This can be seen as assessment *for* learning. Using such a test, the teacher will be able to adapt his or her instruction to increase the students' knowledge and skills. Diagnostic tests also serve this goal of testing. If diagnostic testing is the goal of testing, the interested reader is referred to Rupp, Templin, and Henson (2010). Also tests like the Mathgarden (www.mathsgarden.com), a serious game for primary education arithmetic, and simulation-based learning in aviation can be seen as making judgments about the learning process of the student. Sanders (2011) points out that the distinctions between testing and instruction will become blurred in these tests. Tests that serve this goal have a major impact on the teaching methodology but usually have only indirect impact on the pupils themselves.

Assessment for Making a Judgment about Groups of Students and Schools

Assessments for making a judgment about groups of students and schools take place if the test results of individual students are aggregated to get information about the group or the school (Sanders, 2011). Assessment for making a judgment about groups of students and school can serve different purposes. If the focus is on improvement of the learning of the group of students, it can be seen as an assessment *for* learning. If the focus is on accountability for the results of the group or the school, it can be seen as assessment *of* learning. An example of the former is the situation in which small groups within the class are arranged based on their achievements on a test in order to provide different instruction to each group. An example of the latter is the use of test scores for giving the Inspectorate insight into the quality of the

school. The consequences in this situation are highest for the school instead of for the individual pupil (Stobart, 2008).

Assessment for Making a Judgment about the Quality of Education

The last goal of testing is assessment for making a judgment about the quality of education. In these studies, the goal is to measure the quality of the education in a nation or to compare educational systems in different nations (Sanders, 2011). Such studies, such as PPON and PISA, provide policymakers, the Inspectorate, developers of instructional and assessment material, and so on, insight into the current level of pupils in the nation. Based on the findings, adjustments in policy and materials can be made. Since test results are not used for improvement of education on the individual level, these tests can be seen as assessment of learning. The stakes in these tests are primarily on the national level.

Computerized Classification Testing for Different Testing Goals

Computerized classification testing can be used in many different situations. In this section the use of CCT is explored for the testing goals as defined by Sanders (2011). The efficiency and effectiveness of CCT is compared to linear testing and computerized adaptive testing for those goals. In computerized adaptive testing the goal is to obtain a precise estimate of the respondent's ability level on a continuous scale instead of making a classification decision into one of multiple mutually exclusive categories. But first, some additional information is provided about computerized classification testing.

Computerized Classification Testing

CCT requires two algorithms. The first determines when a classification decision can be made. The second determines which item has to be administered next. Several methods exist for making a classification, such as the sequential probability ratio test (Wald, 1947/1973; Reckase, 1983; Eggen, 1999) and the ability confidence interval method (Weiss & Kingsbury, 1984). The majority of these methods can classify respondents into two levels, but some can also classify respondents into multiple groups. Commonly used item selection methods such as maximization of information at the cutting point and maximization at the current ability estimate can be used if classification into two groups is desired (Eggen, 1999), but some methods can also be used if a classification into multiple levels is required (Eggen & Straetmans, 2000; Van Groen, Eggen, & Veldkamp, 2012).

Assessment for Making a Judgment about Pupils

Computerized classification testing was originally designed for dividing respondents into different groups. If the assessment is used for making a judgment about pupils, computerized classification testing is one of the most efficient methods. Because decisions about pupils have a major impact on students, a high level of accuracy is desired. In CCT, accuracy is maximized while test length is minimized. Depending on the precise testing goal, more or less accuracy is required. If the goal is classification or certification, accuracy is extremely important because of the stakes for the student. CCT cannot be used if selection of students is the goal of the assessment because CCT requires a fixed cutting point instead of a flexible cutting point. When selection takes place, the cutting point is set at the value that results in the specified number of students who pass.

Linear testing can be used for making classification decisions, but many more items are required to make the classification decision as accurate as in CCT. Computerized adaptive testing (CAT) also requires more items than necessary because in CAT precise estimates have to be acquired at all points on the ability scale. In CCT, however, precision is required only on one or more points on the ability scale if a classification decision has to be made. CAT is well suited for making selection decisions because the cutting point can be set at every point on the scale after all tests have been administered.

Assessment for Making a Judgment about the Learning Process

If assessment for making a judgment about the learning process is the goal, computerized classification testing can be used if a precise ability estimate is not required. If a classification decision on subdomains, such as multiplication, division, and so on, is sufficient, CCT can be used; otherwise, CAT or linear testing has to be used. If CAT or linear testing is used, more items will be required for obtaining information about the student's level. Different models can be used in CAT and linear testing that have been designed for diagnostic testing especially (Rupp, Templin, & Henson, 2010).

In assessment for making a judgment about the learning process, the idea is to gather information within a rather short time and use the test results to adapt the instruction to the students. This implies that only a limited number of items will be available for making the classification decision and accuracy is not the most important goal. If diagnostic information has to be gathered on several subdomains, a limited number of items will be available per subdomain, and per subdomain a classification decision has to be made.

Assessment for Making a Judgment about Groups of Students and Schools

If a judgment has to be made about groups of students or schools in the context of assessment for learning, the same conditions apply as for making judgments about the learning process for individual students. In both situations, the ultimate goal is to adapt the instruction the teacher provides to the students' knowledge level. The difference is in the focus on the judgment groups and schools instead of individual students.

If a judgment has to be made about groups of students or schools in the context of assessment of learning, computerized classification testing can be used if a classification decision suffices. If different cutting points have been set for schools due to different student characteristics, this is also possible. If more information is required than CCT can provide, CAT or linear testing has to be used.

Assessment for Making a Judgment about the Quality of Education

If assessment for making a judgment about the quality of education is the goal of testing, whether CCT can be used depends on the specific results policymakers want to be measured. If the goal is to investigate whether the required subjects are mastered by pupils, CCT can be used.

In situations in which the effect of a reform has to be investigated, the policymakers are interested in differences in ability before and after the reform. CAT and linear testing are better suited for evaluation of reforms. In the first situation, the stakes are at the national level instead of at the student level.

Designing Computerized Classification Tests for Different Testing Goals

The relationship between testing goals and components of a computerized classification test were described in Figure 1. The classification method, item selection method, report, and feedback should all be designed so that they are in line with the testing goal. In this section, the four design components are investigated for the different testing goals.

Assessment for Making a Judgment about Pupils

In a computerized classification test for making a judgment about pupils, traditional CCT classification methods can be used for making the decision whether to classify into a certain level or to continue testing. An algorithm can be selected based on the number of cutting points needed for the test. The focus of the item selection method should be on obtaining the most information as quickly as possible to be able to stop testing after as few items as

possible. If one cutting point is used, information can be maximized at the cutting point (Eggen, 1999). If multiple cutting points are used, an algorithm that takes this into account has to be used (Eggen & Straetmans, 2000; Van Groen, Eggen, & Velkamp, 2012). Using simulations, optimal settings for the classification method and item selection method can be determined. The report in a CCT for making a judgment about pupils can be simple. Reporting the actual decision often suffices. Specific feedback is not needed in these situations because the decision is all that matters.

Assessment for Making a Judgment about the Learning Process

If a computerized classification test is used for making a judgment about the learning process, the classification method has to include one decision per subdomain. This implies that per subdomain a classification has to be made about mastering the subdomain or not. It is also possible to include multiple levels in the classification method per subdomain. The number of items that have to be administered before stopping the test is strongly related to the number of subdomains and the number of cutting points per subdomain.

The design of the classification method should be in line with the specific theories behind the topic and should conform to the level of specificity a teacher needs to adapt the instruction to the student's level.

The item selection method should select items for the subdomain for which a classification decision has to be made. This implies that some kind of content control is required within the item selection method. To make decisions with as much information as possible, items should be selected that maximize information at the cutting point that is of interest. If items provide information regarding several subdomains, developing a special item selection method for the test can be more efficient. Simulation studies provide insight into the efficiency and side effects of different item selection methods.

The report should provide the information a teacher needs to adapt the instruction. Per subdomain the classification decision has to be provided. If available, specific feedback for improving the instruction can be given such as references to relevant exercises and instruction material. In a second screen, information could be provided per domain regarding the items that the student answered correctly or incorrectly. Additional feedback can be provided about the types of mistakes a student makes when answering the item. The report and the feedback should be well structured and easy to comprehend; if not, the teacher will look only at the classifications.

Assessment for Making a Judgment about Groups of Students and Schools

If a computerized classification test is used for judging groups of students and schools in the context of assessment for learning, the basic guidelines for CCT for judging the learning process can be followed. Differences should appear in the way the report is presented after the test is administered. The focus should be on groups of students or on the school. Aggregated results can be presented per subdomain with feedback on how instruction could be improved. Instead of providing information about individual students, the number of students that have gathered not enough knowledge about a subdomain could be presented. In addition, clusters of students with similar profiles based on the classifications can be provided. The teacher can provide instruction to the groups of students based on the profiles. If a computerized classification test is used for judging groups of students and schools in the context of assessment of learning, the classification method should be directed toward making a classification after as few items as possible.

The item selection method should also be directed toward gathering evidence for making the classification decision as quickly as possible. The report can be very basic. The percentage or number of students who pass the test should be reported. Feedback is not needed in this situation because the goal is to provide information for accountability purposes only.

Assessment for Making a Judgment about the Quality of Education

If a CCT is used for judgment the quality of education, the design of the components of the CCT can be comparable to the design for accountability. The difference between the two goals is primarily visible in the report. Instead of aggregation to the group or school level, aggregation should be at the national level. Specific feedback is not necessary.

Discussion

Computerized classification testing can be used in many testing situations in which students have to be classified into groups who have gathered knowledge at a certain level. By including subdomains in the classification method, it becomes possible to use CCT in more situations than often realized. The main reasons for not using CCT include requirements for giving scores at a continuous scale and possible objections against computerized testing.

Test developers should always keep the goal of their test in mind when designing the test. This is not different from traditional paper-and-pencil tests or computerized adaptive tests, but only limited theoretical work has been done for computerized classification testing. During the construction phase, the test developer should always keep Figure 1 in mind: the testing

goals define the requirements for the classification method, item selection method, report, and feedback. The classification and item selection methods restrict the information and feedback that can be reported, which implies that these four components have to be designed concurrently.

References

- Bartroff, J., Finkelman, M. D., & Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika*, 73, 473-486. doi: 10.1007/s11336-007-9053-9
- Cito. (2012). *Eindtoets Basisonderwijs [Final test primary education]*. Arnhem: Cito BV.
- Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, 23, 249-261. doi: 10.1177/01466219922031365
- Eggen, T. J. H. M. & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, 60, 713-734. doi: 10.1177/00131640021970862
- Groen, M. M. van, Eggen, T. J. H. M., & Veldkamp, B. P. (Unpublished). *Item Selection Methods Based on Multiple Objective Approaches for Classification of Respondents into Multiple Levels*.
- Reckase, M. D. (1983). A procedure for decision making using tailored testing. In Weiss, D. J. (Ed.), *New horizons in testing: latent trait theory and computerized adaptive testing* (pp. 237-254). New York, NY: Academic Press.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guilford Press.
- Sanders, P. (2011). Het doel van toetsen [The goal of testing] In Sanders, P. (Ed.), *Toetsen op school [Testing in schools]* (pp. 9-20). Arnhem: Stichting Cito Instituut voor Toetsontwikkeling.
- Stobart, G. (2008) *Testing Times: The uses and abuses of assessment*. London: Routledge.

- Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, 69, 778-793. doi: 10.1177/0013164408324460
- Wald, A. (1947/1973). *Sequential analysis*. New York, NY: Dover Publications, Inc.
- Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 4, 361-375. doi: 10.1111/j.1745-3984.1984.tb01040.x