Research Master's programme Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University, the Netherlands

MSc Thesis Marie Buijs (5714702)

The Influence of Teacher-Based Routing on the Accuracy and Classification of Multistage Tests

May 2020

Supervisors:
Dr. D. Hessen (UU)
Prof. dr. A.G. J. Van de Schoot (UU)
K. Lek, MSc. (Cito)
Dr. R. Feskens (Cito

Second grader:
Prof. dr. ir. M. J. C. Eijkemans

Preferred journal of publication: Multivariate Behavioral Research
Word count: 9123

The Influence of Teacher-Based Routing on the Accuracy and Classification of Multistage Tests

M.J. Buijs[a]*

[a] *Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands*

* m.j.buijs@uu.nl

The Influence of Teacher-Based Routing on the Accuracy and Classification of Multistage Tests

Multistage testing (MST) is a test administration method where the difficulty of a test adapts to the ability of each examinee. In the first stage of traditional MST designs, all students take the same module to determine the initial ability on which module allocation in the subsequent stage(s) is based. Routing based on the teacher's judgment may be an interesting alternative to the regular routing stage in MST designs since it resolves the need of the regular routing stage. In this paper, we investigated the effect of replacing the routing stage of a multistage test with teacher-based routing. To make this study as realistic as possible, the simulations were based on the Dutch End of Primary School Test (EPST), which provides track recommendations. Through simulations, we compared the precision of ability estimates and the differences in EPST track recommendation with and without teacher-based routing. The number of modules available in the first stage were varied. Furthermore, there was a focus on the situation in which module allocation based on the teacher was not in line with student's performance. Results indicate that the teacher-based routing design provides students with slightly less precise ability estimates than the regular-routing EPST. Most students obtain an appropriate track recommendation when teacher-based routing is implemented, although this is less often the case for students who were misdirected in the first stage. In conclusion, the gain of information by assigning students to a module that matches their ability does not completely make up for the loss of information by shortening the test. However, with further research, the teacher-based routing MST might be implementable in high-stakes situations as well.

Keywords: teacher-based routing; multistage testing; simulation

Introduction

Multistage testing (MST) is a test administration method where the difficulty of a test adapts to the ability of each examinee (Mead, 2006). Unlike linear tests – in which each examinee is presented with the same items – in an MST, students are presented with different item sets (modules) based on their performance in previous stages. In the first stage of an MST, referred to as the routing stage, all students usually take the same

module. This module is used to determine the initial ability on which module allocation in the subsequent stage(s) is based. The number of stages (i.e. the number of tests parts where each student is assigned a different module) and the number of modules available per stage (i.e. the total number of item sets available) can be varied (Berger et al., 2019; Mead, 2006).

MST designs are used more and more in educational settings. For example, an increasing number of international large-scale assessments use an MST design, such as the Program for International Student Assessment (PISA), the Law School Admission Test (LSAT) (Armstrong et al., 2004; OECD, 2019; Wainer & Wang, 2000; Yamamoto et al., 2009) and many others (Bock & Zimowski, 1998; Luecht et al., 2006; Luecht & Nungester, 1998; Van der Linden & Glas, 2010). In the Netherlands, an example of a large-scale MST assessment is the End of Primary School Test by Cito (College voor Toetsen en Examens (CvTE), 2015). The reason for their increasing popularity is that MST designs have a clear advantage over regular linear testing designs when the population taking the test is diverse and a wide range of abilities is present. To accommodate such a wide range, in regular linear tests items are selected that have varying degrees of difficulty with the mean difficulty centered around the mean ability level of the population. This means that all students are presented with at least a few items which difficulty does not align with their ability level, especially students with relatively high and low abilities. When item difficulty does not align with student ability, less information about the ability of a student is obtained. This results in bigger standard errors and thus less precise ability estimates. In addition, this may lead to poorer performance due to a loss of motivation (Asseburg & Frey, 2013). When the difficulty of a test more closely matches the ability level of a student, such as in MST designs, the difficulty of

items is better aligned with the ability of a student. This results in smaller standard errors and more precise ability estimates.

A drawback of the MST design is the aforementioned routing stage. This stage lengthens the MST and, since each student completes the same routing module, this module suffers from the same disadvantages as those of regular linear tests. Furthermore, by providing each student with the same routing module, we act as if no pre-existing information about the examinee taking the test is available. However, in many instances some knowledge about the ability of students is available, such as previous test scores or teacher observations. As Berger et al. (2019) have shown, using the knowledge and observations of the teacher to determine the starting level of students would eliminate the need of a routing stage. Therefore, routing based on the teacher's knowledge may be an interesting alternative to the regular routing stage in MST designs. However, teacher knowledge may not always be in line with the performance of students. Particularly when the teacher's knowledge is based on both previous test scores and more subjective teacher observations, bias against some students may be introduced (Wang et al., 2018). Therefore, it is especially important that the teacher's view does not influence the eventual test score via the teacher-based routing stage.

In the current article, a simulation study is performed to investigate the effect of replacing the routing stage of a multistage test with teacher-based routing. That is, routing based on a judgment of the teacher; for instance, based on their observations and/or previous test results of examinees. To make this simulation as realistic as possible, it will be based on the earlier mentioned Dutch End of Primary School Test (EPST). This high-stakes educational test is used in the Netherlands to present each student with a recommendation about what track they should pursue in secondary education. Investigating teacher-based routing is especially beneficial in this context, as the current

routing stage of this test is relatively long. The present simulation study specifically focuses on the accuracy of ability estimates (i.e., their standard errors) and the differences in track recommendation of the EPST with and without teacher-based routing. There is a specific focus on the situation in which module allocation based on the teacher is not in line with student's performance to ensure that a biased teacher judgement minimally influences the eventual ability estimates and track recommendations.

Taken together, the following research questions will be answered: (1) What is the effect of replacing the regular routing stage of a multistage test design by teacher-based routing in terms of the precision of the ability estimates and the track recommendations produced by the test? (2) Is this effect different for students who are initially misallocated by the teacher? (3) What is the effect of changing the number of modules in the first stage of a teacher-based routing design?

First, the theory underlying MST, namely item response theory, and the background of teacher-based routing are discussed. After that, the details of the Dutch End of Primary School Test (EPST) by Cito are presented. Next, the methods and the results of the simulation study are presented. Finally, the implications of teacher-based routing are discussed.

[Figure 1 near here]

Theoretic Background

MST-design

As described previously, MST designs consist of several modules and stages. After each stage, a provisional ability estimate is obtained for each student. This provisional ability estimate is based on information from the module(s) that a student has taken (Magis et al., 2017). Based on the provisional ability estimate, the next module is selected. Selection

can be based on a certain cutoff score (i.e. certain number of items correct or thresholds for the provisional ability estimate) or on an information criterion (i.e. selecting the most informative module at the current ability estimate; see Magis et al., 2017). The earlier mentioned PISA student assessment, for example, uses a number correct criterion (Yamamoto et al., 2009). Models from the Item Response Theory (IRT) framework are used to estimate the ability of students in MST designs. IRT models assume that an individual's probability of giving a certain (in)correct response depends on both person characteristics (the ability of a student) and item characteristics (for instance, the difficulty of the item). Depending on which model is used, additional item characteristics are taken into account such as item discrimination (i.e. how well the item is able to discriminate between students with different ability levels) and item guessing (i.e. how likely students are to guess the answer to the item) (Eggen & Sanders, 1993; Harris, 1989; Magis et al., 2017; Van der Linden & Glas, 2010). The current study will use the one parameter logistic model (1PL model) to estimate the ability of students, as this is the most straightforward and simplest IRT model.

The 1PL model only takes the ability parameter of person *i*, θ_i and difficulty parameter of item *j*, β_j into account. The 1PL model assumes that the probability of person *i* answering item *j* correctly, given the ability of person *i* and the difficulty of item *j*, is given by the following equation:

$$P(X_{ij} = 1 \mid \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1+\exp(\theta_i - \beta_j)}, \qquad \text{(Equation 1)}$$

where $X_{ij} = 1$ indicates a correct answer on item *j* by person *i*.

Two observations can be made based on Equation 1. First, the probability of answering an item correctly is monotonically increasing. Thus, when an item gets easier, or when a student's ability gets higher, the probability of answering an item correctly

increases as well. Second, when the difficulty of an item and the ability of a student match perfectly (i.e. $\theta_i = \beta_j$) the probability of answering an item correctly is 0.5 (Eggen & Sanders, 1993; Magis et al., 2017).

The ability estimates as obtained by the 1PL model are not perfect and come with some uncertainty. This uncertainty is indicated by the standard error of measurement (SEM) that is related to the item- and test information function. The item information function is built on the idea that every item in a test provides some information about the ability of a student. The amount of information obtained by an item *j* for examinee *i* as expressed by the item information function is as follows:

$$I_j(\theta_i, \beta_j) = P(X_{ij} = 1 \mid \theta_i, \beta_j) P(X_{ij} = 0 \mid \theta_i, \beta_j) \quad \text{(Equation 2)}$$

Where $P(X_{ij} = 1 \mid \theta_i, \beta_j)$ is the probability of answering item *j* correctly (see Equation 1) and $P(X_{ij} = 0 \mid \theta_i, \beta_j)$ is the probability of answering item *j* incorrectly; *1* - $P(X_{ij} = 1 \mid \theta_i, \beta_j)$. The most information about the ability of a student is obtained when $I_j(\theta_i, \beta_j) = 0.5 \cdot 0.5 = 0.25$, which occurs when the ability parameter of a student perfectly matches the difficulty parameter of the item (at $\theta_i = \beta_j$). (Eggen & Sanders, 1993; Magis et al., 2017). The information gathered over a set of items for examinee *i* is equal to the sum of the item information functions, and is called the test information function:

$$I(\theta_i, \beta_j) = \sum_j I_j(\theta_i, \beta_j). \quad \text{(Equation 3)}$$

The standard error of measurement (SEM) is equal to the square root of the inverse test information:

$$SEM(\theta_i) = \sqrt{\frac{1}{I(\theta_i, \beta_j)}} = \sqrt{\frac{1}{\sum_j P(X_{ij} = 1 \mid \theta_i, \beta_j) P(X_{ij} = 0 \mid \theta_i, \beta_j)}} \quad \text{(Equation 4)}$$

When the ability of a student matches the difficulty of all items throughout the test, the most information about the ability of a student is obtained. In that case, the SEM is lowest, and a more precise ability estimate is obtained. Thus, it pays off to match the difficulty of the test as closely as possible to the ability of a student. For this reason, MST attempts to match the difficulty of the items to the ability of a student. Currently, a routing stage is deemed necessary to obtain an initial ability estimate to match the difficulty. However, the routing stage may be replaced by other available pre-existing information (for instance, from the teacher). By presenting students with items that are adapted to their ability level right from the start of the test, fewer items are needed to obtain estimates with the same precision.

Teacher-Based Routing

There are different ways in which pre-existing information from, for instance, the teacher can be included in (multistage) adaptive testing. Veldkamp & Matteucci (2013), for instance, have investigated the possibility of including pre-existing information in a Bayesian prior. This has been done in the context of (Bayesian) computerized adaptive testing (CAT), where the difficulty of each item adapts to the performance of a student instead of modules as in MST. By including this prior information, the very first item presented is already adapted to the indication of student ability. The study by Veldkamp & Matteucci (2013) suggests a considerable reduction of necessary test duration when prior test scores are included, especially for more extreme ability levels. However, an inaccurate prior in BCAT does result in longer and less informative tests. Furthermore, when BCAT is applied, the ability estimate of each candidate is not only based on their responses but on the prior knowledge as well. This may not be desirable in high-stakes educational testing, especially when the prior information consists of the teacher's knowledge which could be subject to social bias.

As both CAT and MST are based on the same principle of adaptivity and are built on the IRT framework, the Bayesian approach can be extended to multistage testing. However, as in BCAT, directly weighting pre-existing information in the estimation of ability may not be desirable and thus may better be avoided. One way to use pre-existing information without directly weighting it in the estimation of ability is to allocate students directly to an appropriate module based on teacher judgment instead of using a routing stage. As the indication of student ability by the teacher only influences module allocation in the first stage, the MST allows for compensation for a mismatch between the teacher's knowledge and actual performance.

This idea of teacher-based routing has been investigated by Berger et al. (2019) who compared the effect of replacing the current routing stage by multiple modules for which module allocation was based on pre-existing information (for instance from the teacher) in MST with a regular MST. This study found that the MST where the routing stage was replaced with pre-existing information was generally more efficient than the other designs, especially if the pre-existing information variable was a good indicator of students' true ability. Thus, fewer items were needed to obtain the same information as in the other designs.

From a practical point of view, using pre-existing information from the teacher to allocate students to different sets of items is not an entirely novel idea in the Netherlands, nor at Cito. Around 1987, in the first periodic survey of the educational level for mathematics ("Periodieke Peiling van het onderwijsniveau – rekenpeiling") by Cito, a targeted test was used. In this design, teachers were asked to allocate each student to either the easier or the more difficult module (Verhuls, 1989). However, due to a lack of calibration of items, the items that were considered difficult were in reality not much harder than the easier or shared items resulting in little gain in information.

Taken together, the previous studies and applications indicate that teacher-based routing can be more efficient than regular MST designs. Despite these promising results, more insight is needed into the effect of teacher-based routing for students who are misdirected to a module that is not in line with their ability. Although on the population level measurement efficiency might be high, based on previous research it is uncertain whether subpopulations of students (for instance students with relatively high abilities who are allocated to the module with the lowest difficulty and vice versa) can be disadvantaged or whether compensation for the misallocation occurs in the following stages. Furthermore, in previous research only one teacher-based routing MST design was compared to other regular MST designs. In this teacher-based routing design, allocation of students by the teacher was based on three modules (easy, moderate and difficult). However, the effect of increasing or decreasing the number of modules available to the teacher is unknown. Finally, in the study by Berger et al. (2019) the MST was not shortened by replacing the routing stage by teacher-based routing. Students still completed the same number of items in the routing stage and only the difficulty of the items depended on the teacher's judgment. It is currently unknown if the MST-test can be shortened without loss of information by using teacher-based routing.

The current paper adds to the existing literature by highlighting the situation in which students are misallocated. It is investigated if, and if so to what extent, these misallocated students are disadvantaged in terms of their eventual ability estimates. Furthermore, the teacher-based routing MST designs are varied with regard to the number of modules available in the first teacher-based stage. Finally, it is investigated if the typical increase in information due to teacher-based routing is enough to compensate for a shortened test length when the routing stage is removed. To obtain a realistic simulation, data from the high-stakes EPST are used as a basis. Because the EPST test is used to

provide students with a track recommendation, next to the ability estimates it is also investigated how well teacher-based routing designs are able to provide students with such a recommendation. The EPST is discussed in more detail below.

EPST in the Dutch Context

In the Netherlands, the EPST is used to aid in the transition from primary to secondary education. During this transition, students are divided into several tracks. These tracks prepare students for different types of secondary education. The tracks in the Netherlands are VWO (pre-university education), HAVO (general secondary education) and VMBO (pre-vocational education) (see Figure 1). VMBO consists of another four sub-tracks. Each student receives a recommendation from their teacher on what type of secondary education to pursue. Switching between tracks is possible and may lead to a (partial) correction of the recommendation if the initial placement does not match student ability (Inspectie van het Onderwijs, 2019). However, the initial track placement remains an important factor in students' further education (Timmermans et al., 2013). To prevent under-advising of students by teachers, the EPST may be used to raise the teacher's recommendation, but not lower it.

The EPST is taken at the end of primary school and measures ability in two domains: mathematics and language. The EPST presents each student with a recommendation about what type of secondary education they should pursue. This recommendation can be single or composite. A single recommendation consists of one level (e.g. HAVO recommendation), while a composite recommendation consists of two consecutive levels (e.g. HAVO/VWO recommendation). A pitfall of the EPST is that it takes multiple days to complete, which is very intensive to students. The current multistage design of Cito consists of 140 items spread over three testing days, which means approximately 47 items are presented each day.

Investigation of teacher-based routing is especially fruitful in the context of the EPST, as the current routing stage of this test is relatively long, taking up a whole testing day (see Figure 2A). Furthermore, relying on the teacher's expertise, as in teacher-based routing, is not new in the transition from primary to secondary education in the Netherlands. As stated previously, currently the recommendation of the teacher is even leading in the actual placement of students. For the teacher-based routing, this teacher's recommendation could be used for the module allocation of students in the MST (See Figure 2B). However, we should consider that the secondary track placement based on the teacher's recommendation is not without debate. Many people fear that the teacher could be mistaken (De Regt, 2004; Kamerman & Vasterman, 2015; Lek & Van De Schoot, 2019; *Toetsbesluit PO*, 2014; Visser et al., 2019), which is also why the EPST is used as an independent, secondary source of information to raise the teacher's recommendation if necessary (Visser et al., 2019). The fear that the teacher may be mistaken becomes especially relevant when there is a mismatch between the result of the EPST and the recommendation by the teacher. For the MST, a mismatch between the EPST and the recommendation by the teacher would lead to a suboptimal module allocation at the beginning of the test. As a consequence, less information would be obtained about the ability of the student and the standard error of measurement would increase (see Equation 3 and 4). Potentially, such a suboptimal module allocation might also influence module allocation in further stages of the MST-design. Although students are expected to compensate for any misallocation in the following stage, this may be more or less successful depending on the extent of the mismatch. As the EPST is used to prevent any bias or mistakes from the teacher, it's especially important to minimalize the influence of the teacher in the MST-design with teacher-based routing. This study pays

specific attention to the situation in which the teacher's view does not match students' performance on the EPST.

[Figure 2 near here]

Methods

Data Generation

As a starting point for the data generation, non-public microdata on the EPST from Statistics Netherlands (CBS) were used. With this data, it was investigated to what extent certain EPST- and teacher recommendations were provided and how often (mis)matches between the EPST recommendation and the teacher's recommendation occurred in the Netherlands in 2014/2015. To create a simulated dataset as realistic as possible, the EPST- and teacher's recommendation and their (mis)matches were proportionally simulated using the findings from the CBS data. The CBS data were analyzed using R (R Core Team, 2019) in the remote secured online environment of CBS. The simulations were run and analyzed outside the remote secured environment of CBS using R (version 3.6.1). From the 163,794 students in the dataset, 119 754 were included (See Appendix A for a detailed report of the CBS data and the inclusion criteria).

Using the results pertaining to the number of (mis)matches between teacher's and EPST recommendation, random data were generated according to the following procedure. First, the true, unknown student abilities were simulated. Second, for these generated abilities, ability estimates and corresponding track recommendation were obtained from the EPST. That is, the 'traditional' EPST without teacher-based routing. Third, given the EPST track recommendation, a teacher's recommendation was generated for each student, taking the number of (mis)matches between EPST and teacher's recommendation in the CBS data into account (see Figure 3). These three steps were

repeated 1,000 times to obtain 1,000 datasets. Below, the data generation procedure is explained in more detail.

1. *The Truth*

As can be seen in Square A of Figure 3, n = 10,000 ability values were drawn randomly from a normal distribution ($\theta \sim N(0, 1)$). These values are considered the true abilities of students. These true abilities were matched with their corresponding track placement. As the track placements were based on the true abilities, these track placements were considered the most optimal. As it is unknown which true abilities belong to which optimal track placement, the proportions by which the track placements occur in the CBS data were used to match each ability with the optimal track placement. After ordering the true abilities, these track placements were assigned proportionally to the true abilities. For example, as Square A of Figure 3 illustrates, the lowest 6.8% of the simulated abilities were assigned VMBO-BB as the optimal track placement and the next 10.1% lowest abilities were assigned VMBO-BB/KB as the optimal track placement.

2. *EPST without teacher-based routing*

Using the true latent ability values from step one, it was estimated what the corresponding estimated ability would be for the regular multistage EPST, thus without teacher-based routing (see Figure 3, Square B). The R package ‘mstR’ was used for this purpose (Magis et al., 2018). The model that was used is a 1-2-3 MST design like the current design of Cito, with a 1PL IRT model as basis (see Figure 2A and Figure 3, Square B). As it was assumed that the 1PL IRT model fits the data, there was a high correlation (r = .97) between the estimated and true ability of the previous step. Again, the estimated abilities were assigned a corresponding track placement. As in the previous step, the proportion by which the track placements occur in the CBS data were used to match ability and track

recommendation (see Figure 3, Square B).

3. Teacher's recommendation

Given the EPST track recommendation from step two, the teacher's track recommendation was generated (see Figure 3, Square C). We used the conditional frequencies of the teacher's recommendation given the EPST track recommendations in the CBS data. The probabilities of each teacher's recommendation given the EPST track recommendation were obtained. For each student, given the EPST track recommendation from step two, the teacher's track recommendation was sampled using these conditional probabilities (see Figure 3, square C for the exact probabilities with which the teacher's recommendation was sampled given each EPST track recommendation).

Simulation of the Multistage Test with Teacher-based Routing

Based on the teacher track recommendations generated in the previous step, students were allocated to different modules in the MST. The number of available modules is varied; m = 2, 3, 4, 5, 8. With only two modules, all students with a composite VMBO-GT/HAVO teacher recommendation or lower were for instance allocated to module 1 whereas all students with a HAVO teacher recommendation or higher were allocated to module 2. With 8 available modules, students with each single or composite recommendation were allocated to a different module. Figure 4 shows how students were allocated to each design with a different number of modules in the first stage. The decision on which teacher recommendations led to which module allocations was based on the idea that each module was taken by roughly the same number of students. The difficulty of the modules was adapted to the mean ability level of the subgroup taking this module (see Figure 4 for the exact design specifications). For each MST-design (with m = 2, 3, 4, 5, 8 first stage modules), a separate item bank was generated assuming a 1PL model. Modules

were created such that each contains 47 items with a different mean difficulty, in line with the original EPST test (see Figure 4).

For each student within the 1,000 datasets, response patterns were generated given the true ability and the allocated module using the package mstR (Magis et al., 2018). The second stage of each design was fixed and contained 3 modules, with the difficulty parameters distributed as follows[1]: $\beta_{\text{easier module}} \sim N(-1.1, .28)$, $\beta_{\text{intermediate module}} \sim N(0, 1)$ and $\beta_{\text{difficult module}} \sim N(1.1, .28)$. The maximum Fisher information criterion was used to select one of the three modules in the second stage, after the first stage module was completed (Magis et al., 2017). This consists of selecting the most informative module at the current ability estimate. In the context of the 1PL model, the most informative module matches difficulty with the provisional ability estimate of a student as closely as possible. The provisional ability and the final ability after completion of the MST were estimated using weighted maximum likelihood (WML) (Warm, 1989). When all final ability estimates were obtained, they were assigned a track recommendation in the same way as in the data generation phase.

Analysis

Precision of Ability Estimates

To investigate the overall effect of replacing the regular routing stage by teacher-based routing on the precision of the ability estimates, the mean standard errors of measurement (see Equation 4) from each design using teacher-based routing (m = 2, 3, 4, 5, 8 modules in the first stage) and regular-routing were compared. This was assessed using analysis of variance (ANOVA) with the mean standard error as dependent variables and the design

[1] Based on the mean ability of the lower, middle and higher third of the students.

(regular-routing and teacher-based routing with m = 2, 3, 4, 5, 8 modules in the first stage) as the independent variable. Post-hoc testing was performed using a Bonferroni correction. In addition, standard error plots per teacher-based routing design were obtained for students with different levels of mismatch between their optimal track placement and their teacher's recommendation. These plots were used to assess the differences in standard errors of the ability estimates for students who were misdirected to a module in the first stage of the teacher-based routing designs. To assess the misallocation, students were divided into groups based on the mismatch between the optimal track placement and the teacher's recommendation. To quantify this mismatch, the difference between each single advice was considered one level difference (for example, the difference between a HAVO and a VWO recommendation). The difference between an overlapping single and composite advice was considered to be .5 level (for example, a HAVO/VWO and a VWO recommendation).

Track Recommendations

To investigate the effect of replacing the regular routing stage by teacher-based routing on the track recommendations, the recommendation of each design using teacher-based routing (m = 2, 3, 4, 5, 8 modules in the first stage) was compared with the EPST recommendation without teacher-based routing.

For each of the 1,000 datasets and for each design with a different number of modules in the first stage (m = 2, 3, 4, 5, 8), Cramer's V and Goodman-Kruskal were computed over the regular-routing based EPST placement recommendation and the teacher-based routing EPST recommendation. These statistics are all measures of association. Cramer's V is a measure of association between nominal variables and varies between 0 (no association) and 1 (perfect association). As this measure was used for nominal variables, it only assessed if the recommendation obtained from the regular-

routing and teacher-based routing EPST was exactly the same. Goodman-Kruskal γ, on the other hand, is a measure of association between ordinal variables. Thus, this measure also takes the distance between the recommendations into account (i.e. a VMBO-BB recommendation is closer to VMBO-KB recommendation than a VMBO-BB is to a HAVO recommendation). Goodman-Kruskal γ is based on the number of (dis)concordant pairs and takes the ordering of the recommendation into account. Goodman-Kruskal γ shows the relative difference in concordant and discordant pairs on a scale from -1 (negative association) to 1 (perfect association) where a value of 0 indicates the absence of association (Göktas & Isci, 2011).

The above measures were used to compare how well the introduced teacher-based routing designs managed to provide students with the same recommendation as the regular EPST. It was investigated what number of modules in the first stage (m = 2, 3, 4, 5, 8) of the teacher-based routing EPST leads to the most overlap between track recommendations of the regular-routing EPST. This was assessed using a multivariate analysis of variance (MANOVA) with Cramer's V and Goodman-Kruskal γ as dependent variables and the number of modules in the first stage (m = 2, 3, 4, 5, 8), as the independent variable. Post-hoc testing was performed using a Bonferroni correction.

Second, attention was paid to students who were misdirected to a module when teacher-based routing was implemented. To investigate if these students were less likely to obtain recommendation equal to the optimal track placement, the percentage of students who obtained a recommendation equal to or in the category nearest to the optimal track placement was obtained for each combination of teacher's recommendation and optimal track placement.

Ethical Approval

This study was approved by the Ethics Committee of the Faculty of Social and Behavioral

Sciences of Utrecht University (FETC19-215). Furthermore, results from the CBS data were tested by CBS to prevent publication of identifiable information about individuals[2].

[Figure 3 and 4 near here]

Results

The Precision of Ability Estimates

First, the effect of the shortening of the EPST by using teacher-based routing on the standard errors was assessed. An analysis of variance (ANOVA) was performed over the mean standard error of the 1,000 simulation runs under each design. It was investigated whether there was a difference in the mean standard errors of the regular-routing design and the different teacher-based routing designs (m = 2, 3, 4, 5, 8 modules in the first stage). The result indicated that the test design had a significant effect on the mean standard error of measurement ($F(5, 5994) = 633030$, $p < .001$). Additional post-hoc comparisons with a Bonferroni correction indicated that the mean standard errors of the teacher-based routing designs differed significantly from each other and from the regular-routing design. The ability estimates obtained using a regular-routing design were more precise than those obtained using the teacher-based routing designs (see Table 1). However, as the differences between the teacher-based routing designs and the regular-routing design were small, they may not influence the actual outcome of the test too much. Furthermore, as the regular-routing EPST design included more items than the teacher-based routing design (141 items compared to 94 items in the teacher-based routing designs), it is unclear whether this difference was caused by the influence of the teacher

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

or merely by the decreased number of items. The mean standard errors of the different teacher-based routing designs also differed slightly from each other. For the teacher-based routing designs, the design with 8 modules in the first stage resulted in the lowest mean standard error and the design with 2 modules in the first stage in the highest mean standard error.

[Table 1 near here]

Second, the standard errors of the ability estimates were compared between the different designs for students who were misdirected to a module in the first teacher-based routing stage (see Figure 5). Students who were misdirected to a module in the first stage, because their teacher's recommendation differed 3 or more levels from their optimal track placement (for instance, a teacher who recommends VWO for a student whose optimal track is VMBO-KB), obtained less precise ability estimates when teacher-based routing was implemented compared to the regular-routing estimates. For these students, the teacher-based routing design with 2 modules generally resulted in lower standard errors than the other teacher-based routing designs. Students who were directed to a fitting module in the first stage, because their teacher's recommendation matched their optimal track placement, obtained less precise ability estimates when teacher-based routing was implemented compared to the regular-routing only when the ability estimate was inside the -2 – 2 range. Outside this range, the teacher-based routing appeared to lead to slightly more precise ability estimates, with the most precise estimates obtained using the design with 8 modules in the first stage. Students with a mismatch of 1 – 2.5 levels (for instance, a teacher who recommends VWO (1 level difference) or VMBO-VMBO-BB/KB (2.5 level difference) to a student whose optimal track is HAVO) are sometimes directed to a (close to) optimal module, whereas others are directed to a suboptimal module. This results in some students obtaining the same standard error pattern as students without a

mismatch and some students obtaining the same standard error pattern as students with a severe mismatch.

[Figure 5 near here]

Track Recommendations

As an indicator of the ability of the different designs to provide students with a placement recommendation equal to the regular-routing design, a multivariate analysis of variance (MANOVA) was performed. There was a significant multivariate effect for Cramer's V and Goodman-Kruskal γ in relation to the number of modules used in the design ($F(4, 4995) = 1381$, $p < .001$). Thus, the choice for the number of modules in the first stage mattered in terms of the eventual track placement of students.

Univariate analyses indicated that the number of modules had a significant effect on both Cramer's V ($F(4, 4995) = 1393$, $p < .001$) and Goodman-Kruskal γ ($F(4, 4995) = 887$, $p < .001$). Post-hoc comparisons with a Bonferroni correction indicated that all the teacher-based routing designs differed from each other. The design with 2 modules in the first stage led to the lowest Cramer's V and the design with 8 modules in the first stage led to the highest Cramer's V. This indicates that the design with 8 modules most often resulted in a recommendation exactly equal to the regular-routing EPST recommendation. The design with 4 modules resulted in the lowest Goodman-Kruskal γ and the design with 3 modules to the highest Goodman-Kruskal γ , indicating that the design with 4 modules showed the least amount of overlap between the track recommendations of the EPST with and without teacher-based routing and the design with 3 modules showed the most amount of overlap. Thus, overall, the track recommendations obtained with teacher-based routing with either 3 or 8 modules in the first stage, were most in line with the EPST track recommendation when a regular routing stage was used. However, the differences in

Cramer's V and Goodman-Kruskal γ are small, especially the differences in the Goodman-Kruskal γ statistic. (See Table 2 for an overview of the mean Cramer's V and Goodman-Kruskal γ over the 1,000 simulation runs under each of the different designs). These small differences are also reflected in the percentage of students who obtained a recommendation (approximately) equal to the regular-routing EPST recommendation. When 3 modules were available in the first stage, slightly more students obtained a recommendation (approximately) equal to the regular-routing EPST recommendation compared to the other teacher-based routing designs (see Table 3). When 2 modules were available, the smallest number of students obtained a recommendation (approximately) equal to the regular-routing EPST. A small percentage of students obtained a recommendation that differed 2.5 or more levels from the regular-routing EPST recommendation, but differences between designs were small. Overall, the differences between the different teacher-based routing designs were relatively small.

[Table 2 and 3 near here]

Furthermore, it was investigated how often students obtain an appropriate teacher-based routing MST recommendation dependent on the module (mis)allocation in the first stage. For each combination of the teacher's recommendation and optimal track placement, the proportion of students who were presented with a teacher-based routing MST recommendation (approximately) equal to the optimal placement was obtained. As the teacher-based routing recommendations were now compared to the optimal track placement, the regular-routing EPST recommendations were also compared to the optimal track placement to make a comparison between the teacher-based routing and regular-routing EPST recommendations possible.

Overall, around 93-95% of all students obtained a teacher-based routing track recommendation equal to (or in a category nearest) the optimal track placement,

independent of the teacher's recommendation and the number of modules (see Figure 6A). However, there were differences between subgroups of students with different optimal track placements (see Figure 6B). For most optimal track placements, the proportion of students with this track recommendation (or in a category nearest) was about 94%-100%, with the exception for students with VMBO-KB and VMBO-GT as optimal track placement. For these tracks, only 65 – 73% of students (VMBO-KB) and 78 – 84% of students (VMBO-GT), were provided with a teacher-based routing recommendation equal to (or in a category nearest) the optimal track placement. Especially the teacher-based routing design with 2 modules in the first stage resulted in a relatively low proportion of students with a recommendation equal to (or in the category nearest) the optimal track placement for the aforementioned tracks VMBO-GT and VMBO-KB (see Figure 6B). The regular-routing EPST recommendations were also compared to the optimal track placement. The same pattern as above was present where a lot of students were provided with a track recommendation equal to their optimal placement (97 – 100%) with the exception for VMBO-KB (73%) and VMBO-GT students (82%).

The green squares in Figure 7 (A-E) outline students that were allocated to the most optimal module in the first stage of a teacher-based routing design. These figures show that, overall, misallocation to a module in the first stage did not have much impact on the proportion of students who obtained a correct teacher-based routing recommendation. Especially students with a VMBO-BB or VWO track placement recommendation were likely to obtain a recommendation equal to (or in the category nearest) the optimal track placement, regardless of the module they were allocated to. Again, students with a VMBO-KB or VMBO-GT optimal track placement were the exception. For these students, misallocation in the first stage did impact the probability

of obtaining a recommendation equal to (or in the category nearest) the true optimal track placement. In the teacher-based routing design with 2 modules in the first stage (see Figure 7A), only 52-54% of VMBO-KB students obtained an (approximately) equal recommendation when misallocated in the first stage compared to 64%-68% of students who were directed to the most optimal module. In the design with 8 modules, 74% of VMBO-KB students obtained the correct recommendation when they were allocated to the right module, compared to 53 – 67% of students who were allocated to a module too difficult and 80 – 82% of students allocated to a module too easy (see Figure 7E). Thus, for VMBO-KB students it seems most problematic when the teacher's recommendation is too high, and students were allocated to modules that are too difficult. In contrast, for VMBO-GT students it seems more problematic when the teacher's recommendation is too low, resulting in students being allocated to modules that are too easy for them.

When focusing on students with a mismatch of 3 or more levels, a high proportion of students were provided with the a teacher-based routing track recommendation equal to (or in the nearest category) the optimal track placement. As mentioned, for VMBO-BB and VWO students, the proportion of students who obtained a recommendation equal to the optimal track placements barely differed for students who were or were not misallocated to a module in the first stage. For HAVO/VWO and VMBO-BB/KB students, the proportion of students who obtained a recommendation equal to the optimal track placement was lower for students who were misdirected to a module in the first stage than for students who were directed to an appropriate module. Especially for students with a VWO teacher's track recommendation and a VMBO-KB optimal track placement, the difference was large. Depending on the number of modules available in the teacher-based routing MST design, these students were correctly classified in only 53-57% of the cases. This was much lower than the 89 – 100% of the other students with

a mismatch of 3 of more levels that obtained the same track recommendation when teacher-based routing is used. This is in line with the finding above that students with VMBO-KB as optimal track are overall less likely to obtain an appropriate recommendation, especially when students are misdirected to a module in the first stage. However, one should keep in mind that this is the most extreme mismatch possible which occurs for very few students. As an indication, in 2014/2015 only one student obtained a VMBO-KB EPST recommendation with a VWO teacher's recommendation .

[Figure 6 and 7 near here]

Discussion

The main goal of the current study was to assess the accuracy of ability estimates and the differences in track recommendations of the simulated EPST with and without teacher-based routing. The present study specifically focuses on the accuracy of ability estimates (i.e., their standard errors) and the differences in track recommendations of the EPST with and without teacher-based routing. Particular focus is given to situations in which module allocation based on the teacher's recommendation is not in line with student's performance.

With respect to the standard errors, two effects were found. First, for the teacher-based routing designs, there were slight differences between the mean standard errors. This mean standard error was lowest for the design with 8 modules and highest for the design with 2 modules. This difference may be explained by the increase in information about the ability of students when the teacher's recommendation matches students' ability (Eggen & Sanders, 1993; Magis et al., 2017). For most students, the teacher's recommendation matches their ability, resulting in a (near) optimal module allocation if teacher-based routing is implemented. When more modules are available, the difficulty

of these modules will match student ability even closer than when fewer modules are available. This results in an increase in obtained information and in lower standard errors.

Second, it was found that the mean standard error was lowest when a regular-routing design was used. However, this difference is very slight whereas far less items were used in the teacher-based routing EPST compared to the regular-routing EPST (94 items compared to 94 items). Although the gain in information is not enough to make up for the loss of information by shortening the test, this is a positive sign that teacher-based routing could be an effective way to shorten the regular EPST, albeit not with a third of the original test length. Further analysis suggested that both test length and the influence of the teacher's recommendation play a role in the reduced precision of the teacher-based routing EPST. Independent of the (mis)allocation to modules in the first stage, the regular-routing design outperforms the teacher-based routing designs when the estimated ability lies within the -2 – 2 ability range. This suggests that the number of items plays a role in the accuracy of the EPST. However, outside the -2 – 2 ability range, the (mis)allocation by the teacher plays a role. Students who are allocated to the appropriate module are provided with ability estimates that are as precise, or even more precise than the regular-routing design. Students who are allocated to a module that slightly misrepresents their true ability are provided with ability estimates equal to those of the regular-routing design. However, students who are allocated to a module that severely misrepresents their true ability obtain less precise estimates compared to the regular-routing design. Thus, the teacher-based routing design is disadvantageous for students whose teacher's recommendation is not in line with their true performance. As most students have a true ability around 0, it logically follows that most students are provided with the most precise estimate when regular routing is used. However, as a main goal of

the EPST is to provide students with a track recommendation, it is important to understand how these differences influence the track recommendations.

With respect to the track recommendations, results indicate that the choice for the number of modules in the first stage matters in terms of the eventual track placement of students, although differences were small. Depending on the statistic used, the designs with 2 and 4 modules showed the least amount of overlap between track recommendations of the EPST with and without teacher-based routing. The designs with 3 and 8 modules in the first stage showed the most amount of overlap between the track recommendations of the EPST with and without teacher-based routing. Thus, the track recommendations obtained with teacher-based routing with 3 and 8 modules in the first stage were most in line with the regular-routing EPST track recommendation.

There is a small subset of students for whom the information from the teacher does not match their ability and these students may be directed to modules that are too difficult or too easy. Although students are expected to compensate for any misallocation in the following stage, this may be more or less successful depending on the extent of the misallocation. As such a mismatch occurs for only a few students, they may not be properly represented in an overall statistic describing the overlap between the optimal track placement and the teacher-based routing track recommendations. Nevertheless, it is important that such students are not systematically presented with a track recommendation that is influenced by the teacher. Overall, still a high proportion of these students were provided with the same teacher-based routing EPST recommendation as the optimal track placement. Nevertheless, this proportion is lower than for students who were allocated to a (near) optimal module. Especially for students with a VWO teacher's track recommendation and a VMBO-KB optimal track placement, the difference was large. However, this could be explained by the observation that students whose optimal

track placement is VMBO-KB (or VMBO-GT) are overall less likely to obtain the same track recommendation in the teacher-based routing EPST. Especially for these students, the teacher's recommendation seems to have a strong effect on the eventual track recommendation.

Limitations and Future Research

The current study included a restricted set of conditions. First, the item pool in the current study was simulated following a 1PL model where the mean and the variation of the item difficulty in a module targeted the mean and the variation in ability of the group expected to take that module. More variation of the item difficulty or the use of a different, more complex IRT model could provide different results. The EPST, for example, uses the more complex OPLM model that also includes a parameter that describes how well an item can discriminate between students (CvTE, 2015). This parameter weighs the function, which can result in a different test information function (Eggen & Sanders, 1993; Van der Linden & Glas, 2010). Furthermore, more variation of the item difficulty within modules may be advantageous for students who are misdirected in the first stage. It would be interesting to investigate the relationship between the variation of the difficulty within modules on the ability estimates of students who are misdirected in the first stage.

Second, the teacher-based routing used the teacher's recommendation to direct students to modules in the first stage. However, teachers may be more certain about their assessment for some students than for others. It could be useful to not only assess the teacher's recommendation, but also their (un)certainty. For example, if a teacher is uncertain about their judgement of a student, this student could be allocated to a module representing the current routing stage. Thus, preventing students from being directed to the wrong module. One way to assess this uncertainty is through the use of expert

elicitation, which is the process of formalizing experts (teachers) knowledge in a way that can be used statistically (O'Hagan et al, 2006). Specifically, an online application to elicit teacher's knowledge has been developed (Lek & Van De Schoot, 2018). It would be interesting to assess the use of this online application with the introduced teacher-based routing design. However, more research is needed to investigate if the uncertainty of a teacher does indeed relate to a less accurate estimate by the teacher.

Third, in contrast to the previous study by Berger et al. (2019), the current study did not add teacher-based routing to a design but attempted to replace the first stage by the teacher-based routing. Perhaps the current study would also have found an advantage of teacher-based routing over regular routing when the number of stages was kept equal. However, the current study shows that the current design is already quite capable of providing students with accurate ability estimates and track recommendations. A possible addition to the current design would be to add a third stage, but only for students whose standard error is large and/or whose ability estimate confidence interval covers several recommendation possibilities. This would likely ensure that a majority of the students only have two testing days, whereas students who were misdirected get the opportunity to compensate for this misdirection. It would be worthwhile to investigate if a third stage allows students to compensate for any misdirection in the first stage and if perhaps some students whose teacher's recommendation is in line with their ability could even stop after the first stage. However, besides research on its feasibility, research on the desirability of a teacher-based routing MST with different test lengths should also be investigated.

Conclusion and practical implications

Teacher-based routing is a method that could replace the traditional routing stage of an MST design. With this simulation study, we extended previous research on the effect of replacing the traditional routing stage with pre-existing knowledge. In addition to the

effect of the teacher-based routing MST on standard errors and the track recommendation that is provided, insight is gained on the effect of teacher-based routing on the precision of ability estimates and the track recommendation of students who are misdirected to a module in the first stage. The results indicated that the teacher-based routing EPST generally resulted in slightly bigger mean standard errors than the longer, regular-routing design. Only students with more extreme ability estimates who were allocated to a (near) optimal module were not disadvantaged by, or even benefitted from, the teacher-based routing. Students who were misdirected in the first stage were overall disadvantaged by the teacher-based routing. Differences between the different teacher-based routing designs show that more information is obtained about the ability of students when more modules are available in the first stage. Furthermore, the results indicate that most students obtain a track recommendation from the teacher-based routing design that is (approximately) equal to the true optimal track placements of students. Students who were misdirected in the first stage slightly less often obtain an (approximately) equal track recommendation than students who were directed to a fitting module. Slightly more students were provided with a fitting track recommendation in the design with 3 modules in the first stages than designs with fewer modules.

In conclusion, the gain of information by assigning students to a module that matches their ability does not completely make up for the loss of information by shortening the test. Nevertheless, if a teacher-based routing would be implemented in real-life scenarios, a design with 3 modules in the first stage may be slightly more advantageous. As the teacher-based routing design contains as much modules as the traditional EPST, the same number of items need to be developed. Thus, the expected investments are expected to be about equal. Likewise, it is expected that students who are misdirected in the first stage are better able to compensate for this misallocation when

fewer modules are available and the assigned modules cover a broader range of difficulty parameters. Although results show that students in general, are able to obtain the correct recommendation, this is not necessarily true for all students. Therefore, in its current form it is advised to only use the teacher-based routing MST design for low-stakes testing rather than in high-stakes testing such as the Dutch EPST. However, with further research, the teacher-based routing MST might be implementable in high-stakes situations as well. In hindsight, removing the complete first stage (one third of the total number of items) was too strict. Given the small difference between precision of ability estimates and the track recommendations of the regular and teacher-based routing designs, it is possible, for instance, that teacher-based routing could lead to a quarter instead of a third reduction in items without loss of information. Or, that with the addition of a few extra items only for those whose standard error is still above a certain threshold teacher-based routing and the regular MST lead to the same degree of accuracy. This would likely also diminish the influence of the teacher for students who are misdirected to a module in the first stage.

Acknowledgments

I would like to thank my supervisors Kimberley Lek, Rens van de Schoot, Dave Hessen and Remco Feskens for their guidance throughout my master thesis project. My specific thanks go out to Kimberley for her valuable feedback on earlier versions of this article. Moreover, I would also like to thank Daniella Cianci and Maria Schipper for their support and feedback as mentor, as well as my fellow students for their constructive feedback. Lastly, I would like to thank Gabe Naylor-Leyland for his feedback and Veerle Brouwer for both her feedback and moral support.

Declaration of interest statement

The author did not report any conflicts of interest in relation to the work described.

References

Armstrong, R. D., Jones, D. H., Koppel, N., B., & Pashley, P. J. (2004). Computerized Adaptive Testing with Multiple-Form Structures. *Applied Psychologial Measurement*, *28*(3), 147–164. https://doi.org/10.1177/0146621604263652

Asseburg, R., & Frey, A. (2013). Too hard, too easy or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modelling*, *55*(1), 92–104.

Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Improvement of Measurement Efficiency in Multistage Tests by Targeted Assignment. *Frontiers in Education*, *4*, 1. https://doi.org/10.3389/educ.2019.0001

Bock, R. D., & Zimowski, M. F. (1998). *Feasibility studies of twostage testing in large scale educational assessment: Implications for NAEP*. NAEP Validity Studies Panel, american Institute for Research in the Behavioral Sciences.

De Regt, A. (2004). Welkom in de ratrace; over de dwang van de Cito-toets. *Amsterdams Sociologisch Tijdschrift*, *31*(3), 297–320.

Documentatie Kenmerken van deelnemers aan de Eindtoets Basisonderwijs van Cito (CITOTAB). (2019). Centraal Bureau voor de Statistiek (CBS). https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/citotab-kenmerken-deelnemers-eindtoets-basisonderwijs

Documentatie Kenmerken van inschrijvingen in diverse onderwijssoorten (ONDERWIJSINSCHRTAB). (2019). Centraal Bureau voor de Statistiek (CBS). https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/onderwijsinschrtab-kenmerken-onderwijsinschrijvingen

Eggen, T. J. H. M., & Sanders, P. F. (1993). *Psychometrie in de Praktijk*. Cito Instituut voor Toetsontwikkeling.

Göktas, A., & Isci, Ö. (2011). A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation. *Methodolski Zvezki*, *8*(1), 17–37.

Harris, D. (1989). NCME Instructional Module: Comparison of 1-, 2-, and 3-Parameter IRT Models. *Instructional Topics in Educational Measurement, 8(1)*.

Inspectie van het Onderwijs. (2019). *De staat van het primair onderwijs 2019*. Ministerie van Onderwijs, Cultuur en Wetenschap. https://www.onderwijsinspectie.nl/documenten/rapporten/2019/04/10/rapport-de-staat-van-het-onderwijs-2019

Kamerman, S., & Vasterman, J. (2015). De leerkracht weet het vaak écht beter dan de Citotoets. *NRC.Next*. http://tinyurl.com/y549mbxe

Lek, K., & Van De Schoot, R. (2018). Development and Evaluation of a Digital Expert Elicitation Method Aimed at fostering Elementary School Teachers' Diagnostic Competence. *Frontiers in Education*, *3*, 1–14. https://doi.org/10.3389/feduc.2018.00082

Lek, K., & Van De Schoot, R. (2019). Wie weet het beter, de docent of de centrale eindtoets? *De Psycholoog*, *54*(4).

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A Testlet Assembly Design for Adaptive Multistage Tests. *Applied Measurement in Education*, *19*(3), 189–202. https://doi.org/10.1207/s15324818ame1903_2

Luecht, R. M., & Nungester, R. J. (1998). Some Practical Examples of computer Adaptive Sequential Testing. *Journal of Educational Measurement*, 229–249. https://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Magis, D, Duanli, Y., & von Davier, A., A. (2017). *Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR*. Springer.

Magis, D, Yan, D., & von Davier, A., A. (2018). *MstR: Procedures to Generate Patterns under Multistage Testing* (Version 1.2) [Computer software].

Mead, A., D. (2006). *An Introduction to Multistage Testing*. *19*(3), 185–187. https://doi.org/10.1207/s15324818ame1903_1

OECD. (2019). *PISA 2018 Results (Volume I)*. https://www.oecd-ilibrary.org/content/publication/5f07c754-en

O'Hagan, A., & al, et. (2006). *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org

Staatssecretaris van Onderwijs, Cultuur en Wetenschap. (2014). *Toetsbesluit PO*. https://www.rijksoverheid.nl/documenten/besluiten/2014/01/20/toetsbesluit-po.

Timmermans, A. C., Kuyper, H., & Van Der Werf, G. (2013). *Schooladviezen en onderwijsloopbanen: Voorkomen, risicofactoren en gevolgen van onder- en overadvisering*. Gronings Instituut voor Onderzoek van het Onderwijs (GION).

Van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of Adaptive Testing* (1st ed.). Springer-Verlag.

Veldkamp, B., P., & Matteucci, M. (2013). Bayesian Computerized Adaptive Testing. *Ensaio Avaliação e Políticas Públicas Em Educaçaào*, *21*(78), 57–82. https://doi.org/DOI: 10.1590/S0104-40362013005000001

Verantwoording Centrale Eindtoets PO. (2015). College voor Toetsen en Examens (CvTE). https://www.cvte.nl/documenten/publicaties/2015/05/12/verantwoording-centrale-eindtoets-po

Verhuls, N. D. (1989). Informatiewinst bij vertakt toetsen. In Van der Linden & Van der Kamp, *Meetmethoden & data-analyse*. Instituut voor Toetsonwikkeling (Cito).

Visser, D., van den Berge, W., & Visser, D. (2019). *Policy Brief. De waarde van eindtoetsen in het po.* Centraal Plan Bureau (CPB). https://www.cpb.nl/eindtoets-draagt-bij-aan-beter-passend-schooladvies

Wainer, H., & Wang, X. (2000). Using a New Statistical Model for testlets to Score TOEFL. *Journal of Educational Measurement*, *37*(3), 203–220. https://doi.org/DOI: 10.1111/j.1745-3984.2000.tb01083.x

Wang, S., Rubie_Davies, C. M., & Meissel, K. (2018). A Systematic Review of the Teacher Expectation Literature over the past 30 Years. *Educational Research and Evaluation*, 124–179. https://doi.org/10.1080/13803611.2018.1548798

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory with Tests of Finite Length. *Psychometrika*, *54*(427–450). https://doi.org/10.1007/BF02294627

Wet op het voortgezet onderwijs (WVO), Pub. L. No. BWBR0002399. https://wetten.overheid.nl/jci1.3:c:BWBR0002399&z=2020-04-01&g=2020-04-01

Yamamoto, K., Shin, H., J., & Khorramdel, L. (2009). *Introduction of multistage adaptive testing design in PISA 2018* (Vol. 209). OECD.

Appendix A

As a starting point for the data generating, non-public longitudinal microdata from Statistics Netherlands (CBS) were used. In addition to creating the contingency table, several analyses were performed using placement in the first and third year of secondary education too. For this reason, not only data containing information about the EPST and teacher's recommendation of students in the year 2014/2015 (the Citotab file, see (*Documentatie Kenmerken van Deelnemers Aan de Eindtoets Basisonderwijs van Cito (CITOTAB).*, 2019), but also data containing information about the placement in secondary education in the year 2015/2016 and 2017/2018 was used (The Onderwijsinschrtab, see (*Documentatie Kenmerken van Inschrijvingen in Diverse Onderwijssoorten (ONDERWIJSINSCHRTAB)*, 2019))[3]. In the CBS files, the VMBO-GL and VMBO-TL tracks were combined and referenced to as VMBO-GT.

The Citotab file includes information about students whose primary school authorized Cito to share their results with CBS. Only the students who met the following inclusion criteria were included in this study: (1) students are registered in the GBA (Administration of the Municipality); (2) both the teacher's recommendation and the EPST score are available; (3) the placement in the first and third year after completing primary school is known and (4) students are not enrolled in practical education in the first or third year of secondary education[4]. In total, 44,040 students were excluded. After selection, 119,754 from the 163,794 students were part of the analysis (see Figure 8).

[Figure 8 near here]

[3] For information on accessing these files, see: https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/export-van-gegevens

[4] This criterion was used because the lowest category in the EPST is VMBO-BB. Practical education is for students who need a ortho pedagogic or ortho didactic approach (Wet Op Het Voortgezet Onderwijs (WVO), n.d.).

Table 1. mean standard error over the 1000 simulation runs under each of the different designs.

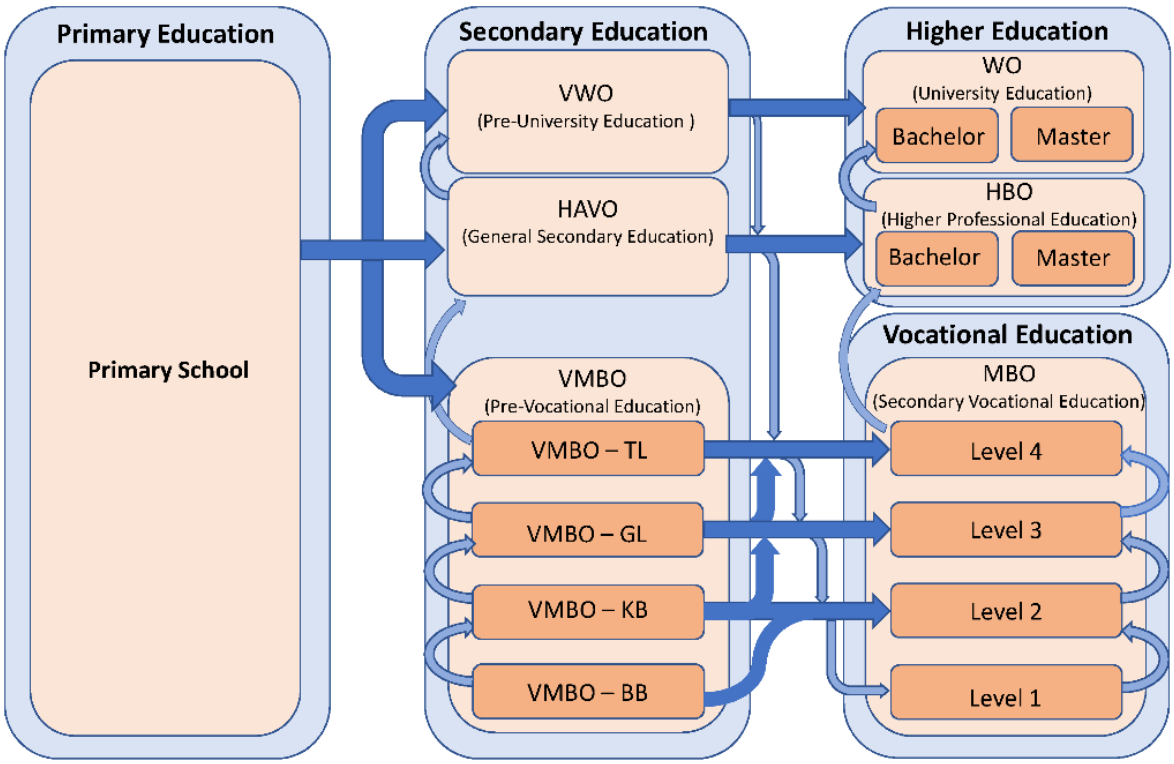
Design	Standard error Mean *(SE)*
Teacher-based routing with $m = 2$	.2185 (.00001)
Teacher-based routing with $m = 3$	.2169 (.00001)
Teacher-based routing with $m = 4$	.2162 (.00001)
Teacher-based routing with $m = 5$	.2158 (.00001)
Teacher-based routing with $m = 8$	.2155 (.00001)
Regular routing stage	.1909 (.00001)

Table 2. mean Cramer's V and Goodman-Kruskal γ over the 1000 simulation runs under each of the different designs.

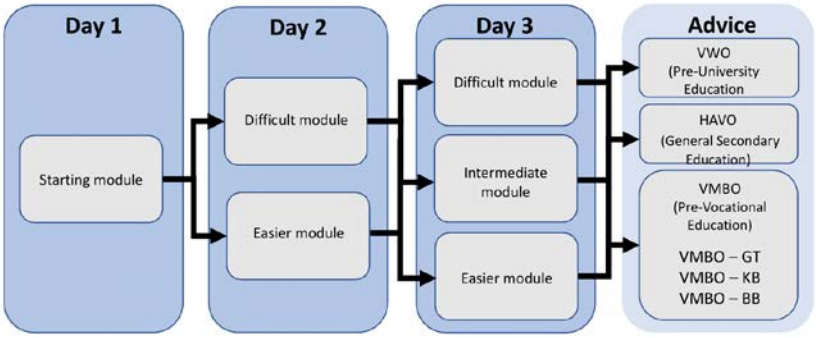
	Cramer's V Mean *(SE)*	Goodman_Kruslkal γ Mean *(SE)*
m = 2	.5753 (*.0002*)	.9439 (*.00007*)
m = 3	.5836 (*.0002*)	.9461 (*.00007*)
m = 4	.5774 (*.0002*)	.9415 (*.00007*)
m = 5	.5800 (*.0002*)	.9421 (*.00007*)
m = 8	.5901 (*.0002*)	.9456 *(.00007)*

Table 3. Difference between regular-routing EPST recommendation and EPST recommendation with teacher-based routing with m = 2, 3, 4, 5, 8 modules in the first stage.

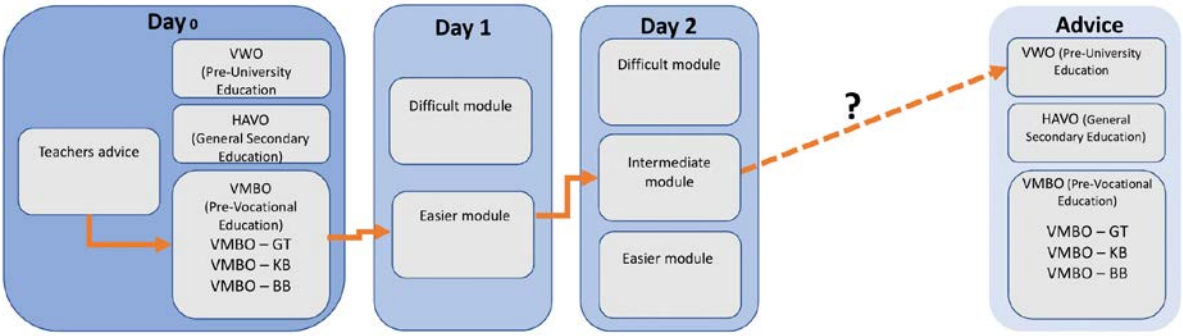
Difference with regular-routing EPST recommendation	2 modules	3 modules	4 modules	5 modules	8 modules
0-.5 levels difference	90.24%	91.21%	90.44 %	90.50%	90.72%
1-2 levels difference	9.75%	8.78%	9.55%	9.49%	9.27%
2.5 or more levels difference	.01%	.01%	. 01%	. 01%	. 01%


Primary Education
Primary School
Secondary Education
VWO
(Pre-University Education)
HAVO
(General Secondary Education)
VMBO
(Pre-Vocational Education)
VMBO – TL
VMBO – GL
VMBO – KB
VMBO – BB
Higher Education
WO
(University Education)
Bachelor
Master
HBO
(Higher Professional Education)
Bachelor
Master
Vocational Education
MBO
(Secondary Vocational Education)
Level 4
Level 3
Level 2
Level 1

[1]

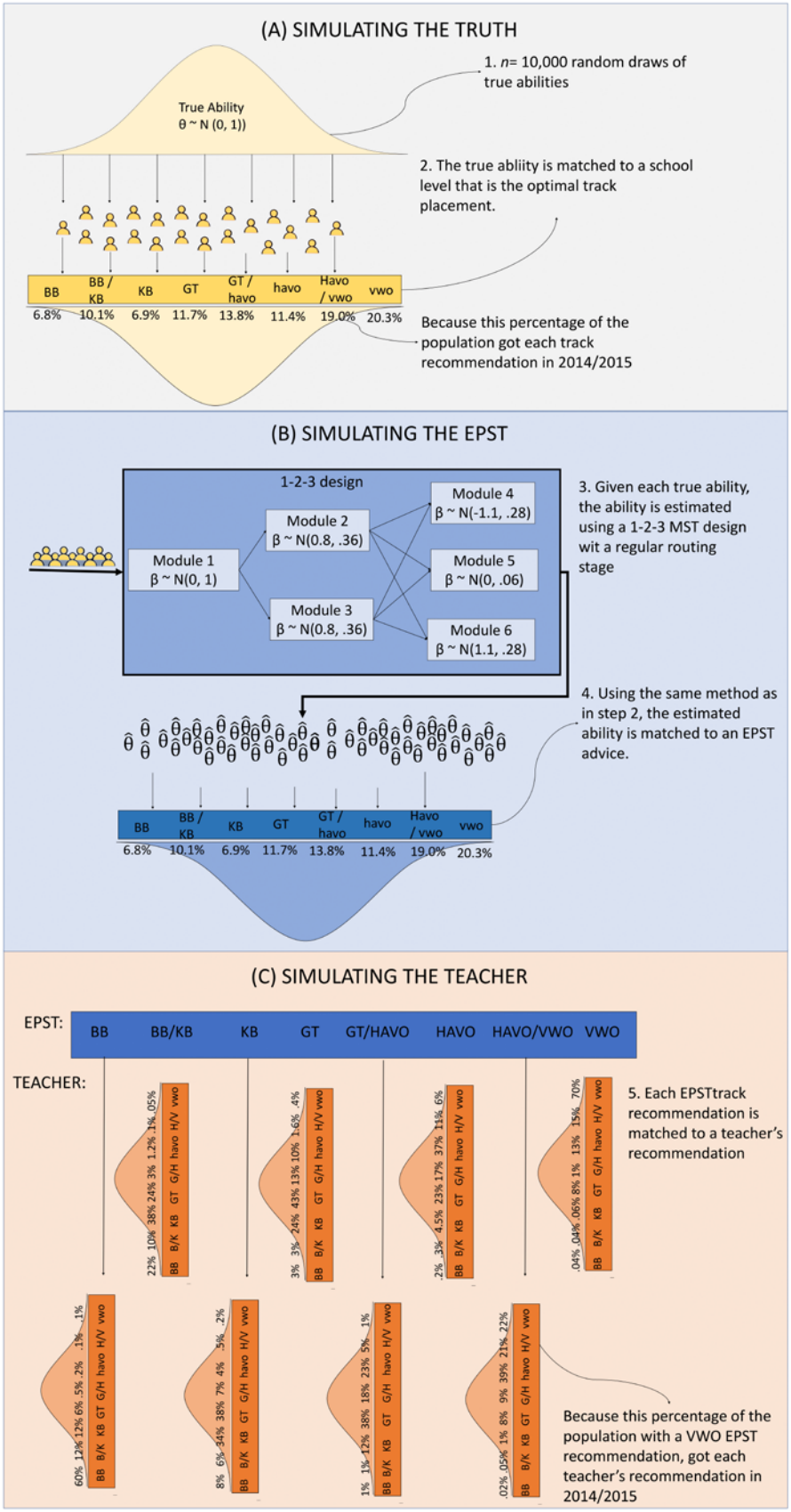

Day 1
Starting module
Day 2
Difficult module
Easier module
Day 3
Difficult module
Intermediate module
Easier module
Advice
VWO (Pre-University Education
HAVO (General Secondary Education)
VMBO (Pre-Vocational Education)
VMBO – GT
VMBO – KB
VMBO – BB

(A)

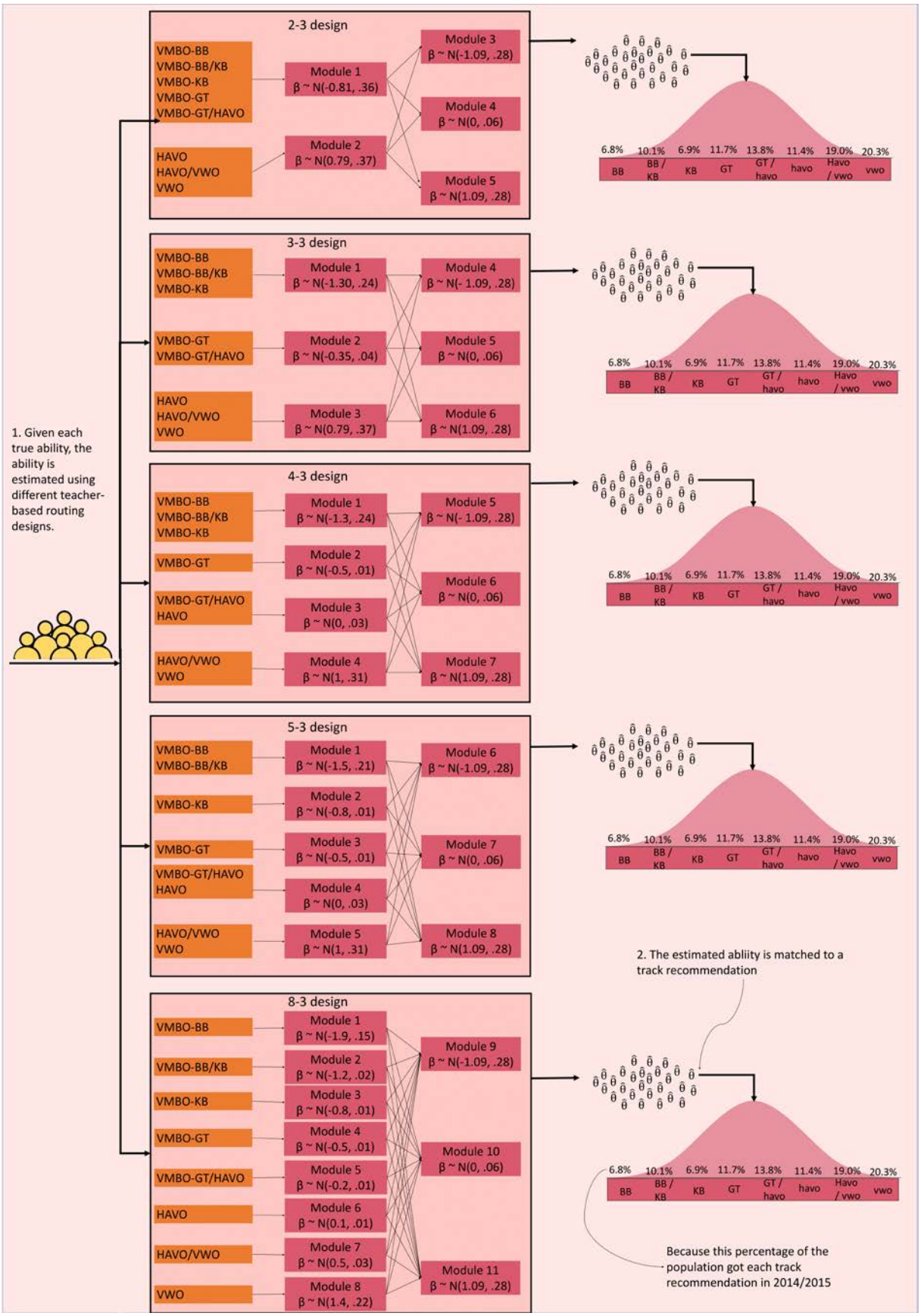

Day 0
Teachers advice
VWO (Pre-University Education
HAVO (General Secondary Education)
VMBO (Pre-Vocational Education)
VMBO – GT
VMBO – KB
VMBO – BB
Day 1
Difficult module
Easier module
Day 2
Difficult module
Intermediate module
Easier module
?
Advice
VWO (Pre-University Education
HAVO (General Secondary Education)
VMBO (Pre-Vocational Education)
VMBO – GT
VMBO – KB
VMBO – BB

(B)

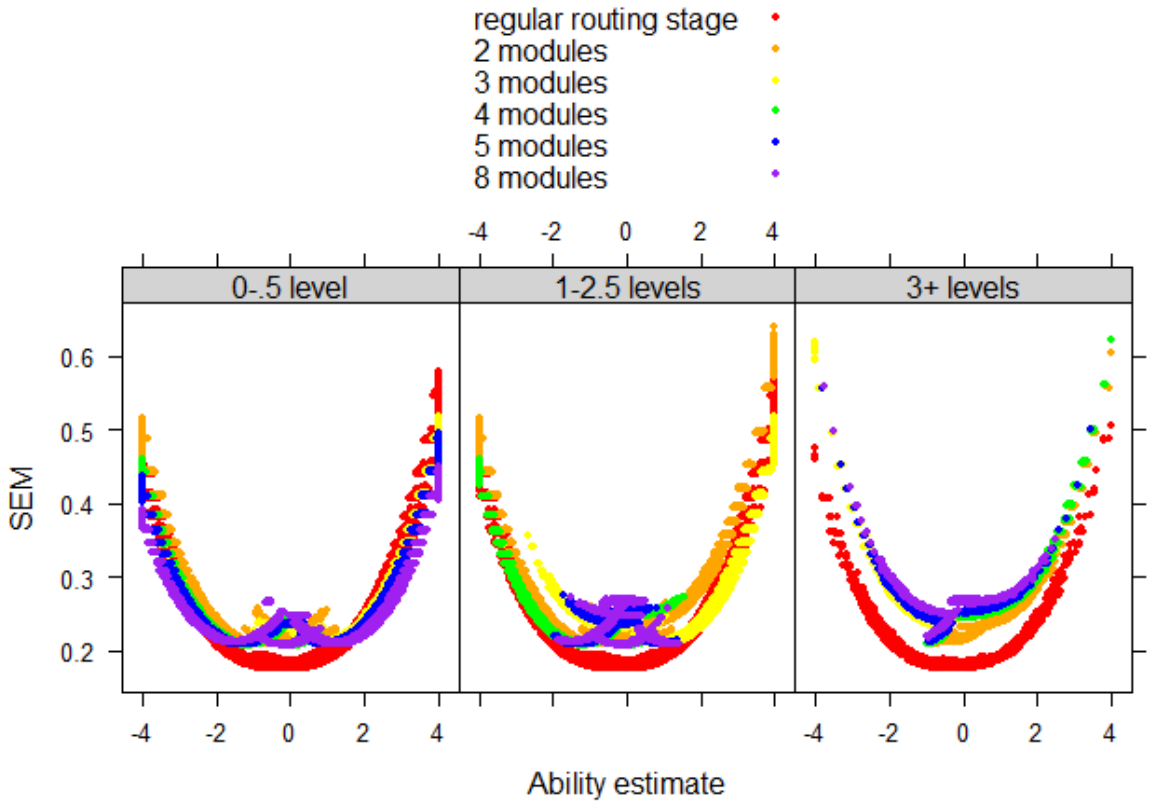
[2]


(A) SIMULATING THE TRUTH
1. n= 10,000 random draws of true abilities
True Ability
θ ~ N (0, 1))
2. The true abliity is matched to a school level that is the optimal track placement.
BB
BB / KB
KB
GT
GT / havo
havo
Havo / vwo
vwo
6.8%
10.1%
6.9%
11.7%
13.8%
11.4%
19.0%
20.3%
Because this percentage of the population got each track recommendation in 2014/2015
(B) SIMULATING THE EPST
1-2-3 design
Module 1
β ~ N(0, 1)
Module 2
β ~ N(0.8, .36)
Module 3
β ~ N(0.8, .36)
Module 4
β ~ N(-1.1, .28)
Module 5
β ~ N(0, .06)
Module 6
β ~ N(1.1, .28)
3. Given each true ability, the ability is estimated using a 1-2-3 MST design wit a regular routing stage
4. Using the same method as in step 2, the estimated ability is matched to an EPST advice.
BB
BB / KB
KB
GT
GT / havo
havo
Havo / vwo
vwo
6.8%
10.1%
6.9%
11.7%
13.8%
11.4%
19.0%
20.3%
(C) SIMULATING THE TEACHER
EPST:
BB
BB/KB
KB
GT
GT/HAVO
HAVO
HAVO/VWO
VWO
TEACHER:
5. Each EPSTtrack recommendation is matched to a teacher's recommendation
Because this percentage of the population with a VWO EPST recommendation, got each teacher's recommendation in 2014/2015

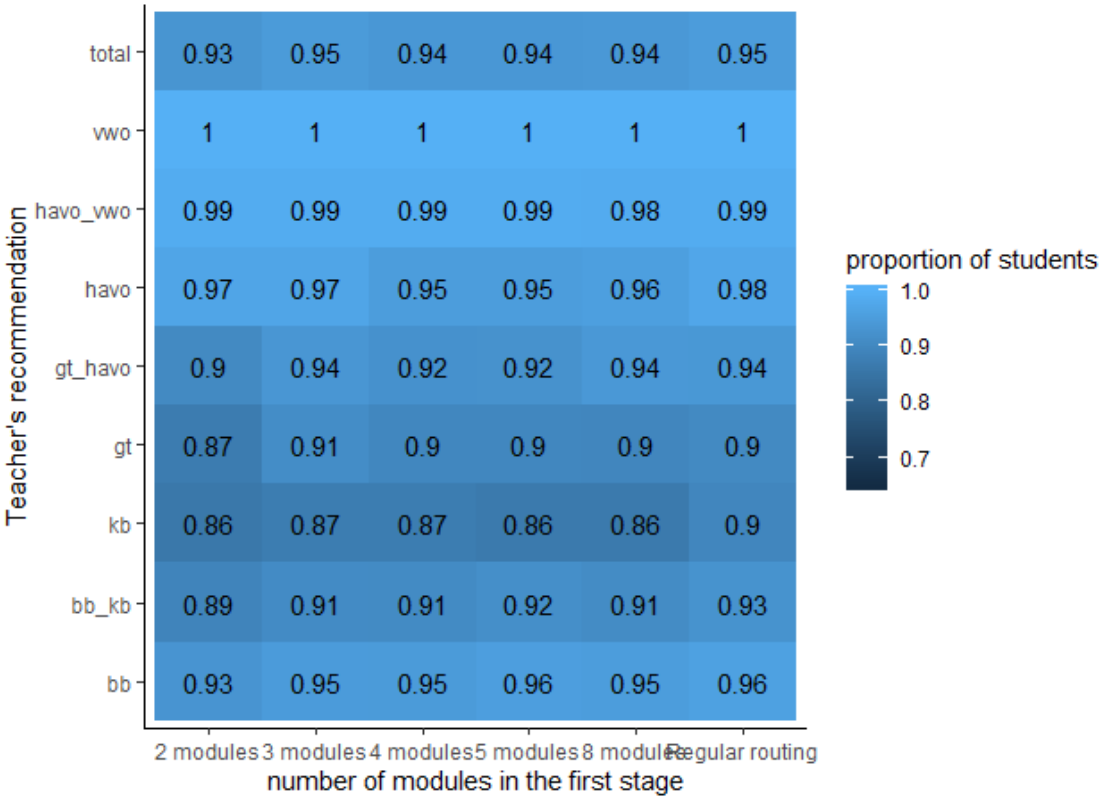
[3]


1. Given each true ability, the ability is estimated using different teacher-based routing designs.
2-3 design
VMBO-BB
VMBO-BB/KB
VMBO-KB
VMBO-GT
VMBO-GT/HAVO
HAVO
HAVO/VWO
VWO
Module 1 β ~ N(-0.81, .36)
Module 2 β ~ N(0.79, .37)
Module 3 β ~ N(-1.09, .28)
Module 4 β ~ N(0, .06)
Module 5 β ~ N(1.09, .28)
3-3 design
Module 1 β ~ N(-1.30, .24)
Module 2 β ~ N(-0.35, .04)
Module 3 β ~ N(0.79, .37)
Module 4 β ~ N(- 1.09, .28)
Module 5 β ~ N(0, .06)
Module 6 β ~ N(1.09, .28)
4-3 design
Module 1 β ~ N(-1.3, .24)
Module 2 β ~ N(-0.5, .01)
Module 3 β ~ N(0, .03)
Module 4 β ~ N(1, .31)
Module 5 β ~ N(- 1.09, .28)
Module 6 β ~ N(0, .06)
Module 7 β ~ N(1.09, .28)
5-3 design
Module 1 β ~ N(-1.5, .21)
Module 2 β ~ N(-0.8, .01)
Module 3 β ~ N(-0.5, .01)
Module 4 β ~ N(0, .03)
Module 5 β ~ N(1, .31)
Module 6 β ~ N(-1.09, .28)
Module 7 β ~ N(0, .06)
Module 8 β ~ N(1.09, .28)
8-3 design
Module 1 β ~ N(-1.9, .15)
Module 2 β ~ N(-1.2, .02)
Module 3 β ~ N(-0.8, .01)
Module 4 β ~ N(-0.5, .01)
Module 5 β ~ N(-0.2, .01)
Module 6 β ~ N(0.1, .01)
Module 7 β ~ N(0.5, .03)
Module 8 β ~ N(1.4, .22)
Module 9 β ~ N(-1.09, .28)
Module 10 β ~ N(0, .06)
Module 11 β ~ N(1.09, .28)
6.8% 10.1% 6.9% 11.7% 13.8% 11.4% 19.0% 20.3%
BB BB / KB KB GT GT / havo havo Havo / vwo vwo
2. The estimated abliity is matched to a track recommendation
Because this percentage of the population got each track recommendation in 2014/2015

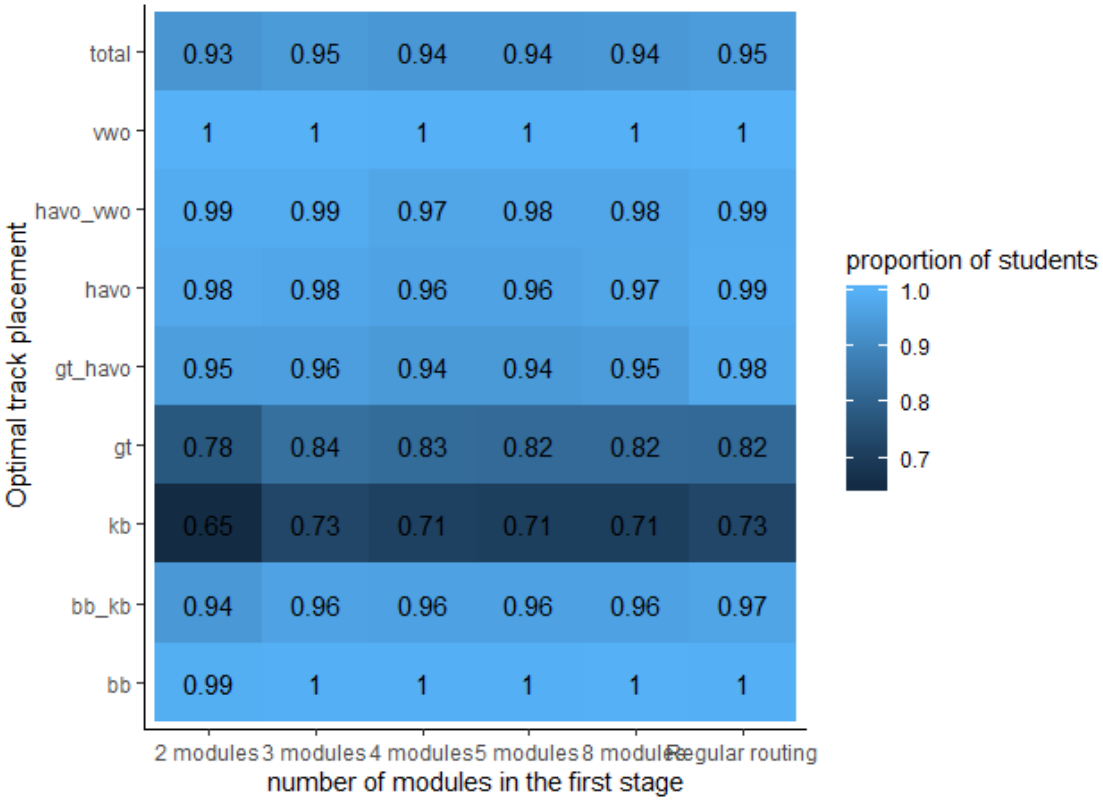
[4]


regular routing stage
2 modules
3 modules
4 modules
5 modules
8 modules
0-.5 level
1-2.5 levels
3+ levels
SEM
Ability estimate

[5]

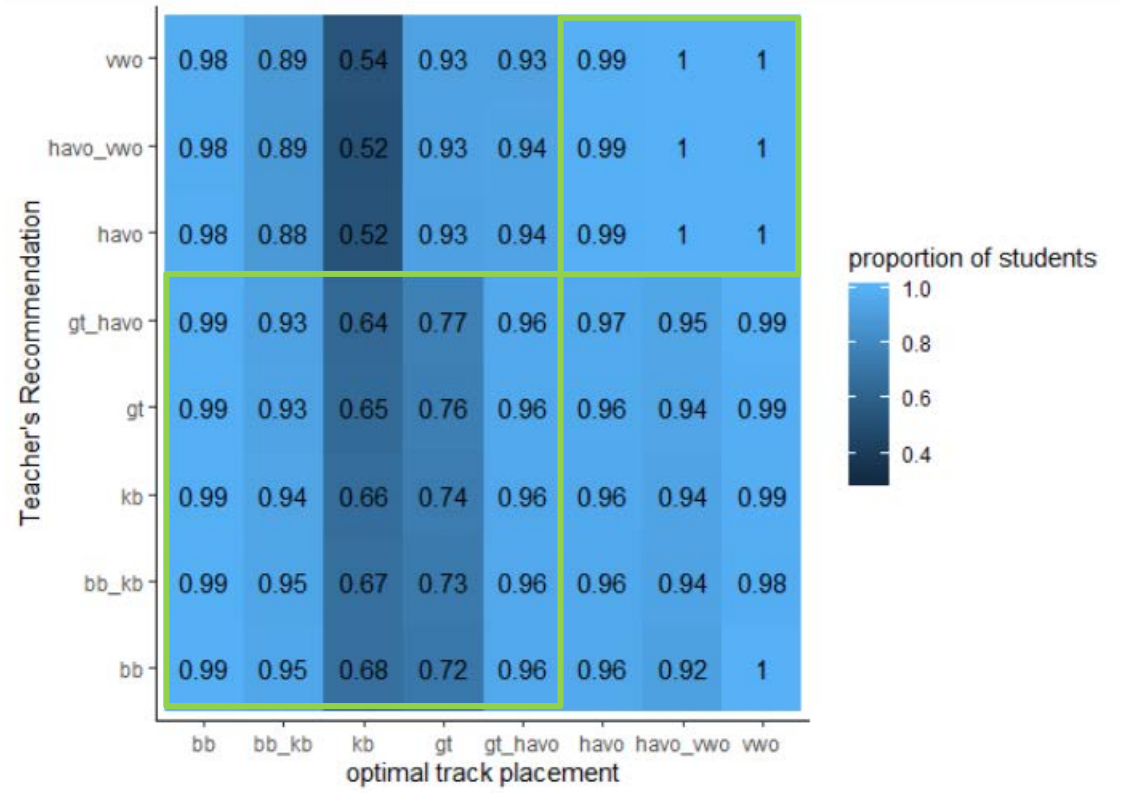

Teacher's recommendation
total
vwo
havo_vwo
havo
gt_havo
gt
kb
bb_kb
bb
0.93 0.95 0.94 0.94 0.94 0.95
1 1 1 1 1 1
0.99 0.99 0.99 0.99 0.98 0.99
0.97 0.97 0.95 0.95 0.96 0.98
0.9 0.94 0.92 0.92 0.94 0.94
0.87 0.91 0.9 0.9 0.9 0.9
0.86 0.87 0.87 0.86 0.86 0.9
0.89 0.91 0.91 0.92 0.91 0.93
0.93 0.95 0.95 0.96 0.95 0.96
2 modules
3 modules
4 modules
5 modules
8 modules
Regular routing
number of modules in the first stage
proportion of students
1.0
0.9
0.8
0.7

(A)

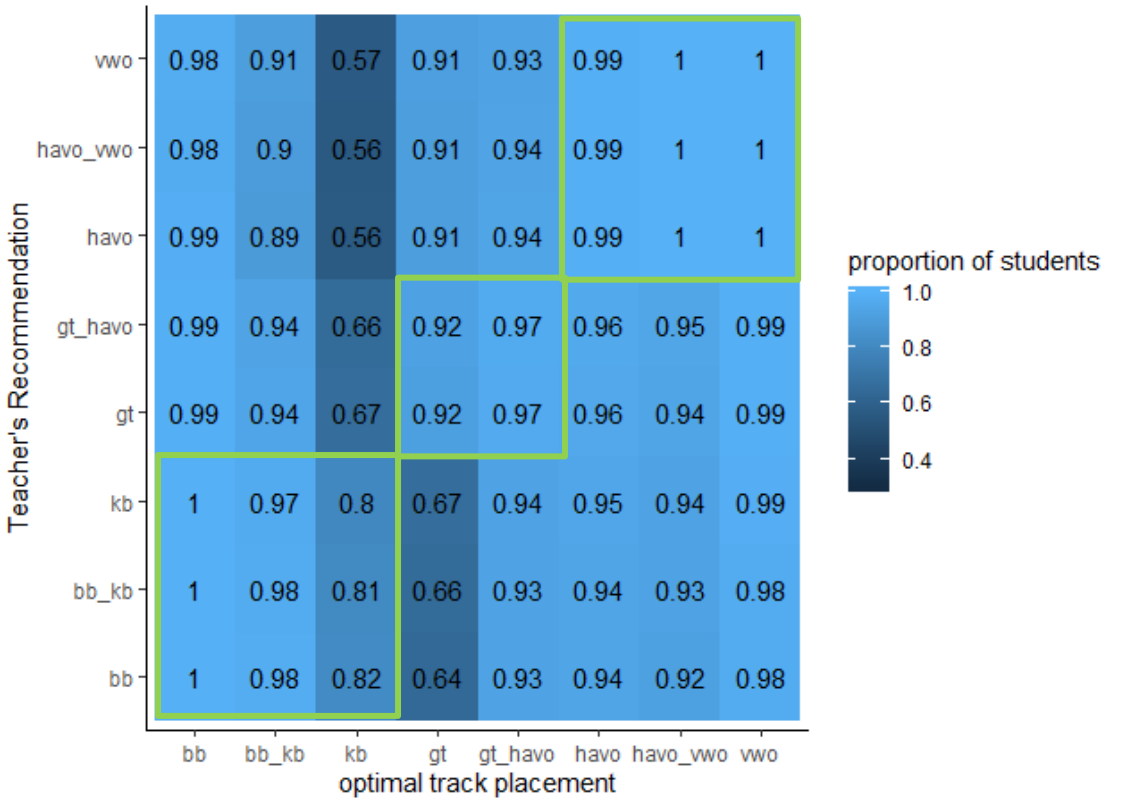

Optimal track placement
total
vwo
havo_vwo
havo
gt_havo
gt
kb
bb_kb
bb
0.93 0.95 0.94 0.94 0.94 0.95
1 1 1 1 1 1
0.99 0.99 0.97 0.98 0.98 0.99
0.98 0.98 0.96 0.96 0.97 0.99
0.95 0.96 0.94 0.94 0.95 0.98
0.78 0.84 0.83 0.82 0.82 0.82
0.65 0.73 0.71 0.71 0.71 0.73
0.94 0.96 0.96 0.96 0.96 0.97
0.99 1 1 1 1 1
2 modules
3 modules
4 modules
5 modules
8 modules
Regular routing
number of modules in the first stage
proportion of students
1.0
0.9
0.8
0.7

(B)

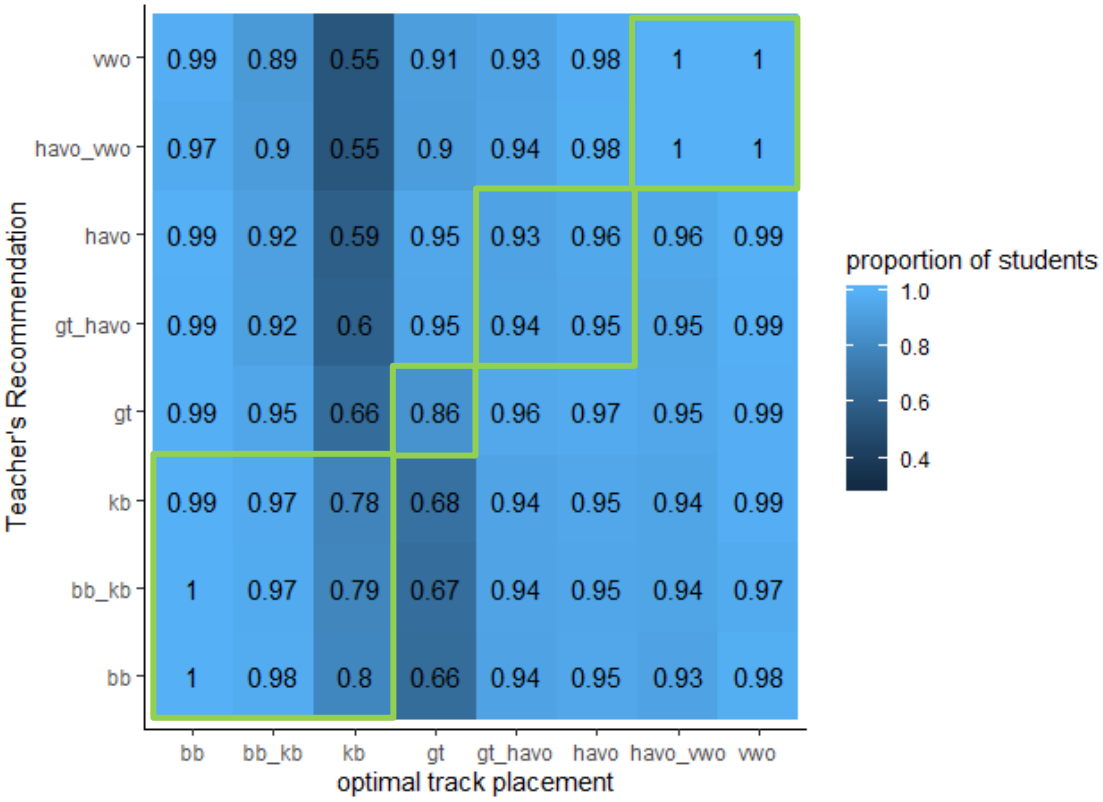
[6]


Teacher's Recommendation
vwo 0.98 0.89 0.54 0.93 0.93 0.99 1 1
havo_vwo 0.98 0.89 0.52 0.93 0.94 0.99 1 1
havo 0.98 0.88 0.52 0.93 0.94 0.99 1 1
gt_havo 0.99 0.93 0.64 0.77 0.96 0.97 0.95 0.99
gt 0.99 0.93 0.65 0.76 0.96 0.96 0.94 0.99
kb 0.99 0.94 0.66 0.74 0.96 0.96 0.94 0.99
bb_kb 0.99 0.95 0.67 0.73 0.96 0.96 0.94 0.98
bb 0.99 0.95 0.68 0.72 0.96 0.96 0.92 1
bb bb_kb kb gt gt_havo havo havo_vwo vwo
optimal track placement
proportion of students
1.0
0.8
0.6
0.4

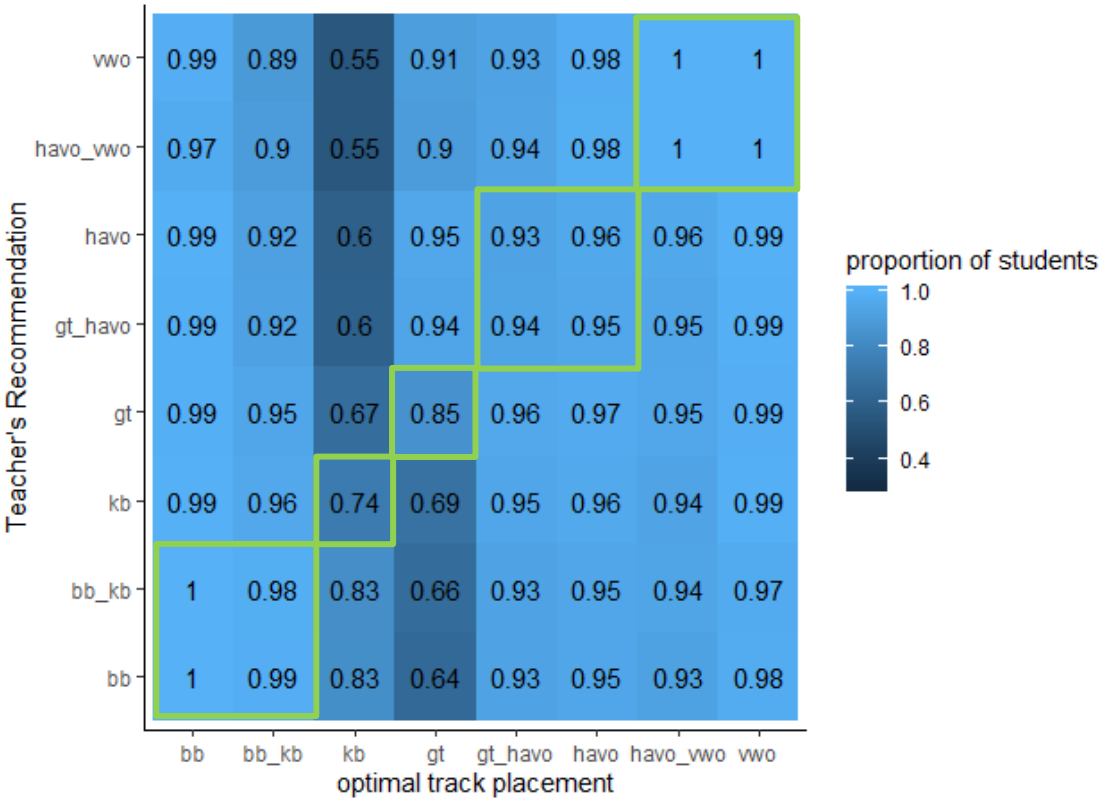
(A)


Teacher's Recommendation
vwo 0.98 0.91 0.57 0.91 0.93 0.99 1 1
havo_vwo 0.98 0.9 0.56 0.91 0.94 0.99 1 1
havo 0.99 0.89 0.56 0.91 0.94 0.99 1 1
gt_havo 0.99 0.94 0.66 0.92 0.97 0.96 0.95 0.99
gt 0.99 0.94 0.67 0.92 0.97 0.96 0.94 0.99
kb 1 0.97 0.8 0.67 0.94 0.95 0.94 0.99
bb_kb 1 0.98 0.81 0.66 0.93 0.94 0.93 0.98
bb 1 0.98 0.82 0.64 0.93 0.94 0.92 0.98
bb bb_kb kb gt gt_havo havo havo_vwo vwo
optimal track placement
proportion of students
1.0
0.8
0.6
0.4

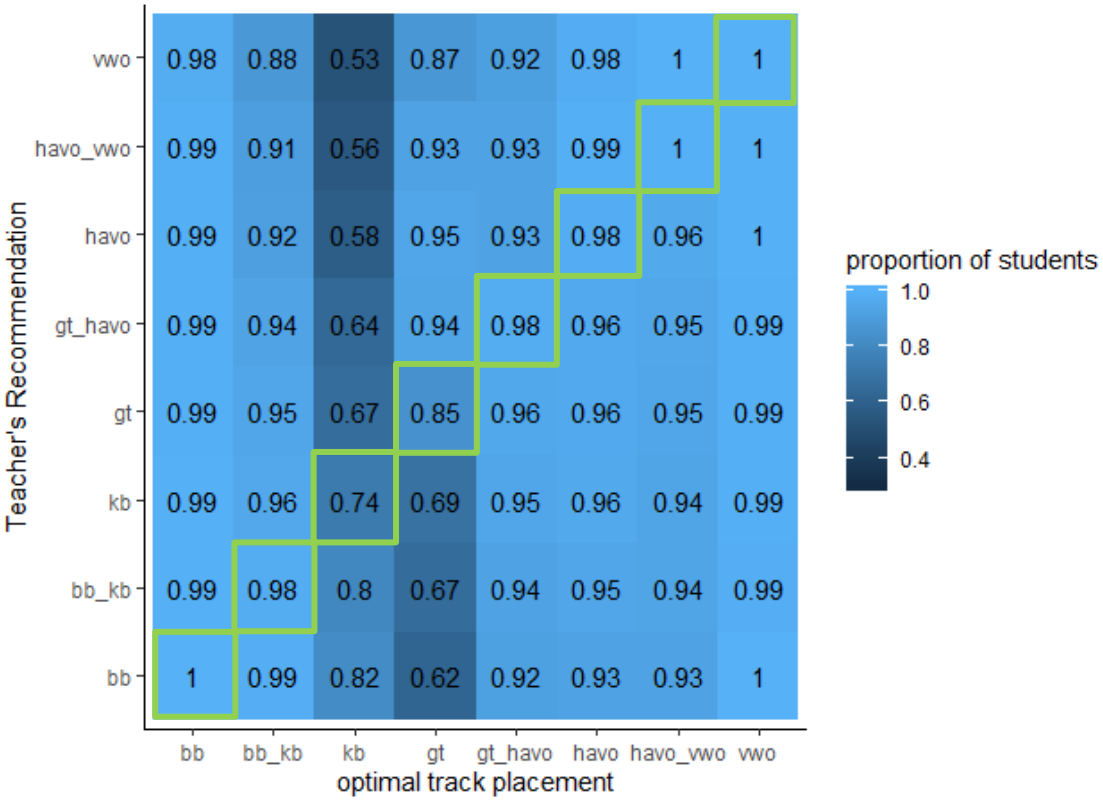
(B)


Teacher's Recommendation
vwo 0.99 0.89 0.55 0.91 0.93 0.98 1 1
havo_vwo 0.97 0.9 0.55 0.9 0.94 0.98 1 1
havo 0.99 0.92 0.59 0.95 0.93 0.96 0.96 0.99
gt_havo 0.99 0.92 0.6 0.95 0.94 0.95 0.95 0.99
gt 0.99 0.95 0.66 0.86 0.96 0.97 0.95 0.99
kb 0.99 0.97 0.78 0.68 0.94 0.95 0.94 0.99
bb_kb 1 0.97 0.79 0.67 0.94 0.95 0.94 0.97
bb 1 0.98 0.8 0.66 0.94 0.95 0.93 0.98
bb bb_kb kb gt gt_havo havo havo_vwo vwo
optimal track placement
proportion of students
1.0
0.8
0.6
0.4

(C)

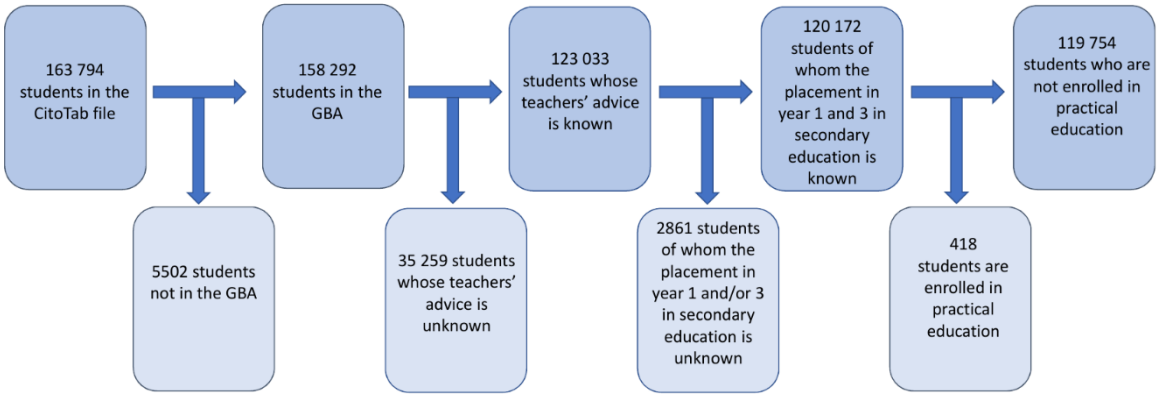

Teacher's Recommendation
vwo 0.99 0.89 0.55 0.91 0.93 0.98 1 1
havo_vwo 0.97 0.9 0.55 0.9 0.94 0.98 1 1
havo 0.99 0.92 0.6 0.95 0.93 0.96 0.96 0.99
gt_havo 0.99 0.92 0.6 0.94 0.94 0.95 0.95 0.99
gt 0.99 0.95 0.67 0.85 0.96 0.97 0.95 0.99
kb 0.99 0.96 0.74 0.69 0.95 0.96 0.94 0.99
bb_kb 1 0.98 0.83 0.66 0.93 0.95 0.94 0.97
bb 1 0.99 0.83 0.64 0.93 0.95 0.93 0.98
bb bb_kb kb gt gt_havo havo havo_vwo vwo
optimal track placement
proportion of students
1.0
0.8
0.6
0.4

(D)


Teacher's Recommendation
vwo
havo_vwo
havo
gt_havo
gt
kb
bb_kb
bb
0.98 0.88 0.53 0.87 0.92 0.98 1 1
0.99 0.91 0.56 0.93 0.93 0.99 1 1
0.99 0.92 0.58 0.95 0.93 0.98 0.96 1
0.99 0.94 0.64 0.94 0.98 0.96 0.95 0.99
0.99 0.95 0.67 0.85 0.96 0.96 0.95 0.99
0.99 0.96 0.74 0.69 0.95 0.96 0.94 0.99
0.99 0.98 0.8 0.67 0.94 0.95 0.94 0.99
1 0.99 0.82 0.62 0.92 0.93 0.93 1
bb bb_kb kb gt gt_havo havo havo_vwo vwo
optimal track placement
proportion of students
1.0
0.8
0.6
0.4

(E)

[7]


163 794 students in the CitoTab file
158 292 students in the GBA
123 033 students whose teachers' advice is known
120 172 students of whom the placement in year 1 and 3 in secondary education is known
119 754 students who are not enrolled in practical education
5502 students not in the GBA
35 259 students whose teachers' advice is unknown
2861 students of whom the placement in year 1 and/or 3 in secondary education is unknown
418 students are enrolled in practical education

[8]

List with Figure Captions

[1] A graphic representation of the Dutch educational system. After finishing primary school, which children attend from approximately age 4 to12, 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).

[2] (A) The current adaptive multistage design used by Cito. (B) Example of a new (hypothetical) design, where the teacher's expectation about a student determines the difficulty of the module on day 1. The orange line shows a possible route of a student who is capable of VWO but is expected to score low by their teacher.

Note. The GL and TL track recommendations from the EPST are combined into the GT track recommendation.

[3] A graphical representation of the data simulation.

[4] A graphical representation of the simulation of the introduced teacher-based routing designs.

[5] Plots of the standard errors per teacher-based routing design for students with 0-.5 levels mismatch between their teacher's recommendation and their optimal track placement, 1-2.5 levels mismatch and 3 or more levels mismatch. As each combination of modules results in a slightly different information function and thus in different standard error curves, only the standard error curves from the panels consisting of the middle module in the first stage and the middle module in the second stage and the panels consisting of the easiest and most difficult modules in the first stage and both the most difficult and easiest modules in the second stage were plotted.

[6] The proportion students who obtained a teacher-based routing EPST recommendation equal to (or in the category nearest) the regular-routing EPST recommendation for (A) all students with a certain teacher's recommendation independent of the regular-routing

EPST recommendation (B) for all students with a certain regular-routing EPST recommendation independent of the teacher's recommendation.

[7] The proportion of students who obtained a teacher-based routing EPST recommendation equal to (or in the category nearest) the optimal track placement for the design with 2 (A), 3 (B), 4 (C), 5 (D) and 8 modules (E) in the first stage.

[8] A schematic overview of the number of students in the datafile (n = 163794) and the numbers of students included in the present study (n = 119754).