frontiers in Education

ORIGINAL RESEARCH
published: 11 December 2018
doi: 10.3389/feduc.2018.00111

Multi-Segment Computerized Adaptive Testing for Educational Testing Purposes

Theo J. H. M. Eggen [1,2*]

[1] Cito, Arnhem, Netherlands, [2] University of Twente, Behavioral and Management Sciences, Enschede, Netherlands

Computerized adaptive testing (CAT) was initially developed as a psychometric tool for the efficient estimation of the ability of a student. Because of technological and psychometrical developments, CAT can now meet practical conditions and can also have other purposes than the mere estimation of ability. It can be applied in not only summative but also formative settings. In this paper, attention is given to the goals of testing in education and the different CAT algorithms serving these purposes. In particular, the approach of multi-segment adaptive testing is described. A multi-segment CAT consists of a number of segments, each with its own algorithm and branching rules. In this approach, practical constraints can be implemented in CATs. Furthermore, having a different testing purpose per segment for possibly different parts of the population is possible. The method will be illustrated with a CAT that has been developed as a part of an operational student monitoring system, which is a spelling test of Dutch words.

Keywords: computerized adaptive testing, multi-segment testing, goals of educational testing, spelling ability, item response theory

OPEN ACCESS

Edited by:
Okan Bulut,
University of Alberta, Canada

Reviewed by:
Ulf Kroehne,
German Institute for International Educational Research (LG), Germany
Shenghai Dai,
Washington State University, United States

****Correspondence:***
Theo J. H. M. Eggen
theo.eggen@cito.nl

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *19 April 2018*
Accepted: *27 November 2018*
Published: *11 December 2018*

Citation:
Eggen TJHM (2018) Multi-Segment Computerized Adaptive Testing for Educational Testing Purposes. Front. Educ. 3:111. doi: 10.3389/feduc.2018.00111

INTRODUCTION

In developing tests in education, the most important phase is specification. In this phase, the purpose and the practical conditions in which testing should take place are clarified. The specific intended uses of test results put special demands on the way tests are composed. The same is the case if computerized adaptive testing (CAT) is applied in education.

CAT was initially developed as an individualized testing system which focuses on the efficient estimation of the ability of a student (Lord, 1970; Wainer, 2000; Van der Linden and Glas, 2010). Using an item response theory (IRT)-calibrated item bank, a CAT algorithm ensures that each test taker receives an optimal test. The algorithm selects items from the item bank tailored to the ability of the test taker, as determined from the test taker's responses during the testing. In applications the efficient measurement on one subject or dimension in one testing session was the main focus.

Because of technological and psychometrical developments and successful applications (Reckase, 1989), CAT has evolved from a mere psychometric tool for the efficient estimation of student ability to a testing mode that can meet practical constraints and serve different testing purposes. Meeting practical constraints are traditionally successfully implemented as modifications of the item selection part of the algorithm. To serve different educational purposes numerous customized parts of CAT algorithms, models or procedures, have been developed.

In this paper, the multi-segment adaptive testing approach will be presented. This approach can be used to develop CATs for possibly serving a variety of educational testing goals and meeting practical restrictions in one testing session. It offers a general framework for existing or new ways to develop CATs with possibly related multiple parts. First, however, a summary of the basic CAT elements and some of its extensions will be given. Second, the main purposes of testing in education will be presented. Finally, the approach will be presented and illustrated extensively with an example CAT developed as a part of an operational pupil monitoring system. In simulation studies of this multi segment CAT the expected measurement accuracy and the relation with the testing goals will be discussed.

BACKGROUND

CATs presuppose the availability of a calibrated item bank. An item bank is a collection of items constructed to measure a well-defined construct or ability. It also contains various characteristics of each item. These characteristics may relate to content or administrative information, but the item parameters derived from the calibration, or the estimation and establishment of model fit with an IRT model, are most important. In the case of dichotomously scored items, a commonly used IRT model is the two-parameter logistic model (2PL) (Birnbaum, 1968). In this model, the probability of correctly answering item i, $X_i = 1$is related to the ability θ of a student:

$$p_i(\theta) = \mathrm{P}(X_i = 1|\,\theta) = \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))}, \qquad (1)$$

where b_i is the location or difficulty parameter, and a_i is the discrimination parameter of item i.

CATs are governed by a testing algorithm. This algorithm is a set of rules determining the way CATs are started, continued and terminated. In **Figure 1**, a schematic representation of a CAT algorithm is given.

Item selection, estimation and stopping are the main psychometric parts of CATs. In basic CAT algorithms, the likelihood function of θ is used to estimate the ability of a student (Cheng and Liou, 2000). Considering the scores on $k$items $x_i, i = 1, ..., k$, this function is given by

$$L(\theta; x_1, ..., x_k) = \prod_{i=1}^{k} p_i(\theta)^{x_i}(1 - p_i(\theta)^{1-x_i}). \qquad (2)$$

In estimating θ after every administered item, this likelihood function is maximized with respect to θ , giving a maximum likelihood estimate of a student's ability. Because of the bias of this estimate (Warm, 1989), the maximization of the weighted likelihood is preferred. This estimate after $k$administered items is given by

$$\hat{\theta}_k = \max_{\theta}(\sum_{i=1}^{k} I_i(\theta))^{1/2} L(\theta; x_1, ..., x_k). \qquad (3)$$

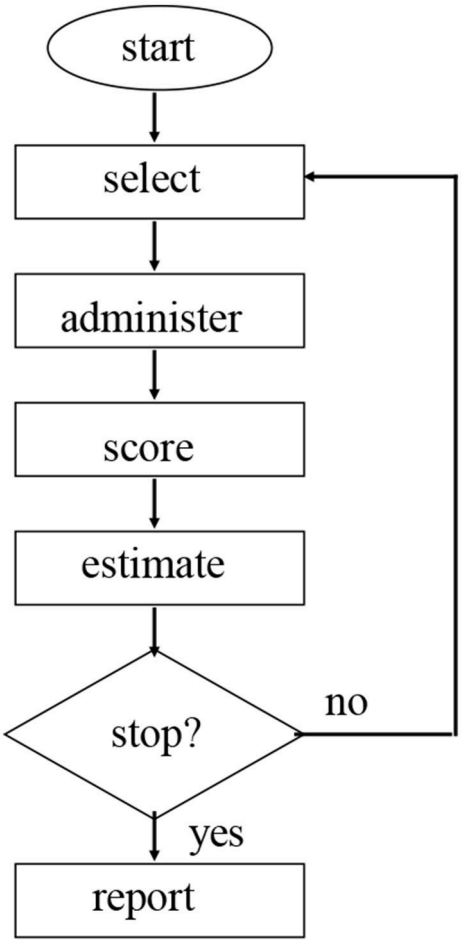

FIGURE 1 | Schematic representation of an adaptive test.

With the estimate $\hat{\theta}_k$, the standard error of the estimate$se(\hat{\theta}_k)$is determined also (see (4)), which is an indication of the accuracy by which the ability is estimated. In basic CATs, $se(\hat{\theta}_k)$is often used in a stopping criterion (Babcock and Weiss, 2012); if $se(\hat{\theta}_k)$is below a certain level, testing is stopped.

In (3), the likelihood $L(\theta; x_1, ..., x_k)$ is weighted by another function of the ability: $L_i(\theta)$.This item information function plays a major role in the selection of items. This function expresses the contribution that an item can give to the accuracy of the measurement of a student. This is readily seen because the standard error of the ability estimate can be written in terms of the sum of the item information of all the administered items.

$$se(\hat{\theta}_k) = 1/\sqrt{\sum_{i=1}^{k} I_i(\hat{\theta}_k)}. \qquad (4)$$

The larger is the item information, the smaller is the contribution to the standard error. In 2PL, the information

function is given by (5):

$$\mathrm{I}_i(\theta) = a_i^2 p_i(\theta)(1 - p_i(\theta)) = \frac{a_i^2 \exp(a_i(\theta - \beta_i))}{(1 + \exp(a_i(\theta - \beta_i)))^2}. \quad (5)$$

In the first developed CATs, item selection was based solely on the psychometric criterion of maximum information. The increasing number of CAT applications has resulted in greater consideration given to content-based and practical requirements or conditions (Reckase, 1989). Constraints are successfully implemented as modifications of the item selection part of the algorithm. For item selection, a wide range of methods is available to meet practical conditions [see, e.g., Eggen, 2008]. The most important imposed restrictions are content, exposure and difficulty control.

With content control, the desired content specification of the test, such as demands that subdomains of the measured ability are represented in a certain proportion, is met in each CAT. Ways of implementing this easily and effectively are given by Kingsbury and Zara (1991) and Van der Linden (2010a). In unconstrained CATs, although every student, in principle, receives a different test, a group of items is commonly found to be administered very frequently, whereas other items are hardly ever administered or never at all. Measures controlling over- and under-exposure overcome practical problems, such as security problems, with this (see Revuelta and Ponsoda (1998) and Barrada et al. (2009) for an evaluation of a number of exposure control methods). In difficulty control the goal is to have a certain desired of success probability for student on items. In non-constrained CATs, the items are chosen for an individual student at the current ability estimate where this probability is 50%. In practice, students perceive CAT tests as very difficult, and this could have negative effects, such as enhanced test anxiety. An algorithm that adapts the success probability on items without substantially losing measurement precision can be found in the work of Eggen and Verschoor (2006).

Common to these conditions is that they all have a small detrimental effect on the measurement accuracy of the CAT, as found in the references mentioned above. However, the size of the loss in accuracy is generally not considerable in comparison to the advantages of fulfilling practical requirements.

Even more important is the assertion that the choices for certain elements in CAT algorithms can be chosen in a way that specific goals of testing can be better served. Initially, CATs were developed for an individual efficient estimate of the ability of a student, and this aligns only with a limited number of testing purposes in education. Educational testing can have many goals, which will be described in the following section.

Goals of Educational Testing

The goals of testing or assessment at the individual level in education can be classified into two main areas: assessment of learning and assessment to support the learning of an individual (William and Black, 1996). The results of assessment of learning, also known as summative assessments, are used to allocate educational resources and opportunities among individuals. Typically, after a period of education or training, a decision on individuals are made based on test results. The major kinds of decisions are selection, classification, placement and certification decisions. Selection takes place, for instance, during admission to a university. In contrast to selection, with classification decisions, all students tested will proceed in education; they are assigned to different programs in schools. Placement decisions involve whether individuals participate in remedial programs or in programs for very talented students. In certification, the main emphasis is on establishing whether minimum competencies for a (part of a curriculum leading to a) certain school diploma or profession have been reached.

In formative assessments or assessments to support learning, three main approaches are presented in the educational literature (Van der Kleij et al., 2015). The first is data-based decision making (Schildkamp et al., 2013), in which the basic idea is that decisions made by teachers on the progress of the learning of students should not be made only intuitively but should be based on data. The results of testing are the most important data. The tests of student monitoring systems have this as their main goal. The major concern of testing here is to establish what is learned. This characteristic distinguishes it from the second approach, assessment for learning (Stobart, 2008), in which the focus is not on the outcomes of learning but on learning itself. The item-based learning environments of MathGarden (Klinkenberg et al., 2011) are an example of an application of this approach in an automated environment. If the main focus of assessment is on how students learn, we have the third approach, which is diagnostic testing. In diagnostic testing, detailed information is gathered about the learning process on the basis of a cognitive theory, and the purpose is to identify the steps in the development or misconceptions of learning among students. For more details on the theoretical underpinnings and the common and different elements of these three approaches, please see the work of Van der Kleij et al. (2015).

CAT has evolved to a testing mode that can serve different testing purposes, whilst meeting practical conditions. Traditionally, it has been mainly applied for summative assessment. Although efficient estimation of the ability is common in CATs for summative assessment, but, for instance, specific algorithms are used if the goal to classify students (Eggen and Straetmans, 2000). In formative assessment, a wide variety of algorithms attuned to the purpose of testing has been developed. Examples are that in the work of Cheng (2009) for diagnostic testing and that in the study of Wauters et al. (2010) for the assessment for learning. The current paper uses an example of data-based decision making in the multi-segment CAT approach, which will be introduced next.

MULTI-SEGMENT COMPUTERIZED ADAPTIVE TESTING

CATs are commonly considered an efficient measurement procedure of a construct in one testing session. In the practical development of CATs, however, meeting test specifications can result in considering a testing session as consisting of several connected segments. The multi-segment adaptive testing

framework was developed to serve a variety of goals and or to meet practical restrictions in one testing session.

A multi-segment CAT consists of a number segments, which can have a variety of relations with each other, and possible branching rules between segments used when the testing procedure is administered. Each segment is a clearly identified part of the testing procedure. The main reasons for building separate elements are as follows:

1. Fulfilling different goals of testing during one testing session becomes possible by distinguishing different segments. An example is that in one segment, a classification is made, and in another, an accurate estimate of the ability of a person is determined. Another example is that some segments only consisting of new items on which data are gathered only for calibration purposes (seeding) could be added.
2. Item content. In computerized testing, many different item types are available (see, e.g., Scalise and Gifford, 2006). Often, items have a specific stimulus or answer mode, and they need a special instruction to be able to answer them. For this reason, they cannot be intermixed during the same test administration session with other item types belonging to the same item bank.
3. Subtest content. In educational testing, a clear content structure in the item bank is often available. Grouping items with the same content in a segment can enable possible demands for administering these items subsequently in a testing session or giving subscores with enough accuracy to be met. If only representativeness for subdomains is needed content control within a segment will suffice, but often there is special interest in subdomains.
4. Segments can be built using the psychometric structure available in the item bank, that is, using the subsets of items fitted to separate one-dimensional, possibly differently modeled,scales. An example of this could be a reading test in which the first segment is on technical reading modeled with the Poisson counts model (Rasch, 1960), and the second segment is on reading comprehension for the items are modeled with the 2pl (equation (1)). Furthermore, it could represent the structure established by the fitting of a multi-dimensional IRT model or a sequence in an adaptive test battery [see, e.g., Brown and Weiss (1977) and Van der Linden (2010b)].
5. Segments can be built to meet practical constraints. An obvious example is the time available for testing: a separate segment for each available time slot. Another is that for meeting exposure control demands segments could be built on the basis of stratification of the discrimination of items as in the a-stratification method of Chang (1999).
6. Parts of the content of the item bank are sometimes not suitable for all groups in the population tested because, for instance, they have not yet been taught in school, or these parts are far too difficult for some students. With separate segments, these problems can be dealt with, and efficiency in testing can be achieved.

Each segment in a multi-segment CAT is seen as a CAT with its own algorithm (see **Figure 1**) serving the specific purpose of the (segment of the) test. The items in different segments on which the algorithm operates can come from a simultaneous calibration (e.g., in the case of different response types on items per segment) or from separate calibrations (e.g., in the case different psychometric models are used per segment). Each segment can have a quite complex algorithm with, for instance, many restrictions in the item selection, but it can also be a linear testing segment. The segments are connected in a flow in which branching is possible. The flow between segments can be simply based on the time needed or available for a segment, But more often also branching rules, based on results in an earlier segment determine the flow in the session. These rules can be based on the underlying psychometrical structure between the segments or on the specific goals of testing but they can also be motivated by the inadequacy or inefficiency of a segment for specific subgroups of the population for whom the complete CAT is designed. In general, the approach is very flexible and is implemented in the DOT software package (Verschoor, 2012).

In recent years, interest has increased in the development of adaptive tests, such as multi-stage testing (MST), in which the adaptivity is not on the item level but on the level of groups of items Zenisky et al. (2010). From the results of a routing test, students are administered an easier or a more difficult test in the next stage. MST can also be considered and developed as a special case of multi-segment testing. In the MST case, the segments themselves are usually linear tests (see, for specific information on MST, Yan et al., 2014). Furthermore, more sophisticated variants of MSTs, like (Zheng and Chang's, 2011) on the fly multistage test can also be considered as and fit in the framework of the multi-segment approach.

WORD SPELLING TEST IN DUTCH

The multi-segment method will be illustrated with CATs that are developed as a part of the operational student monitoring system called Cito[1]-LOVS. The monitoring and evaluation system of Cito consists of a coherent set of nationally standardized tests (paper and computer based) for the longitudinal assessment of a pupil's achievement throughout primary education. In this system a reading and spelling test are used for the detection of dyslexia (Keuning et al., 2011).

The development, test properties and performance of the spelling test of Dutch words, developed as multi-segment CAT will be discussed in some detail in the following.

The Item Bank

The item bank is based on descriptions of Dutch orthography by Bosman et al. (2006). From these basic principles of Dutch spelling, Keuning and Verhoeven (2007) derived the main relevant categories for the measurement of the development of Dutch spelling ability in primary schools. Distinguished are words that are phonetic, analogy based, rule based, or visual imprint words. Operationally, the spelling items (or target words) are presented orally in a short sentence. For instance, "We take the bus to school" … write down … "bus." Children were asked to spell the target word, such as "bus" in the example, by using

[1]The Institute for Educational Measurement in the Netherlands.

the computer keyboard. In addition, a repeat button is shown on the computer screen so that the children can listen to the item once more. The tests are administered individually. After a short introduction, pupils work on their own.

In the study by Keuning and Verhoeven (2008), the measurement structure of the item bank was first established. Their results show that children's spelling development can be conceptualized as a one-dimensional and continuous learning process across elementary grades 4 through 8 in the Netherlands. The results of the factor analyses showed that the spelling tests are highly dominated by a single factor. Furthermore, the results of IRT analyses showed that the initially constructed separate scales (i.e., the phonetic, visual-imprint, analogy-based, and rule-based scales) are strongly related and it is reasonable to conceptualize a single latent ability to underlie all of the spelling items. Nevertheless, the results also revealed the tendency of the children to master the four types of spelling items at different points in their development.

In the final calibration phase of the item bank, 450 items were administered to 10 subgroups in the primary school population, particularly in grade levels 4 to 8, during the middle (M) and at the end (E) of the school year, i.e., 4M, 4E, 5M, 5E, 6M, 6E, 7M, 7E, 8M, and 8E, in an incomplete design consisting of 30 modules with items. The total sample size of 4,977 students was equally distributed among the 10 subgroups (Keuning et al., 2011).

The total data set was calibrated in the one-dimensional 2PL model (See equation 1) using the OPLM computer program (Verhelst et al., 1995). In this software the incomplete dataset can be run in one concurrent estimation of the 2PL model with a limited number of discrimination values. For fitting the model graphical model checks (observed vs. expected item response curves) and statistics, with proven statistical properties are available (Glas and Verhelst, 1995). The misfit of 24 items was detected using M and S statistics. After several calibration runs 426 items gave a good item fit (graphical, and M and S statistics) and a reasonable fit with the global statistic, $R_{1c} = 3{,}150$, $df = 2{,}137$, $p < 0.01$, on one dimension (Glas and Verhelst, 1995). These 426 items are distributed across the content categories as phonetic words (68 items), rule-based words (125 items), analogy words (131 items), and visual-imprint words (102 items). Although the model fit on one dimension was established, the average difficulty of the items belonging to the content categories can be arranged from easy to more difficult.

The difficulty of the items shown a good spread on the scale. Phonetic words are generally less discriminating and are also the easiest items. Analogy and rule-based words are more difficult. Visual-imprint words are, on average, the most difficult. The calibration results from the use of marginal maximum likelihood with eight normal distributed subpopulations (Verhelst et al., 1995) are summarized in **Tables 1**, **2**. It is seen that the populations show on average a little declining growth in ability each half year. It can be concluded that the item bank is suitable for the measurement of the development of spelling ability in these populations.

Figure 2 shows the distribution of the difficulty parameter and the mean difficulty for each content category.

TABLE 1 | Item bank spelling item parameters per subcategory.

Spelling category	Mean a	SD a	Mean b	SD b
Phonetic (ph)	2.22	0.66	−1.14	0.70
Analogy based (an)	3.65	1.25	0.04	0.68
Rule based (ru)	3.10	1.48	0.20	0.79
Visual imprint (vi)	3.72	1.59	0.48	0.69
Total	3.28	1.44	0.00	0.90

TABLE 2 | Item bank spelling mean and SD in normal distributed subpopulations.

Subpop	Sample	Mean	SD
4M	499	−0.70	0.35
4E	474	−0.50	0.32
5M	460	−0.22	0.34
5E	459	−0.05	0.38
6M	511	0.10	0.38
6E	503	0.23	0.38
7M	544	0.33	0.41
7E	538	0.49	0.39
8M	505	0.62	0.36
8E	484	0.69	0.37

The Spelling CAT

The spelling CAT is developed as a multi-segment test. The content categories give the main rationale for the different segments of this test. The common instruction for students to answer the items within a category for each content category is one reason. But the differences between the (average) difficulty of the categories was an even more important reason for putting these items together in a segment. Because this makes it possible to test students efficiently without bothering them with items from content categories that are not important for them yet in their phase of development of the spelling ability. Being a part of a monitoring system intended to support learning this is very acceptable and favorable. The association between the difficulty and the content of the items was considered in the branching rules of the segmented test. The branching rules in this test are based on a criterion related to the mastery of a certain content subdomain. The stopping rules in each segment are based on considerations on measurement accuracy of the individual spelling ability and on available testing time.

In **Figure 3**, the flow of the multi-segment spelling CAT is given. The test starts with phonetic words (the category with the on average easiest items). If, after 15 items, the estimated ability is below the 50% mastery point of all phonetic words in the item bank ($\theta_{c\_ph} = -1.138$) testing is continued with 15 other phonetic words in order to get an accurate estimate of the spelling ability. For these students testing stops after 30 items. If the ability estimate after the first 15 items is higher than $\theta_{c\_ph} = -1.138$, 20 items from all analogy or rule-based items are administered. From both categories 10 items are administered. Then, the current ability estimate, based on these 35 items, is

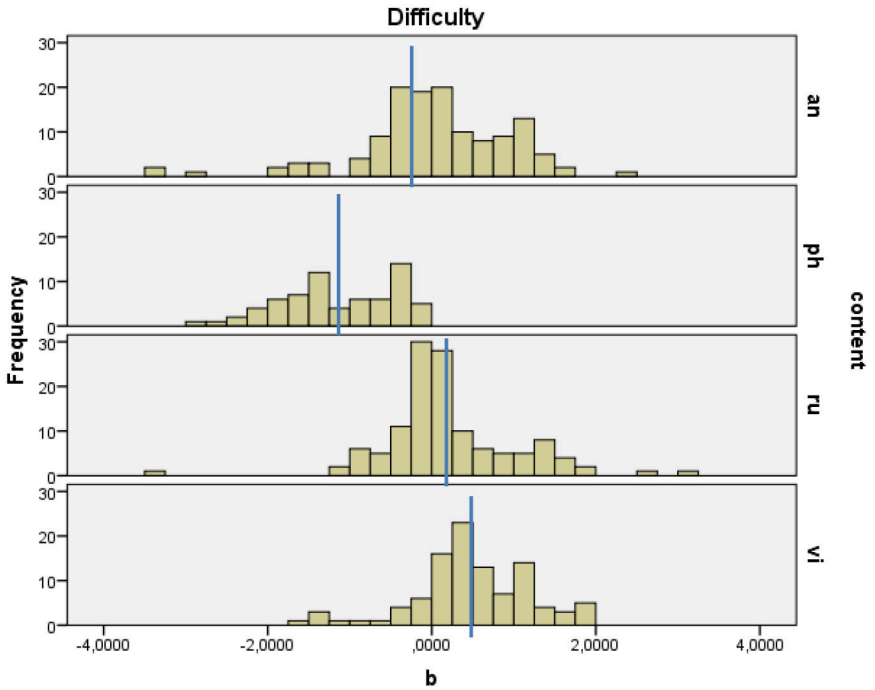

FIGURE 2 | Distribution of difficulty parameters per subcategory, analogy-based (an), phonetic (ph), rule-based (ru), and visual-imprint (vi) words.

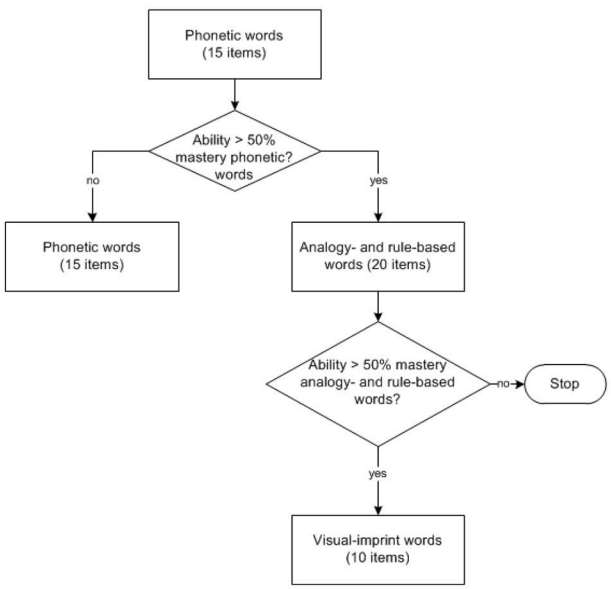

FIGURE 3 | Flow of the CAT spelling test.

tested against the 50% mastery point of all analogy and rule-based words in the bank ($\theta_{c\_arb} = 0.127$). If the ability estimate is below this cutting point testing is stopped. If it is above this threshold, testing is completed with the administration of 10 items of the most difficult category of visual-imprint items.

Within each segment, items are administered according to a CAT algorithm, in which items are selected on the demonstrated performance during testing. The algorithm has the following elements. In the first segment, the start is a randomly selected phonetic word. Then, in all segments, item selection is based on the maximum information at the current ability estimate, which is, in turn, based on the responses in all items answered thus far. Because of the low stakes character of the testing in the monitoring system, only a light form of exposure control (Sympson and Hetter, 1985) with maximum exposure of 0.7 was implemented. Furthermore, a fixed-test length is applied in each segment. Only in the segment with analogy and rule-based words does content control is applied to ensure the administration of 10 items from each category.

Expected Performance of the Spelling CAT

In the development of the CATs, pre-operational simulation studies play an important role. In these studies, the expected performance of the test is predicted. The possible fulfillment of the specifications and the cost of constraints of the CAT can be established. The expected measurement or decision accuracy is also set.

The expected performance of this multi-segment CAT was determined through simulation studies with the use of the DOT software (Verschoor, 2012). In the first phase of development the properties of each segment were set. On the basis of a number of simulation studies, the final settings of the presented algorithm were established. The results for the performance of the complete multi-segment spelling CAT are summarized in **Tables 3**, **4**. The basis for these results are simulations with 10,000 draws from each subpopulation given in **Table 2**.

From **Table 3**, the mean and standard deviation of the estimated $\widehat{\theta}$ are clearly almost the same as those in the population ability distributions from which the simulees were sampled

TABLE 3 | Measurement accuracy and reliability spelling CAT.

Population	Mean $\widehat{\theta}$	SD$\widehat{\theta}$	RMSE	RHO
4M	−0.70	0.35	0.14	0.87
4E	−0.54	0.32	0.12	0.88
5M	−0.24	0.34	0.10	0.93
5E	−0.03	0.38	0.09	0.95
6M	−0.01	0.38	0.09	0.95
6E	0.23	0.38	0.09	0.95
7M	0.37	0.40	0.10	0.94
7E	0.46	0.40	0.11	0.93
8M	0.59	0.36	0.12	0.90
8E	0.65	0.37	0.12	0.90

TABLE 4 | Test length (tl) in the spelling CAT with percentages of population with test length.

Population	Mean tl	SD tl	% with tl = 30	% with tl = 35	% with tl = 45
4M	34.37	2.06	15	84	1
4E	35.02	2.03	5	92	3
5M	36.42	3.65	1	84	15
5E	38.24	4.73	1	67	32
6M	39.68	5.01	0	53	47
6E	41.01	4.91	0	40	60
7M	41.85	4.65	0	31	69
7E	43.14	3.89	0	19	81
8M	44.07	2.90	0	9	91
8E	44.32	2.52	0	7	93

(compared with **Table 2**). For the accuracy of the measurement of the CATs, the root mean square error (RMSE) is considered. With $\theta_j$and $\hat{\theta}_j$, the true and the estimated ability (3) of simulee $j = 1, ..., n$ is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{j}=1}^{\mathrm{n}} (\theta_{\mathrm{j}} - \widehat{\theta_{\mathrm{j}}})}{\mathrm{n}}} \tag{6}$$

The multi-segment CAT shows that the best measurement results can be expected in the middle groups (5, 6, 7). This can also be concluded from the estimated reliabilities of the CAT. In simulation studies, this familiar index from classical test theory is easily estimated by

$$\mathrm{rho} = 1 - \left[(\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{se}(\widehat{\theta_{\mathrm{j}}})/\mathrm{n})/(\mathrm{sd}(\theta_{\mathrm{j}})) \right]^2. \tag{7}$$

In (7), $se(\hat{\theta}_j)$is the standard error of the ability estimate defined in (4).

From **Table 3**, although the youngest groups and the oldest children are measured slightly less precisely, the reliability of the spelling CAT is clearly high in all subpopulations.

In **Table 4**, the expected test lengths in the different populations are given. As expected, especially in the lower groups, the students will have advantages of shorter test lengths. In group 4E, for instance, only 1% of the population is expected to have the maximum test length of 45 items. Because of the goal of the spelling CAT, this is a very satisfactory property.

DISCUSSION

Because of technological and psychometrical developments, CAT has evolved from being a mere psychometric tool for the efficient estimation of ability to a testing mode that can serve different purposes of testing. CAT algorithms that serve the traditional summative functions of testing in education and that meet practical constraints have become available. In this study, three main approaches in formative testing are distinguished, and these are assessment for learning, diagnostic testing and data-based decision making. Applying CATs in these settings is demanding for other algorithms than commonly available. In assessment for learning applications for instance the optimal learning path is more important than efficient measurement; in diagnostic testing the emphasis could be more on the detections of misconceptions of students.

The multi-segment adaptive testing approach can be used to address the practical restrictions of CATs in summative settings but also to possibly serve a variety of testing purposes in formative settings. The multi-segment adaptive testing approach introduced in this study considers CATs as consisting of different segments that can each have its own algorithm and that is connected by branching rules. Many existing CAT applications can be considered as a cases fitting in the multi-segment approach.

In this study, the approach is illustrated by an operational spelling test of Dutch words used in a data-based decision-making formative testing application. The results show that multi-segment spelling CAT has potential to perform very well in actual practice.

The multi-segment approach has been shown to be very flexible. which will enable CATs to meet a variety of specific testing goals in education.

AUTHOR'S NOTE

The reported results in this paper are based on simulation studies. The input parameters for the simulations are based on published studies.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Babcock, B., and Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: do variable-length CATs provide efficient and effective measurement? *J. Comput. Adap. Test.* 1, 1–18. doi: 10.7333/1212-0101001

Barrada, J. R., Abad, F. J., and Veldkamp, B. P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema* 21, 313–320.

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theory of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, PA: Addison-Wesley).

Bosman, A. M. T., de Graaff, S., and Gijsel, M. A. R. (2006). "Double Dutch: the Dutch spelling system and learning to spell in Dutch," in *Handbook of Orthography and Literacy,* eds R. M. Joshi and P. G. Aron (Mahwah, NJ: Lawrence Erlbaum Associates), 135–150.

Brown, J. M., and Weiss, D. J. (1977). *An adaptive Testing Strategy for Achievement Test Batteries (Research Report 77-6)*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Chang, H_„ and Ying, Z. (1999). A-stratified multistage computerized adaptive testing *Appl. Psychol. Measure.* 23, 211–222.

Cheng, P. E., and Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Appl. Psychol. Measure.* 24, 257–265. doi: 10.1177/01466210022031723

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74, 619–632. doi: 10.1007/s11336-009-9123-2

Eggen, T. J. H. M. (2008). "Adaptive testing and item banking," in *Assessment of Competences in Educational Contexts,* eds J. Hartig, E. Klieme, and D. Leutner (Göttingen: Hogrefe & Huber), 215–234.

Eggen, T. J. H. M., and Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educ. Psychol. Measure.* 66, 713–734. doi: 10.1177/00131640021970862

Eggen, T. J. H. M., and Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Appl. Psychol. Measure.* 30, 379–393. doi: 10.1177/0146621606288890

Glas, C. A. W., and Verhelst, N. D. (1995). "Testing the rasch model," in *Rasch Models Foundations, Recent Developments, and Applications,* eds G. H. Fischer and I. W. Molenaar (New York, NY: Springer Verlag), 69–95.

Keuning, J., and Verhoeven, L. (2007). Screening for word-reading and spelling problems in elementary school: an item response theory perspective. *Educ. Child Psychol.* 24, 44–58.

Keuning, J., and Verhoeven, L. (2008). Spelling development throughout the elementary grades: the Dutch case. *Learn. Individ. Differ.* 18, 459–470. doi: 10.1016/j.lindif.2007.12.001

Keuning, J., Vloedgraven, J., and Verhoeven, L. (2011). *Wetenschappelijk verantwoording Screeningsinstrument dyslexie [Scientific justification]*. Arnhem: Cito.

Kingsbury, G. G., and Zara, A. R. (1991). A comparison of procedures for content sensitive item selection in computerized adaptive tests. *Appl. Measure. Educ.* 4, 241–261. doi: 10.1207/s15324818ame0403_4

Klinkenberg, S., Straatemeier, M., and Van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Comput. Educ.* 57, 1813–1824. doi: 10.1016/j.compedu.2011.02.003

Lord, F. M. (1970). "Some test theory for tailored testing," in *Computer-Assisted Instruction, Testing, and Guidance,* ed W.H. Holtzma (New York, NY: Harper and Row), 139–183.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.Copenhagen: The Danish Institute of Educational Research*. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

Reckase, M. D. (1989). Adaptive testing: the evolution of a good idea. *Educ. Measure. Issues Pract.* 8, 11–15. doi: 10.1111/j.1745-3992.1989.tb00326.x

Revuelta, J., and Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *J. Educ. Measure.* 35, 311–327. doi: 10.1111/j.1745-3984.1998.tb00541.x

Scalise, K., and Gifford, B. R. (2006). Computer-based assessment in e-learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms. *J. Technol. Learn. Assess.* 4, 1–44.

Schildkamp, K., Lai, M. K., and Earl, L. (Eds.) (2013). *Data-Based Decision Making in Education: Challenges and Opportunities*. (Dordrecht: Springer). doi: 10.1007/978-94-007-4816-3

Stobart, G. (2008). *Testing Times: The Uses and Abuses of Assessment.* London, UK: Routledge. doi: 10.4324/9780203930502

Sympson, J. B., and Hetter, R. D. (1985). "Controlling item-exposure rates in computerized adaptive testing," *Paper Presented at the Annual Conference of the Military Testing Association* (San Diego, CA).

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., and Eggen, T. J. H. M. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assess. Educ. Principl. Policy Pract.* 22, 324–343. doi: 10.1080/0969594X.2014.999024

Van der Linden, W. J. (2010a). "Constrained adaptive testing with shadow tests," in *Elements of Adaptive Testing,* eds W. J. van der Linden and C. A. W. Glas (NewYork, NY: Springer), 31–55.

Van der Linden, W. J. (2010b). "Sequencing an adaptive test battery," in *Elements of Adaptive Testing,* eds W. J. van der Linden and C. A. W. Glas (New York, NY: Springer), 103–119.

Van der Linden, W. J., and Glas, C. A. W. (eds.). (2010). *Elements of Adaptive Testing*. NewYork, NY: Springer. doi: 10.1007/978-0-387-85461-8

Verhelst, N. D., Glas, C. A. W., and Verstralen, H. H. F. M. (1995). *One-Parameter Logistic Model (OPLM) [Computer software]*. Arnhem: Cito.

Verschoor, A. J. (2012). *Design of Optimal Tests (DOT) [Software (update 2012) and user's manual]*. Arnhem: Cito(Freely available for scientific purposes)

Wainer, H. (Ed.) (2000). *Computerized adaptive testing. A primer, 2nd Edn*. Hilsdale, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9781410605931

Warm, T. A. (1989). Weighted maximum likelihood estimation of ability in item response theory. *Psychometrika* 54, 427–450. doi: 10.1007/BF02294627

Wauters, K., Desmet, P., and Van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *J. Comput. Assisted Learn.* 26, 549–562. doi: 10.1111/j.1365-2729.2010.00368.x

William, D., and Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *Br. Educ. Res. J.* 22, 537–548.

Yan, D., Von Davier, A. A., and Lewis, C. (Eds.) (2014). *Computerized Multistage Testing: Theory and Applications*. Boca Raton, FL: Chapman and Hall/CRC.

Zenisky, A., Hambleton, R. K., and Luecht, R. M. (2010). "Multistage testing:Issues, designs, and research," in *Elements of Adaptive Testing,* eds W. J. van der Linden and C. A. W. Glas (New York, NY: Springer), 355–372.

Zheng, Y., and Chang, H.-H. (2011). "Automatic on-the-fly assembly for computerized adaptive multistage testing," *Paper Presented at the International Association for Computerized Adaptive Testing Conference* (Pacific Grove, CA).

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.