Designing and Developing a Multistage Test in a High-Stakes Situation

Maaike M. van Groen[1]

Abstract

Computerized multistage tests adapt the test to the student's ability. All students take the same, or a very similar, first stage module. After this initial module, one of the modules available for the second stage is selected based on the student's previous responses. These follow-up modules differ in difficulty. Test developers need to make many decisions during the test development process about the design of a multistage test. Several design choices are discussed in the paper in the context of high-stakes testing, with choices that can be made about the configuration of multistage tests, reporting, and test assembly discussed in more detail. The discussion of these decision choices aims at providing practical guidance for test developers who want to design and develop a high-stakes multistage test.

Keywords: multistage testing, high-stakes testing, test development

Introduction

More and more testing organizations want to develop tests that are better tailored to the individual student or test taker. Computerized adaptive tests have been designed to tailor the test to the student's proficiency level. One type of computerized adaptive tests is the multistage test (Yan, Lewis, & von Davier, 2014). Multistage tests consist of several stages that include one or more modules. All test takers take the same, or a very similar, starting module. Afterward, the test taker's proficiency is computed and a subsequent module is selected for the second stage. The second stage contains multiple modules of different difficulty. The testing algorithm selects the module that appears to be the most suitable for the test taker. Modules of subsequent stages are selected in a similar way.

Test developers need to make many decisions during the test development process about the test design. Examples include decisions about the number of stages, the number of modules per stage, the selection of items, the type of test report, and so on. Depending on the purpose of the test, different decisions can be made about the test design. However, all choices that are made by test developers regarding the test design influence the precision of the test outcomes, the perceived value of the test, and the decision accuracy. This implies that decisions should be made with care and after careful investigation.

In high-stakes testing, the consequences of a wrong decision or a less precise estimate of the test taker's proficiency can have serious consequences for test takers (van Groen & Eggen, 2019). High-stakes tests have a summative purpose. Summative tests are used to make inferences about individual students (Haertel, 2013); make a decision about the test taker's mastery of a content domain (van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015);

[1] Maaike M. van Groen
Cito, Amsterdamseweg 13, 6814 CM Arnhem, the Netherlands, email: maaike.vangroen@cito.nl

determine entry into education; aid in college admissions decisions (Haertel, 2013); or determine whether a test taker graduates (van Groen & Eggen, 2019). Therefore, test reports need to be extremely precise and decisions need to be made with a high accuracy rate.

This paper focuses on the type of decisions that test developers make when designing and developing a multistage test for high-stakes testing. Insights are provided about the choices and decisions that need to be made during the test design and development process. The paper starts with an introduction to multistage testing. In the subsequent sections, several design choices are discussed in more detail. The paper ends with a discussion.

Multistage Testing

Multistage testing allows the difficulty of the test to be adapted to the proficiency level of the individual test takers (Yan, von Davier, & Lewis, 2014). Yan et al. (2014) also stated that multistage testing is at the confluence of linear and computerized adaptive testing and embeds features of both test designs. However, multistage tests have some unique features that are not shared with either test design. In the following subsections, the components that make up a multistage test are discussed, as are some of the choices that need to be made regarding these components. In the last subsection, the pros and cons of multistage testing as compared to linear and computerized adaptive testing are discussed.

Components of Multistage Tests

Multistage testing consists of and requires many components: modules, stages, paths, an item bank, a test blueprint, and methods for test assembly, routing, statistics, and reporting method. The main characteristic of multistage testing is that sets of items are assembled before the test taker begins with the first item of the first set. These sets of items are called modules. In contrast to testlets, these items do not necessarily share a common theme or stimulus; however, their difficulty falls within the same difficulty range. Modules are arranged into stages. Test takers take one module per stage, even though a stage contains multiple modules. A multistage test consists of two or more stages. The route that a test taker takes through the multistage test is called a path.

A high-quality item bank is required when assembling a multistage test for high-stakes testing. Multistage testing requires an item bank that covers a wider range of difficulty than a bank used in linear testing. The item bank needs to include high-quality items that are suitable for the easiest and the most difficult modules, and it should also contain a sufficient number of items that fully cover the multistage test blueprint. More information about item banks for multistage testing can be found in Zenisky and Hambleton (2014).

Tests are developed using a test blueprint. The test blueprint specifies how many items should be selected for the test, which domains should be covered, how many items should be selected per domain, how difficult the modules should be for the intended population, and so on. The module difficulty for stage two and higher should be computed for the subpopulation that is intended to be routed toward that module. This, in conjunction with the multistage test design, makes the construction of a test blueprint for multistage testing much more complicated than the design of a test blueprint for linear testing. The test blueprint determines which items can be selected by the test assembly method. This method forms a heuristic or algorithm for selecting the items.

After the test taker has taken the first module, a subsequent module needs to be selected for the test taker based on his or her performance so far. This routing decision is made by a routing method. Many routing methods have been described in the literature (Yan, Lewis, & von Davier, 2014). These methods can be very simple or elaborate and are often based on a statistical method.

Multistage tests for high-stakes testing require a statistical method to ensure the quality of the items, the test, and the test outcomes. One such method is item response theory (Hambleton, Swaminathan, & Rogers, 1991), which informs test developers about the statistical characteristics of the items and the test and provides an estimate of the test taker’s proficiency. A calibrated item bank results from the item response theory analysis.

One of the most important components of a multistage test, as in all testing formats, is the test report based on a reporting method. The test report is used by relevant stakeholders to make a decision about the test taker, such as a decision about domain mastery or a pass/fail classification. Depending on the test design and the statistical method, certain types of reporting are available.

Design Choices in Multistage Test Development

When developing a multistage test, test developers need to make choices about the design of all the components. The design that is the most appropriate for the testing situation at hand needs to be selected, but the design choices should also be in accordance with the purpose of the test. This implies that test developers need to make a different choice when designing formative tests than in summative testing. The focus of test developers when making design choices in summative testing should be on ensuring that the most accurate decisions or the most precise proficiency estimates are being provided to the student within the opposing practical and political restrictions.

Unfortunately, when making a choice about one component, other components are also affected due to the strong interrelationships between components. For example, a decision about the number of domains to be measured influences the number of items to be administered. A decision about the test report outcomes might influence the way subsequent modules are being selected or assembled.

During the test development process of a multistage test, many design choices need to be made. Some of those choices should have also been made when developing a linear test, some choices are affected by the choice of a multistage test design, and other choices are unique to multistage testing. Here, we limit ourselves to choices about the test design that are affected by or unique to multistage testing. Some of the design choices will be discussed in more detail in the next (sub)sections.

One of the choices that test developers need to make is about the number of items to be administered to each test taker. This choice has a major impact on the precision of the test outcomes. Longer tests can provide higher precision than shorter tests when the additional items are of similar statistical quality as the items that were already included in the shorter test. The desired test length depends on the configuration of the multistage test (Zenisky & Hambleton, 2014), but also on the number of domains included in the test report. Multistage tests allow for more efficient and precise measurement across the proficiency scale compared to linear testing (Yan, Lewis, & von Davier, 2014).

Test developers should decide how many stages are needed, how to distribute the number of items per stage, and the number of modules per stage. Furthermore, test developers have to decide whether test takers can take all paths through the test design, or if they can take only some paths. These choices will be discussed in the “Multistage Configuration” section.

When assembling a multistage test, test developers rely on test blueprints for selecting the items. The test blueprint specifies how many and which type of items need to be selected for each module, stage, and path, and also whether overlap in items is allowed between modules within the same stage. The design of the test blueprint itself requires a lot of choices to be made. The multistage test design complicates the design of the test blueprint due to the complicated configuration of multistage tests. More information about the assembly of multistage tests can be found in the “Multistage Test Assembly” section.

Several statistical methods exist for developing, assembling, and analyzing multistage tests, including number correct scoring, item response theory, and three-based methods (Yan, Lewis, & von Davier, 2014). Depending on the choice of a certain statistical method, choices can be made about reporting, test assembly, routing, and so on. The decision to use a certain statistical method for a specific purpose should be made carefully. When test developers use item response theory to make routing decisions, items need to be pretested and calibrated beforehand. Using number correct scoring does not require pretesting. Nevertheless, if the testing situation allows pretesting, pretesting can provide valuable information about the quality of the items and can aid test developers in designing a high-quality test for high-stakes testing. Interested test developers can obtain more information about statistical methods for multistage testing in Yan, von Davier, and Lewis (2014).

One of the factors that influences the precision and accuracy of test outcomes is the choice of a routing method. The routing method determines which module will be selected next for the candidate. Many routing methods have been developed (see Yan, Lewis, & von Davier, 2014). The choice of a routing method in high-stakes testing should be based on test situation specific simulation studies. Depending on the testing situation, one should select the method that will result in the most precise test outcomes given the test design. Often, routing methods based on complicated statistical methods will result in the most precise test outcomes. Test developers should also consider how to implement a certain routing method. Some routing methods might require additional choices to be made about the specific implementation of the method, such as whether all or just the previous module should be included in the ability estimation for the maximum information routing method. More details about additional choices for routing were presented at the 2017 IACAT conference by Straat and van Groen.

One of the most important choices that needs to be made by test developers when designing a multistage test is which test outcomes will be reported to the test takers. Given the importance of this choice, reporting will be discussed separately in the “Multistage Test Reporting” section.

All choices that are made by test developers influence the precision of the test outcomes, the perceived value of the test, and the decision accuracy. This implies, especially in high-stakes testing, that decisions about the test design should be made after careful consideration, deliberation, and investigation. The unique combination of high-stakes and multistage testing challenges test developers to make the most appropriate choices while considering practical and political constraints.

The Pros and Cons of Multistage Testing

Multistage testing can be seen as a hybrid between linear and computerized adaptive testing. The test design shares a sequential structure with linear testing, but at the same time, the content is tailored to the test taker’s proficiency as in computerized adaptive testing. This implies that multistage testing shares advantages and disadvantages with linear testing and computerized adaptive testing.

A major advantage over linear testing is that multistage testing adapts the test content to the test taker’s proficiency. However, multistage tests are by design less adaptive than computerized adaptive tests. The amount of adaptivity that multistage tests can provide strongly depends on the number of stages and modules with varying difficulty per stage.

Another major advantage of computerized adaptive and multistage testing is that fewer items are required to test with the same precision of linear testing. Also, both types of adaptive testing result in more precise test outcomes given a specified test length as compared to linear testing. When compared to computerized adaptive testing, multistage testing is only slightly less efficient and less precise (Yan, Lewis, & von Davier, 2014).

One of the major challenges with computerized adaptive testing is that item parameters need to be estimated with sufficient precision before testing can take place. Therefore, pretesting is a prerequisite for adaptive testing. Multistage and linear testing, on the other hand, do not necessarily require items to be pretested. If the testing situation allows pretesting, it can provide valuable insight into the quality of the items and the test before test administration. However, if pretesting is not possible, multistage and linear tests can be developed using expert judgments.

Multistage testing requires a larger item bank than linear testing due to the construction of multiple modules per stage. Or put otherwise, a linear test consists of one or more stages, each consisting of one module, and thus requiring fewer items. Whether multistage testing requires a smaller item bank than computerized adaptive testing depends on the testing situation. In theory, the same item bank could be used to develop a computerized adaptive or a multistage test. However, the item bank should be consistent with the test blueprint. Depending on the test blueprint, computerized adaptive testing might require more items because more items might need to be available within each difficulty range, whereas multistage testing needs items for a limited set of difficulty ranges.

One of the major problems associated with computerized adaptive testing is that items cannot be revised after the selection of a subsequent item. Although response revision is a topic of current research (Han, 2013; Wang, Fellouris, & Chang, 2019), it is not clear yet whether response revision will be feasible for high-stakes adaptive testing. Multistage testing has the advantage that response revision is possible within a stage. This allows test takers to carefully review their responses after providing preliminary responses to items.

Multistage tests are more complicated to design and develop than linear tests due to their design configuration. Also, developing test administration software is more complicated for multistage testing. Compared to computerized adaptive testing, it might be less complicated to develop and design a multistage test. Computerized testing relies heavily on complicated algorithms, whereas in multistage testing test developers can make many design choices themselves. They can also review the item sequence by hand, whereas this is not possible in computerized adaptive testing. Test administration software might be less complicated for multistage testing.

Some Design Choices for Multistage Testing

Test developers need to make many design choices when developing a multistage test. Three of those (sets of) design choices are discussed next. Choices regarding the multistage configuration deal with determining the number of stages, the number of modules per stage, the permissible paths through the multistage test, and so on. The choices that are made about the reporting of multistage tests influence the usability of the test as well as the precision of the test outcomes. A multistage test assembly method needs to be chosen to be able to select the most suitable items for each position within the multistage test.

Multistage Configuration

Many important design choices that need to be made when designing and developing multistage tests. The implementation of multistage testing is very flexible, but the exact configuration of a multistage test involves “a series of critical decisions with consequences for the relative efficiency of the test” (Zenisky, Hambleton, & Luecht, 2010). One important word of caution before discussing the choices regarding the multistage configuration is necessary. The optimal choice for a specific configuration is highly dependent on the purpose of the test. Although multistage testing in a high-stakes situation requires a focus on enhancing the precision of the test outcomes, different decisions need to be made depending on the purpose of the test. It is strongly advised to use simulations to investigate the optimal configuration for the testing situation at hand. Many practical guidelines and considerations are provided in Zenisky et al. (2010) and in the multistage testing book edited by Yan, von Davier, and Lewis (2014).

One important configuration choice deals with the number of stages in the test. An obvious advantage of including more stages in the test is that the test will be adapted more frequently when additional stages are included. Also, measurement precision will be higher due to the increased tailoring of each item's difficulty to the test taker's proficiency. On the other hand, including more stages might complicate test assembly, administration, and analysis. However, practical and statistical considerations are not the only important factors; policy considerations also need to be taken into account: stakeholders might consider a two-stage multistage test unfair in high-stakes testing because of the perception that test takers cannot recover from routing errors (Zenisky et al., 2010). In high-stakes testing, all decisions should be aimed at increasing the precision of the test outcomes. This might suggest the need to set the number of stages as high as possible given practical and political constraints.

Another important choice concerns the number of modules for each stage. So far, the assumption has been made that the first stage consists of one module. However, in high-stakes testing it might be desirable to construct several parallel modules in order to reduce cheating possibilities. Adding more modules to subsequent stages with a greater variety of item difficulties allows for more adaptivity (Yan, Lewis, & von Davier, 2014). The increased level of adaptivity will increase the precision of the test outcomes. Unfortunately, adding modules also implies adding additional high-quality items to the item bank. The optimal number of modules per stage also depends on the heterogeneity of the population: a broader population suggests more modules per stage than a narrow population. Depending on the testing purpose and corresponding test outcome measures (see the next section), high-stakes multistage tests should include a sufficient number of modules in order to obtain the required precision.

When determining test length and the number of modules per stage, it is important to consider the number of items per stage. Adding additional items to a stage will increase the precision of proficiency estimates after finishing the stage. This will decrease the chance that a routing error is made and will increase the precision of the test outcome for the last stage. In some testing situations, a short first module is considered desirable if there are large differences in proficiency within the population. A short first module will allow test takers with proficiency at the extremes of the population to be routed to a more suitable module as soon as possible. However, a shorter first module implies that more routing errors will be made. Simulations and discussions are needed to determine the number of items per module in order to make an optimal choice for the testing situation at hand.

Test developers also need to decide whether all possible paths through the multistage test will be available to the test takers. In reality, some paths through the test will (hardly) ever be taken due to the settings of the routing method, or will require aberrant test taker behavior. It does not often occur that a test taker will be routed from a very difficult to a very easy module and vice versa. Only a few or perhaps even zero test takers will take such an extreme path. Test developers need to pay attention to each path through the multistage test during test development and design. Eliminating test paths will save time during test development. Simulations can be used to determine whether all paths can be reached and how many test takers might take each path.

Multistage Test Reporting

The purpose of a test has a major impact on the outcomes that need to be reported by the test. The test report enables stakeholders and test takers to use the outcomes of administering the test (van Groen & Eggen, 2019). This implies that the choices that are being made about the reporting should be suitable for the purpose of the test and they should support the stakeholders in using the outcomes. High-stakes testing implies that test outcomes need to be precise, accurate, and preferably efficient as well. The test should be as short as possible while achieving the required precision.

Two main types of test outcomes can be reported for multistage testing. An item response theory ability estimate can be provided, or a classification decision can be made. Item response theory can provide an ability estimate that has the same interpretation regardless of the path that the student took through the multistage test. Item response theory does require a calibration of test results. A calibration needs to be done before test administration using pretest or previous test results, or reporting can be done after a sufficient number of tests have been administered. The ability estimate can also be used to make a classification decision about the test taker: does the test taker fail or pass the test or at which level does the test taker perform. Special classification methods exist that are not based on the ability estimate. A method like the sequential probability ratio test requires fewer items than a classification based on the ability estimate. When the multistage test design is tailored to obtain classification decisions, test length will be decreased and decision accuracy will be increased. More information about classification methods for multistage testing can be found in Lewis and Smith (2014).

Number correct scoring is not meaningful in multistage testing because a number correct score for a specific test taker is only comparable with scores from test takers that took the same path through the multistage test. Using equating procedures, it is possible to predict the number correct score on a reference test, but this requires complicated statistical procedures.

Multistage tests can also report on subsets of items, such as domain reports. The same methods as those for reporting on the entire test can be used. Test developers should take care that sufficient precision is reached while using fewer items for reporting on domains. An elaborate discussion of the possibilities for scoring can be found in the “Routing, Scoring, and Equation” section of Yan, von Davier, and Lewis (2014).

Multistage Test Assembly

One of the most important decisions a test developer needs to make when developing a multistage test is to decide which items to include in the test. The selected items determine whether the test taker’s proficiency can be measured efficiently and accurately. Item selection is even more complicated in the case of multistage testing than in linear testing. The items themselves need to be selected, but the developer also needs to determine in which stage and in which module the items should be administered. Two approaches exist for selecting items: manually, by test experts, or automatically, using automated test assembly algorithms.

Test developers can select items manually using a test blueprint. The test blueprint provides the specifications for the test. The test specifications define what a test is designed to measure and how the assessment will be accomplished (Parshall, Spray, Kalohn, & Davey, 2002). Specifications include the number of items per domain, the number of items per module, the intended difficulty of the test, the item characteristics of the modules, and so on. Some specifications can conflict each other. Imagine a test with ten items of which eight items should have a drag-and-drop format and six items should measure subtraction. If no drag-and-drop items actually measure subtraction, item selection becomes impossible. Selecting the most suitable items in linear testing is already complicated, but it is even more complicated in multistage testing. Test developers need to make sure that all paths through the multistage test meet the test specifications. Given that those paths share items, changing the selection of one item affects multiple paths. This makes it really complicated to select items for multistage paths, especially if the test has more than two stages. The test developer also needs to decide whether items with desirable measurement will be selected in the first or last stages (Zenisky et al., 2010). Specialized software or dashboards can provide some support when selecting items manually.

Automated test assembly algorithms select items from the item bank according to the test blueprint, using both content and statistical specifications (Parshall et al., 2002). Automated test assembly algorithms were developed for linear testing, and were adapted for computerized adaptive testing later on. Automated test assembly algorithms for multistage testing are still being developed (Han & Guo, 2014; Luecht, Brumfield, & Breithaupt, 2010; van der Linden & Diao, 2014; Verschoor, 2019; Verschoor & Eggen, 2014; Zheng, Wang, Culbertson, & Chang, 2014). Some multistage tests have been assembled using Verschoor's 2019 algorithm (van Groen, Straat, & Keizer-Mittelhaëuser, 2018). More information about automated test assembly in general can be found in the reference work of van der Linden (2005). At this time, automated test assembly with a careful manual review appears to be the most suitable approach in high-stakes multistage testing. A careful manual review ensures that suitable items are selected, and the automated selection process ensures that the test blueprint is followed.

Discussion

The focus of this paper was on providing insight into the choices that test developers need to make when developing a multistage test for high-stakes testing. The paper presented several of the many design choices that need to be made and discussed some of the design choices in more detail. Many of those choices in test design and test development are interrelated. A decision about the number of stages in the test influences the number of items per stage. The decision about the type of test reporting influences which items should be selected for which module and stage. The interrelated nature of the decision making makes it difficult to oversee the consequences of making one specific choice. Test developers should be aware of this complexity and should consider interrelatedness during the entire test development process.

One of the methods for investigating the effects of certain design choices is simulation research. Settings can be configured in simulation research in such a way that the effects of those choices can be investigated. A thorough simulation study can provide useful insights for designing multistage tests. To obtain those insights, test developers should design their simulation study to be as close to the actual testing situation as possible. Several tools are available for simulations of multistage testing. One of the most flexible tools is the mstR package (Magis, Yan, & von Davier, 2018) in the open source environment R (R Core Team, 2019). Magis, Yan, and von Davier (2017) have provided a thorough explanation of how to simulate multistage tests.

An aspect of multistage testing that was not discussed much here deals with the IT environment for administering the test. Although some multistage tests are administered on paper, for example, some of Cito's student monitoring tests (Cito, 2019) or Professor Chang's automatically generated multistage tests in China (Chang, 2018), most multistage tests are administered via computer. Administration on the computer requires that an IT system needs to be purchased or developed that is able to administer the multistage test. This might seem like a simple step, but in reality, IT systems can impose major restrictions on possible test designs. The limitations and possibilities of the IT system need to be taken into account when designing and developing the test. Parshall et al. (2002) have provided some guidance in selecting or developing software that is appropriate for the testing situation at hand.

One important aspect that is often neglected in the literature regarding multistage testing is that important stakeholders need to be involved when making decisions during the test development process. Involving important stakeholders in the decision process ensures that the test will be considered to have high validity. It is said that “validity is in the eye of the beholder” (Mattern, Kobrin, Patterson, Shaw, & Camara, 2009). This implies that developers need to make sure that the test is considered to be valid in the eyes of those that will use the test. By involving these stakeholders early on, developers can make sure that they will consider the test to be valid for its intended purposes.

When developing a multistage test for high-stakes testing, test developers need to be constantly aware of the fact that each choice that is being made can have a major influence on the decisions that are being made based on the test reports. This implies that a test developer has fewer degrees of freedom during test development than in a low-stakes situation. Therefore, the test developer needs to be constantly aware of the possible consequences of choices and needs to carefully investigate those consequences before making a final choice. When all decisions are made with care and after careful investigation, multistage testing can provide the means to make informed decisions about the test takers.

References

Chang, H.-H. (2018, September). *From adaptive testing to adaptive learning*. Paper presented at the Frontiers in Educational Measurement Oslo Conference, Oslo, Norway.

Cito. (2019). Leerlingvolgsysteem 4.0 [Student monitoring system 4.0]. Arnhem: Cito B.V.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, *11*, 1–18. doi:10.1080/15366367.2013.783752

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Han, K. T. (2013). Item pocket method to allow response review and change in computerized adaptive testing. *Applied Psychological Measurement*, *37*(4), 259–275.

Han, K. T., & Guo, F. (2014). Multistage testing by shaping modules on the fly. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 119–133). Boca Raton, FL: CRC Press.

Lewis, C., & Smith, R. (2014). Multistage testing for categorical decisions. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 189–203). Boca Raton, FL: CRC Press.

Luecht, R., Brumfield, T., & Breithaupt, K. (2010). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*, 189–202. doi:10.1207/s15324818ame1903_2

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. New York, NY: Springer International Publishing.

Magis, D., Yan, D., & von Davier, A. A. (2018). mstR: Procedures to generate patterns under multistage testing (R package version 1.2) [Computer software]. Retrieved from https://CRAN.R-project.org/package=mstR

Mattern, K. D., Kobrin, J. L., Patterson, B. P., Shaw, E. J., & Camara, W. J. (2009). Validity is in the eye of the beholder: Conveying SAT research findings to the public. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 213–240). Charlotte, NC: Information Age Publishing.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, assessment for learning, and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice 22*(5), 324-343. doi:10.1080/0969594X.2014.999024

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York, NY: Springer. doi:10.1007/0.387.29054.0

van der Linden, W. J., & Diao, Q. (2014). Using a universal shadow-test assembler with multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 101–118). Boca Raton, FL: CRC Press.

van Groen, M. M., & Eggen, T. J. H. M. (2019). *Educational test approaches: The suitability of computer-based test types for assessment and evaluation in formative and summative contexts*. Manuscript accepted for publication by Journal of applied testing technology.

van Groen, M. M., Straat, J. H., & Keizer-Mittelhaëuser, M. (2018, November). *Routing in the multistage End of Primary School Test*. Paper presented at the 19th Annual Conference of the Association for Educational Assessment – Europe, Arnhem, the Netherlands.

Verschoor, A. J. (2019, June). *An ATA model for multistage testing*. Paper presented at the International Association for Computerized Adaptive Testing Conference, Minneapolis, MN.

Verschoor, A. J., & Eggen, T. J. H. M. (2014). Optimizing the test assembly and routing for multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 135–150). Boca Raton, FL: CRC Press.

Wang, S., Fellouris, G., & Chang, H.-H. (2019). Statistical foundations for computerized adaptive testing with response revision. *Psychometrika, 84*, 375-394. doi:10.1007/s11336-019-09662-9

Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 3–20). Boca Raton, FL: CRC Press.

Yan, D., von Davier, A. A., & Lewis, C. (2014). Preface. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. xxi–xxiii). Boca Raton, FL: CRC Press.

Zenisky, A., & Hambleton, R. K. (2014). Multistage test designs: Moving research results into practice. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 21–37). Boca Raton, FL: CRC Press.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York, NY: Springer. doi:10.1007/978-0-387-85461-8_18

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H.-H. (2014). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 87–99). Boca Raton, FL: CRC Press.