Using Expert Elicitation and Adaptive Testing to Shorten the End of Primary School Test (EPST)

Research Report

Marie Buijs, Utrecht University
Supervised by prof. dr. A. G. J. Van De Schoot

December 15, 2019

1 Introduction

During the transition from primary to secondary education in the Netherlands, students are divided into tracks. Several tracks prepare students for different types of secondary education: VWO (pre-university education), HAVO (general secondary education) and VMBO (pre-vocational education) (see Figure 1). Each student receives an advice on what type of secondary education to pursue. This advice is based on two sources: the teacher and the End of Primary School Test (EPST) that is taken at the end of primary school. The EPST takes multiple days to complete and measures ability in two domains; language and mathematics. Both the advice from the EPST and the teacher can be single or composite. A single advice consists of one level (e.g. HAVO advice), while a composite advice consists of two consecutive levels (e.g. HAVO/VWO advice).

Switching between tracks is possible and relatively easy in the first three years of secondary education. This may lead to a (partial) correction of the advice when the initial placement does not match student ability [1]. Although switching between tracks is possible, the initial track placement is an important factor in a students' further education. Therefore, the advice a student receives on what education they should pursue needs to be as accurate as possible. However, what should be decided when there is an extreme mismatch between the result of the EPST and the advice by the teacher? The current study will introduce a new way of combining the EPST with the teachers' advice. In what follows, we provide a historical overview of the ongoing debate in the Netherlands about whether the teacher or test provides the better advice.

1.1 Historical overview

Before 2015, the advice students received on what track to pursue, was based on the score they achieved on the EPST. The EPST was designed by Cito, the main educational testing service in the Netherlands. Not everyone was in favor of basing the advice solely on the EPST. Critics argued that the test captures just one moment in time [2]. In addition, they argued that the test does not to take other factors into account such as discipline, motivation and attitude [3].

In 2015, there was a change in policy [4]. The main change being that the teachers' advice would be leading in school advisement. However, the result of an EPST may be used to raise the advised level by the teacher, but not to lower it. In addition to making the teachers' advice leading, multiple test providers would be allowed to publish their version of the EPST [4], [5], [6]. Furthermore, taking an EPST became obligatory after the teachers advice is communicated. The idea of making the teachers' advice leading in school advisement was not without controversy. Opponents of this idea argued that teachers are not sufficiently capable of weighing information and obtaining the right decision [7].

In the summer of 2019, the House of Representatives published a letter [8] evaluating the policy change since 2015. The letter cites the Centraal Plan Bureau (CPB), who argue that the EPST should play a bigger role in the transition to secondary education [6]. The authors state that inequality has increased due to the opportunity for wealthy parents to arrange additional training when they disagree with the advised level.

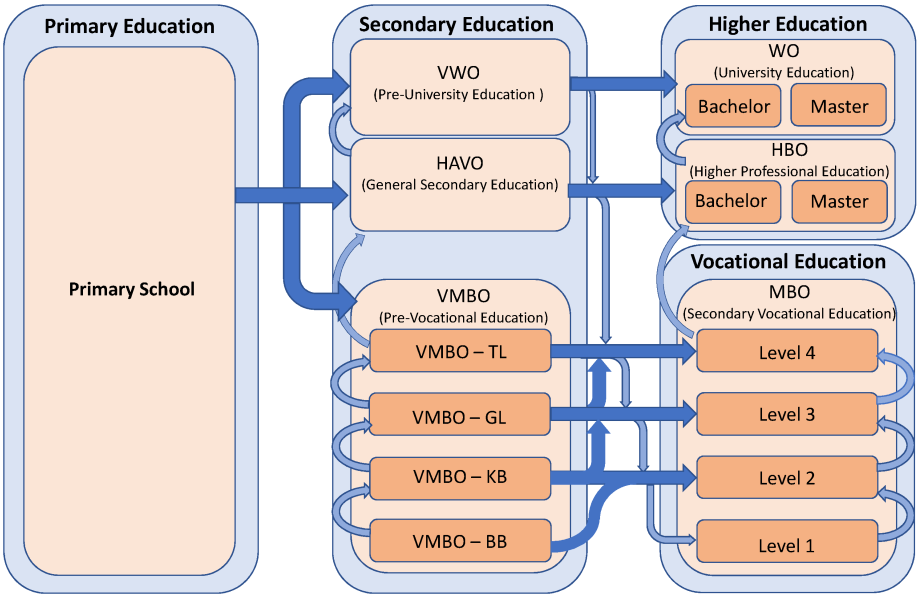


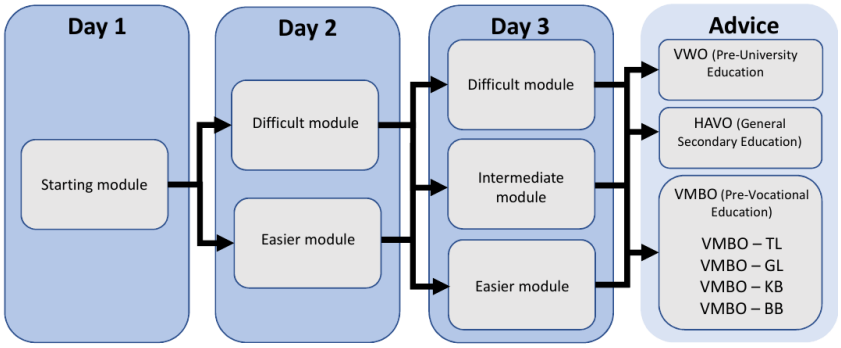
Figure 1: A graphic representation of the Dutch educational system. After finishing primary school, that children attend from approximately age 4 to 12, children transition to secondary education that consists of several levels. VWO (pre-university education) prepares students for University education (WO). HAVO (General Secondary Education) prepares students for higher professional education (HBO). VMBO (Pre-Vocational Education) consists of 4 different subtypes ranging from the most theoretical (TL track) to the most vocational (BB track). The different VMBO levels prepare students for different levels of Vocational Education and Training (MBO).

Furthermore, they argue that the incentive to perform well has decreased for students who are content with the advised level. In contrast, Lek & Van De Schoot are in favor of using both the EPST and teachers' advice [9]. The authors show that both have their advantages and consider a combination of both to be best. The House of Representatives decided not to make the EPST leading again, but to change the timing of the EPST and teachers' advice. By this change, the EPST is taken two weeks after the teachers decide on their advice to prevent additional training for the test. Furthermore, it would be possible to refrain from communicating the teachers' advice to keep students motivated [8].

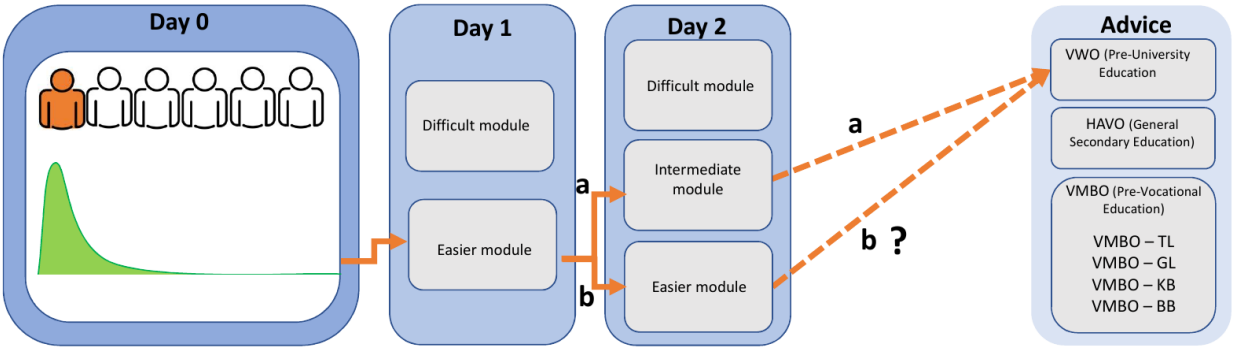
1.2 Adaptive testing

Since 2018, one of the EPSTs schools can choose from, is the multistage adaptive version of the EPST by Cito [10]. In adaptive testing, the selection of an item is adapted to an individuals performance using Item Response Theory (IRT). IRT assumes that an individual's probability of giving a certain response depends on person- and item characteristics (e.g. ability and item difficulty) [11, [12]. The latent ability of each individual is estimated after completion of each item. Subsequent items are selected to match this estimated ability level. The most information about the ability of an individual is obtained when item difficulty aligns with the true ability, resulting in more precise ability estimates.

For the adaptive EPST, Cito implemented a variation on adaptive testing that administers various sets of items (modules), called Multistage testing (MST) [13]. Instead of adapting what item is presented, an entire module is presented based on the match of its difficulty with the estimated student ability. In MST the number of stages and modules can be varied. In the adaptive multistage EPST that is implemented by Cito, each day corresponds to one stage that has respectively one, two and three modules per stage (see Figure 2A).



(A)



(B)

Figure 2: (A) Current adaptive multistage design used by Cito. (B) Example of a new (hypothetical) design, where the teachers' expectation (as elicited on day 0 using a digital tool [15]) about a student determines the difficulty of the module on day 1. Both lines a and b show a possible route of a student who is capable of VWO but is expected to score low by their teacher.

In the current design, the first stage (routing stage), takes place on day one. On this day, all students take the same module which gives an initial estimate of ability. This ability estimate determines what module is presented to the students in the second stage: an easier or a more difficult module. Performance in the second stage provides a new estimate of ability. In the final stage on day three, based on this new estimate, the student is once more presented with an easier, intermediate or more difficult module [10].

1.3 Combining the adaptive EPST and teachers opinion using expert elicitation

In the future, the teachers' opinion could be integrated with the multistage adaptive EPST of Cito. Ideally, not only the teachers' opinion, but also their (un)certainty about their opinion is taken into account. A method to formalize the teachers' opinion to take uncertainty into account, is expert elicitation. Expert elicitation is the process of formalizing experts opinions in a way that these can be used statistically [14]. Lek and Van De Schoot tested an online application with icons to elicit teachers' judgments about their students' ability [15]. The application presents a distribution that represents the teachers' knowledge and their (un)certainty. The information from this application could potentially enrich the EPST design by combining expert elicitation [14] with multistage adaptive testing [13] (see Figure 2B), see [16]. Both the paper-based EPST and the current version of the multistage EPST take three days to complete [10]. A method to shorten the overall duration of multistage EPST by one day, could be to use expert knowledge of the teacher to determine the starting level of students on day one (see Figure 2B, day 0).

However, immediately implementing this method would be a political challenge. As the previous debate underlines, allowing the teachers' opinion to influence the EPST score of students is a sensitive matter. One argument is that it would be unfair if a student is provided with a completely different advice because of their teachers' judgment. Therefore, it will be investigated, if the teachers' knowledge is used to determine the starting level of students, what the effect will be on the final advice when an extreme mismatch between teacher and test occurs. For example, consider a student who is capable of obtaining VWO advice in the current adaptive EPST of Cito (see Figure 2A). Now assume that the teacher expects this student to score low (See Figure 2B). The teachers' judgment results in the student starting with the easier module on day one. This student performs adequately on both day one and two, which would result in the same VWO advice as the student would have obtained in the original design (Figure 2B, situation a). However, if the same student had performed suboptimal on day one, for example due to stress, but nevertheless makes the module on day two well, this same student should be able to obtain VWO advice (Figure 2B, situation b).

1.4 The current study

First, it is important to determine how often a severe mismatch occurs between the EPST advice and the teachers' advice. Although research about the (dis)agreement between EPST and teachers' advice has been conducted before, information about the frequency of extreme mismatches is missing [9]. When it is clear how often extreme mismatches occur, synthetic data will be generated and used in a simulation study. This simulation study will examine the effect of combining expert elicitation with the adaptive testing procedure when extreme mismatches occur (see Figure 2B). It will be estimated how many items are needed to provide students with the same advice as the regular adaptive EPST. The number of items needed will be compared to the 140 items used in the current test version of Cito [17]. In order to implement expert elicitation methods in the actual testing procedure of Cito, two conditions have to be met. First, the simulation study has to show that using expert elicitation is a feasible method to determine the starting level. Second, the duration of the test has to stay reasonable. If these conditions are met, using teachers' expertise would lead to a shorter duration of the overall test, aligning test items with students' capabilities.

The current report focuses on the first part of the study. Using longitudinal data from Statistics Netherlands (CBS), it is investigated how often extreme mismatches between the EPST advice and the teachers' advice occur in the Netherlands. Furthermore, it is investigated what track students with an extreme mismatch follow in the third year of secondary education. The current study provides a starting point to generate synthetic data that will be implemented in a simulation study. This simulation study will investigate the effectiveness of using expert elicitation to determine the starting level in multistage adaptive testing.

2 Methods

2.1 Data

The data used for this study are non-public microdata from Statistics Netherlands (CBS). All analyses were performed using R [18] in the remote secured online environment of CBS. The used datafiles are the "Citotab" data from the school year 2014/2015 and the "Onderwijsinschrtab" data from the year 2015/2016 and 2017/2018, corresponding to the first and third year of secondary education.[1] The Citotab file contains information about the results of the EPST and the teachers' advice of students in the Netherlands [19]. The Onderwijsinschrtab file contains information on the enrollment of all students in secondary and higher education in the Netherlands [20]. All files were combined to match the EPST score and teachers' advice from the Citotab file with the placement in secondary education one and three years after completing primary school from the Onderwijsinschrtab files. In the CBS files, the VMBO-GL and VMBO-TL tracks were combined and referenced to as VMBO-GT. These tracks are similar as the VMBO-GL track has one less theoretical course and one more vocational course than the VMBO-TL track.

[1] For information about accessing these files:
https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research

2.2 Inclusion criteria

The Citotab file includes information about students whose primary school authorized Cito to share their results with CBS. Only students who met the following inclusion criteria were included in this study: (1) students were registered in the GBA (Administration of the Municipality); (2) both the teachers' advice and the EPST score were available; (3) the placement in the first and third year after completing primary school was known and (4) student were not enrolled in practical education in the first or third year of secondary education[2]. After selection, 119754 students were included (see Figure 3).

2.3 Analysis

This study investigated how often extreme mismatches between the EPST advice and the teachers' advice occurred in the Netherlands. To quantify this mismatch, the difference between each single advice was considered one level difference. The difference between an overlapping single and a composite advice was considered to be .5 level difference. For example, an overlapping composite advice of the teacher (HAVO/VWO) and a single advice of the test (HAVO) was considered .5 level difference, whereas the difference between two single advices (HAVO and VWO) was considered one level difference. The mismatches between teachers' advice and EPST advice are obtained by creating a contingency table. For the subset of students whose EPST- and teachers' advice severely mismatch, their advices were compared with the placement in the third year of high school as an indication of their true ability.

2.4 Ethical approval

This study is approved by the Ethics Committee of the Faculty of Social and Behavioral Sciences of Utrecht University (FETC19-215). Before publication, results were tested by CBS to prevent publication of identifiable information about individuals[3].

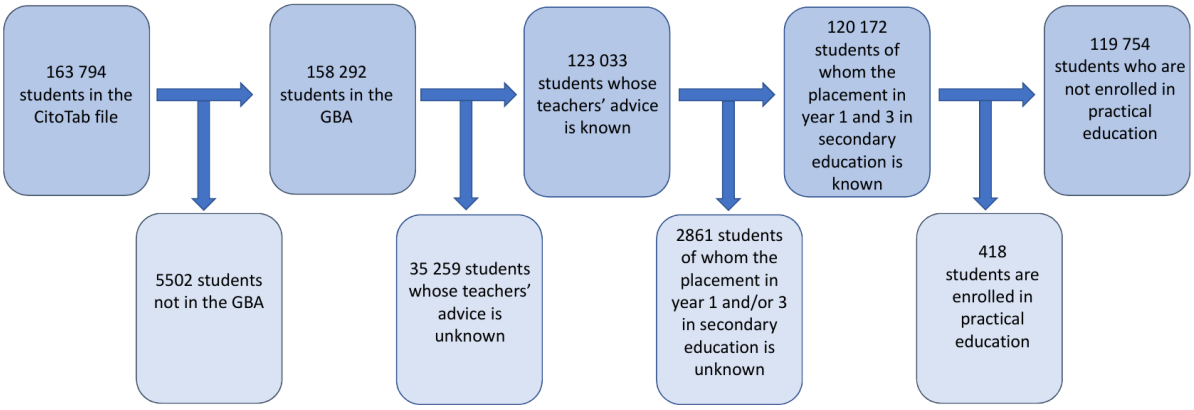


Figure 3: Schematic overview of the number of students in the datafile (n = 163794) and the numbers of students included in the present study (n = 119754).

[2]Practical education focusses on children who are unlikely to be able successfully finish VMBO-BB. This criterion was used because the lowest category in the EPST is VMBO-BB

[3]Also see:
https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/export-van-gegevens

3 Results

When examining the amount of overlap between teachers' and EPST advice, it showed that in 72.8% of the cases, the teachers' and EPST advice were equal or differ by .5 level (see Table 1). This result indicates that most of the time, the EPST and teacher are in reasonable agreement. However, the EPST and teacher do not always agree. The most severe mismatch possible, where one advice is VWO and the other advise is VMBO-BB, occurred only twice (.002%). A less extreme, but still severe mismatch of three or more levels affected 120 students (.1%).

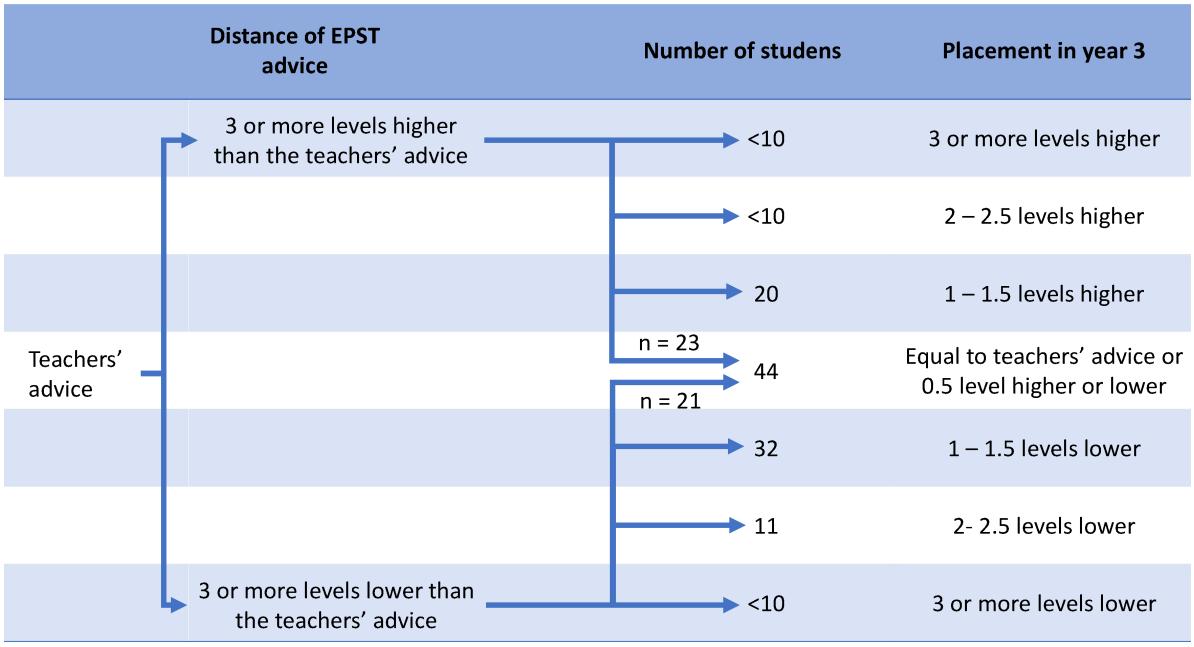
For most students whose EPST- and teachers' advice differed by three or more levels, the eventual placement in the third year of secondary education lied 1.5 level of less away from the teachers' advice (see Figure 4). This result implies that the advice of the teacher is a better indication for eventual track placement than the EPST advice for most students.

Table 1: Number of times that each level of mismatch occurs (n=119754)

Levels of mismatch	Frequency (Percentage)
EPST advice equals teachers' advice	44975 (37.5%)
EPST- and teachers' advice differ 0.5 level	42238 (35.3%)
EPST- and teachers' advice differ 1 level	20371 (17.0%)
EPST- and teachers' advice differ 1.5 levels	8612 (7.2%)
EPST- and teachers' advice differ 2 levels	2757 (2.3%)
EPST- and teachers' advice differ 2.5 levels	681 (0.6%)
EPST- and teachers' advice differ 3 levels	104 (0.086%)
EPST- and teachers' advice differ 3.5 levels	14 (0.012%)
EPST- and teachers' advice differ 4 levels	2 (0.002%)

Table 2: Contingency table of EPST advice and teachers' advice

EPST → Teacher ↓	VMBO-BB	VMBO-BB/KB	VMBO-KB	VMBO-GT	GT/HAVO	HAVO	HAVO/VWO	VWO	Marginal Teachers' advice
VMBO-BB	4865	2694	727	470	174	32*	5*	1*	8968
VMBO-BB/KB	1017	1255	489	454	195	40	12*	1*	3463
VMBO-KB	1706	4591	2839	3402	1978	586	230	15*	15347
VMBO-GT	525	2935	3183	6178	6365	3167	1863	185	24401
GT/HAVO	40	414	634	1852	2996	2316	2045	283	10580
HAVO	15*	153	325	1400	3814	5095	8844	3155	22801
HAVO/VWO	1*	16*	44	236	818	1592	4771	3680	11158
VWO	1*	7*	14*	57	229	768	4984	16976	23036
Marginal EPST advice	8170	12065	8255	14049	16569	13596	22754	24296	119754



Note. Fewer than 10 students were placed in a broad track (undifferentiated track).

Figure 4: Graphical representation of the eventual placement of the students whose EPST- and teachers' advice differ by three or more levels (n=120). 96 students (80%) are placed within 1.5 level of the teachers' advice.

4 Discussion

The advice students receive on what educational track they should pursue, can be based on the teachers' advice and the EPST results. The debate on what source provides the better advice is still ongoing. This paper introduced a new idea to combine both sources of information to shorten the EPST, while still all advices can be obtained by students. To investigate the feasibility of this method, a simulation study will be performed. In order to create realistic starting points for this simulation, the current study investigated how often a severe mismatch occurred between the EPST advice and the teachers' advice occurs.

The results show that for most students the EPST- and teachers' advice are in agreement. Although only a small percentage (0.1%) of students is affected by a severe mismatch of more than three levels, this mismatch may have a vast influence on students' school career. For most students who are affected by a severe mismatch, eventual track placement in the third year of secondary education matches closer to the teachers' advice than the EPST advice. For these students, the influence of the teachers' knowledge on their EPST score when using expert elicitation to determine the starting level of the EPST would not necessarily be disadvantageous. However, for a few students with a severe mismatch eventual track placement in the third year of secondary education matches closer to the EPST advice than the teachers' advice. For these students, the influence of the teachers' opinion on their EPST could be disadvantageous. Thus, it is important that students are able to obtain any advice regardless of their teachers' opinion.

4.1 Future research

The current results will be used a starting point for the data generation in a simulation study. This simulation study will examine the effect of combining expert elicitation with the adaptive testing procedure. It will be investigated if the EPST can be shortened by using teachers' opinions to determine the starting level while each student is able to obtain any advice regardless of their teachers' opinion. The teachers' advice will be used as an approximation for the teachers' expert knowledge that could be elicited using the digital expert elicitation application [15]. The regular EPST advice will be used as an approximation for the advice a student would obtain in the multistage adaptive EPST [10]. In the simulation study, it will be estimated how many items are needed to provide the student with the same advice as the regular EPST. This number of items will be compared with the 140 items that are currently used in the adaptive EPST to determine if the duration of the test is improved compared to the original design.

References

[1] Onderwijsinspectie, "De staat van het primair onderwijs 2019," 2019.

[2] S. Kamerman and J. Vasterman, "De leerkracht weet het vaak écht beter dan de Citotoets." *NRC.next*, Mar. 2015.

[3] A. De Regt, "Welkom in de ratrace; over de dwang van de Cito-toets." vol. 31, no. 3, pp. 297–320, 2004.

[4] Staatssecretaris van Onderwijs, Cultuur en Wetenschap, "Toetsbesluit PO," 2014.

[5] "Wijzigingswet Wet op het primair onderwijs, enz. (Centrale eindtoets en leerling- en onderwijsvolgsysteem primair onderwijs)." Aug-2014.

[6] "Policy Brief. De waarde van eindtoetsen in het po." Centraal Plan Bureau (CPB), 2019.

[7] S. Niessen and R. Meijer, "De leerkracht is geen meetinstrument." *NRC.next*, Apr. 2016.

[8] A. Slob, "Eindevaluatie Wet eindtoetsing po." Jun-2019.

[9] K. Lek and R. Van De Schoot, "The Transition from Primary to Secondary Education in the Netherlands: Who Knows best, the Teacher or the Test?" *De Psycholoog*, vol. 54, no. 4, 2019.

[10] "Terugblik 2018: Resultaten Centrale Eindtoets 2018." College voor Toetsen en Examens (CvTE), 2018.

[11] W. Van der Linden and Glas, C. A. W., *Elements of Adaptive Testing*, 1st ed. New York: Springer-Verlag, 2010.

[12] D. Harris, "NCME Instructional Module: Comparison of 1-, 2-, and 3-Parameter IRT Models," *Instructional Topics in Educational Measurement*, 1989.

[13] A. Mead D, "An Introduction to Multistage Testing," vol. 19, no. 3, pp. 185–187, 2006.

[14] A. O'Hagan and et al., *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons, 2006.

[15] K. Lek and R. Van De Schoot, "Development and Evaluation of a Digital Expert Elicitation Method Aimed at fostering Elementary School Teachers' Diagnostic Competence." *Frontiers in Education*, vol. 3, p. 82, 2018.

[16] S. Berger, A. J. Verschoor, T. J. H. M. Eggen, and U. Moser, "Improvement of Measurement Efficiency in Multistage Tests by Targeted Assignment," *Frontiers in Education*, no. 4, p. 1, 2019.

[17] "De centrale eindoets in 2019; informatie voor scholen." College voor Toetsen en Examens (CvTE), 2018.

[18] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2019.

[19] "Documentatie Kenmerken van deelnemers aan de Eindtoets Basisonderwijs van Cito (CITOTAB)." Centraal Bureau voor de Statistiek (CBS), 2019.

[20] "Documentatie Kenmerken van inschrijvingen in diverse onderwijssoorten (ONDERWIJSINSCHRTAB)." Centraal Bureau voor de Statistiek (CBS), 2019.