Cut-off Scores in Adaptive Learning Systems

Jane (G.C.E) de Boer

Master Thesis – Applied Cognitive Neuroscience

S3133338

July, 2024

Department of Psychology

University of Groningen

Examiner:

prof. dr. D.H. (Hedderik) van Rijn

Second Supervisor:

T.J. (Thomas) Wilschut, MSc

A thesis is an aptitude test for students. The approval of the thesis is proof that the student has sufficient research and reporting skills to graduate, but does not guarantee the quality of the research and the results of the research as such, and the thesis is therefore not necessarily suitable to be used as an academic source to refer to. If you would like to know more about the research discussed in this thesis and any publications based on it, to which you could refer, please contact the supervisor mentioned.

Abstract

This study investigated the extent to which it is justifiable to consistently employ cut-off scores of 100% in adaptive learning systems (ALSs) and whether it might be more desirable to employ more flexibility in these scores. To answer this question, we collected teachers' opinions on the topic through a poll, a questionnaire and interviews. In addition, we conducted an experiment in which we tested participants' knowledge retention a day after they studied facts under two different cut-off scores (high and low), within Memorylab (an ALS). We found the teachers participating in our study generally opposed to scores of 100%. Instead, they prefer lower scores that they assumed would result in a balance between protecting student well-being and a minimum workable knowledge level. In addition, most teachers prefer to differentiate between students and learning tasks when setting cut-off scores. The results of our experiment provide evidence that studying under different cut-off scores within Memorylab can effectively achieve such differentiation: learning based on the high cut-off score was associated with significantly higher levels of long-term knowledge retention one day later compared to the low cut-off score. In addition, we found that learning based on the low cut-off score was associated with higher learning efficiency compared to the high cut-off score. Therefore, learning based on a lower cut-off score could help to reduce time spent on studying and reduce stress among students, thereby further underscoring the benefits of adoption of lower cut-off scores. Overall, the results of this study indicate it would be more desirable to adopt lower and more flexible cut-off scores, as opposed to consistently employing cut-off scores of 100%.

Keywords: Adaptive learning systems, cut-off scores, score-setting, student motivation, student well-being.

Cut-off Scores in Adaptive Learning Systems

Adaptive learning systems (ALSs) offer an alternative to the "one size fits all" approach that has been dominant within traditional education. This one size fits all approach focusses on what skill and knowledge students must learn but ignores students' individual needs and characteristics that influence how a student learns most effectively (Katsaris & Vidakis, 2021). In contrast, ALSs alter their behavior based on the individual users' needs and characteristics, thereby providing optimized learning paths for each individual student (Ennouamani & Mahani, 2017). For example, ALSs that focus on factual knowledge acquisition, which are the ALSs we will focus on within this paper, often use user performance data to personalize and optimize the scheduling of fact presentation within a lesson. For instance, these ALSs might present difficult facts more often and further enhance learning gains by ensuring that repetitions of facts are properly spaced within a lesson (e.g., *Anki - Powerful, Intelligent Flashcards*, n.d.; Sense et al., 2021).

Cut-off Scores

While ALSs can provide unique learning paths for individual users, they do not always offer the same flexibility in the learning goals they set. More specifically, ALSs might differ regarding the strictness of the cut-off scores they employ. These cut-off scores refer to the level of knowledge acquisition a user must demonstrate before the system decides a user knows the materials well enough and has successfully completed a lesson.

An example of an ALS that does offer considerable flexibility in setting cut-off scores is Anki, a flashcard practice system (*Anki Manual*, n.d.). During a learning session in Anki, users assess their own recall accuracy for every card presented by the system. When a user repeatedly

indicates that they know a card well, the card gets "promoted" to a review card and removed from the current deck. A lesson is successfully finished when the deck is empty. When a user indicates they forgot a card, the system will present the card more frequently. If a user forgets a card more than eight times, the card is tagged as a "leech". Users have the option to delete leeches from the deck when the importance of its content does not warrant a big-time investment. This means that within Anki, flexible cut-off scores are employed, since users can complete a lesson without having to demonstrate they know every fact.

An example of an ALS that employs more strict cut-off scores is Memorylab (Van Rijn et al., 2009; Sense et al., 2016). Memorylab is a model-based ALS that uses performance data of a user (accuracy and response times) to estimate how well a user knows facts within a lesson. When a fact is not well known yet, it will be practiced again. Only when the system's model predicts that the user is likely able to remember a fact 24 hours later, a fact is considered mastered. To complete a lesson within Memorylab, mastery must be achieved for all items within a lesson, resulting in a cut-off score approaching 100%. While employing cut-off scores of 100% might have several benefits, it also can bring forward several challenges, as I will outline below.

Arguments in Favor of Cut-off Scores of 100%

Employing cut-off scores of 100% may have multiple benefits. It could, for example, help to make the cut-off score setting process more straightforward and simpler. This simplification could provide a solution to some of the challenges related to cut-off score-setting as highlighted by literature on large-scale language assessment (for overview see, Eckes, 2012). For example, many widely used score setting methods, like the Angoff method (Angoff, 1971), require judges to make estimations about what minimally competent examinees might score on a test. These estimations are then used to determine the cut-off scores used to categorize test scores

into proficiency levels as, for example, “sufficient” or “not sufficient”. However, the process of estimating the performance levels of minimally competent examinees, is known to be cognitively challenging and can be prone to biases and inaccuracies (Brandon, 2004, as cited in Eckes, 2012). In contrast, 100% cut-off scores offer an unambiguous proof of mastery of the materials, thus mitigating some of the challenges related to choosing appropriate cut-off scores.

Challenges Related to Cut-off Scores of 100%

On the other hand, cut-off scores of 100% could bring forward several challenges. To begin with, acquiring and demonstrating the last 10-20% of knowledge can take a relatively long time (cf. Dunford & Tamang, 2014). Having to spend a long time on achieving 100% scores could increase negative (achievement) emotions like frustration and stress, especially in students for whom the materials are already challenging. Importantly, these negative emotions can lead to further negative effects. Namely, frustration can decrease motivation or interest in learning and hinder the learning trajectory in multiple ways (Pekrun, 2006). In addition, increases in academic stress can be undesirable as such increases are associated with lower levels of student well-being (e.g., Barbayannis, 2022).

Another challenge related to fixed cut-off scores of 100% is that school success is judged based on predefined numbers that might not match students’ own learning goals. In fact, when students’ own goals are not considered because all students are expected to perform the same, this resembles a “one-size-fits-all” approach (Guterman, 2020), which is the very approach that ALSs are supposed to offer a solution to (Katsaris & Vidakis, 2021).

Moreover, scores of 100% do not align with scores required to pass in-class tests. In the Netherlands, a score of below 100% is sufficient, as long as it results in a 5.5 (*Dutch Grading*

System, n.d.). This raises the question why students would need to master higher percentages when studying in ALSs.

Finally, assessing students' competence with ALSs may have advantages over traditional testing. Namely, traditional methods measure competence at a single point in time, which allows students to engage in mass studying and "cram" in all the test materials the day before the test is due (e.g., McIntyre & Munson, 2008). Cramming is a common learning strategy among students and can be effective for short-term knowledge retention, thereby resulting in good test results. However, cramming does usually not result in long-term knowledge retention, causing most study materials to be forgotten soon after the test took place (e.g., Simon & Bjork, 2001, as cited in Penn, 2019). In contrast, rehearsal of materials through spaced learning sessions leads to more durable memories. This phenomenon is known as the spacing effect (e.g., Dempster, 1989). Testing with ALSs can help to promote spaced learning. For example, instead of testing students one time, students could be required to repeat study materials by completing the same lesson on several different days. This approach not only helps ensure sufficient proficiency levels at the time of testing but also enhances the chances of long-term knowledge retention. But here again the question arises: what should the passing score be? Do students need to master all the materials or should passing scores align with those employed for traditional tests? These questions, together with the aforementioned points, raise the question of whether it is justifiable to consistently employ cut-off scores of 100% or whether it might be desirable to employ more flexibility in these cut-off scores.

Teachers' Perspectives

To answer the question raised in the previous paragraph, it is crucial to incorporate teachers' insights. Teachers are experts in the field of education, so when their perspectives are

taken into consideration, this will help to understand how to optimally improve ALSs and increase the research's practical relevance. Importantly, when the research findings have higher practical relevance, this will likely increase teachers' willingness to implement research findings in their practice (Mohajerzad et al., 2021). To our knowledge, there is no literature published on teachers' perspectives on cut-off scores employed by ALSs.

Empirical Data

In addition to investigating teachers' preferences regarding cut-off scores in ALSs, it is crucial to empirically test whether ALSs can meet these preferences. Namely, further advancements in ALSs can only be made once this aspect is also investigated. Again, to our knowledge there is little to no research of this kind conducted yet.

Current Study

As previously outlined, there is a need to determine whether fixed cut-off scores of 100% are justifiable or whether it might be desirable to employ more flexibility in these cut-off scores. The current literature lack data to adequately address this issue. Therefore, in this study, we aim to address this literature gap by exploring teachers' perspectives on the topic and by conducting a controlled experiment.

To get insight into teachers' perspectives, multiple approaches for data collection will be utilized, involving questionnaires, polls, and interviews. Questionnaires and polls will be employed to learn what scores teachers find acceptable, while interviews will be used to uncover the teachers' arguments for choosing certain scores and to contextualize their choices. The data gathered will address several key questions:

Q1: What cut-off scores do teachers deem acceptable for different learning tasks and why?

Q2: Are teachers open to using different cut-off scores for different students?

Q3: Would teachers allow students' involvement when setting cut-off scores?

Q4: Would teachers consider replacing traditional tests with ALSs and what scores would then be acceptable?

The empirical study will consist of a two-day long experiment. On the first day, participants will learn different sets of facts, based on two different cut-off scores within in an ALS. On the second day participants will be tested on their knowledge retention for facts studied under both cut-off scores. Based on the data collected in this study, the following research questions will be answered:

Q5: What differences in learning effects can be observed when participants study facts under different cut-off scores? Here, we will specifically examine the levels of long-term knowledge retention and learning efficiency associated with both cut-off scores.

Q6: To what extent is the ALS used in the experiment able to meet teachers' preferences, based on the findings for Q1 through Q5?

Methods

Before any participants were recruited, this study was submitted as a fast-track procedure and approved by the Ethics Committee of the University of Groningen (research code: PSY-2324-S-0273). Because we submitted this study as a fast-track procedure, it was exempt from a full ethical review

Poll

Participants

The poll was targeted at secondary school teachers within the LinkedIn network of my supervisor dr. H. van Rijn. In total 27 teachers filled in the poll. No further information was collected on the teachers' demographics.

Materials

In the LinkedIn poll, teachers were asked how much of the study materials students should master when learning vocabulary at home, under the assumption that it would never be tested. The answer options were: 100% ("a 10"), 80% ("good but not perfect"), 60% ("a sufficient grade"), and <= 50% ("this is not what it's about").

Procedure

For this project a descriptive research design was employed. The poll was posted on the LinkedIn account of my thesis supervisor. Every secondary school teacher who encountered the post was invited to vote. The poll remained active for one week.

Questionnaire

Participants

The questionnaire was handed out to high school teachers who attended a presentation about ALSs during an information day on AI in education. In total 10 high school teachers filled in the questionnaire. No further information was collected on the participants' demographics.

Materials

The questionnaire included items that asked teachers about their opinions on cut-off scores in ALSs and whether they would differentiate between students when setting these scores. An example of a question from the questionnaire is: “What scores do students need to reach when the learning is part of the homework?”. To access the full questionnaire, see Appendix A.

Procedure

For the questionnaire, a descriptive research design was employed. The questionnaire was distributed at the end of a presentation about ALSs during an information day on AI in education. The questionnaires were printed out on paper prior to this information day. After the presentation, teachers were requested to fill in the questionnaire by pen. Filling in the questionnaire would take no more than a few minutes.

Interviews

Participants

Participants were teachers teaching at either secondary schools or secondary vocational education institutions (MBO). The participants were selected based on the condition that they had experience in teaching a foreign language. Recruitment of the participants took place through personal contacts of both my supervisor and me, as well as through outreach on LinkedIn. In addition, participants were asked to recommend other teachers who might be interested in participating. The participants’ demographics are summarized in Table 1.

Table 1

Summary of the Interview Participants' Demographics

Participant ID	Years Experience	Subject	Years/Classes Taught	Type of School
1	3.5	English Language	HAVO and VWO, years 2,3 and 4	Regular high school
2	3	English Language	First year lyceum and HAVO year 5 + VWO year 2	Regular high school
3	6	German Language	HAVO and VWO, year 2 through 4	Montessori high school
4	3	Spanish Language	HAVO and VWO, year 1 through 5	Regular high school
5	16	English Language + other courses not disclosed to prevent identifiability.	Levels 3 and 4 at MBO years 1 through 3	Currently MBO
6	14	English Language + other courses not disclosed to prevent identifiability.	MBO level 4 + several high school classes	Regular high school + MBO
7	7	English Language	Upper secondary years HAVO and VWO, except for 4 VWO	Montessori high school

Note. VWO stands for pre-university education, HAVO stands for higher general continued education, MBO stands for vocational education.

Materials

We developed a semi-structured interview guide specifically for this study. The questions withing this guide focused mainly on teachers' perceptions of acceptable cut-off scores and how they decide what acceptable scores are. An example of a question is: "what cut-off score do you want your students to reach when the learning is part of the homework?". To access the full interview guide, see Appendix B.

Procedure

For the interview a qualitative study was employed. Five of seven interviews were conducted in person. Usually, I would visit the teachers at their work, where we would sit in private classrooms or in the school canteen. One teacher visited me at the University of Groningen, where we sat in a private classroom. Two of the interviews were conducted online, via Google Meet. Before the interview started teachers were asked to sign the consent form. To enable offline transcription, the interviews were audio recorded. After transcription, all audio files were deleted. Each interview lasted approximately 20 minutes. Before the interview ended teachers were given the opportunity to address issues they thought were insufficiently discussed or unclear.

Experiment

Participants

Participants were 17 first year psychology students from the university of Groningen. The participants were recruited through the university's online SONA system and compensated with SONA-credits. One participant did not complete the full study due to illness, resulting in a final

sample size of 16. No further information was collected on the participants' demographics.

Materials

Participants completed both the learning and testing phases using computers in the labs at the University of Groningen.

Study Materials. The study materials were facts from the first-year course biopsychology at the rug (PSBE1-04). Every fact consisted of a cue describing a term or process from biopsychology (for example: "Episodic memory relies on the _____.") and the corresponding answer (for example: "Hippocampus"). At the time of the study, participants were halfway through the biopsychology course. To minimize the chance of prior familiarity with items, this study included only facts that would be covered during the second half of the course. Two sets of facts (set A and B) were created, each containing 30 unique items in total. To access both fact sets, see Appendix C.

Organizing and Testing Software. On the first day Qualtrics (Qualtrics, Provo, UT) was used to redirect learners to the online learning sessions. On the second day Qualtrics was used to test participants on the studied materials. All items were tested on separate pages so response times could be recorded for every item on the test.

Learning Software. Participants studied items withing the adaptive learning system Memorylab (Van Rijn et al., 2009; Sense et al., 2016). Memorylab optimizes testing and spacing effects within a single lesson by estimating the activation of items. The activation of an item refers to how likely the user can retrieve the item from their memory. In other words, the higher the activation, the better an item is known. Activation is highest when an item is currently presented and drops when it disappears from the screen. Memorylab uses two measures of

performance data to predict the activation of items: accuracy and reaction time. Items to which the user gives correct and fast responses have higher predicted activation and items for which incorrect and slow responses are given have lower predicted activation (Memorylab, 2024).

Within Memorylab, the presentation of items is scheduled based on the estimated activation of items: the item with the activation closest to the forgetting threshold is presented first. When an item's activation drops below this threshold, users are unable to retrieve the information from their memory. When all currently practiced items have estimated activation substantially above this forgetting threshold, new items will be presented (Braam, 2024).

Besides the forgetting threshold that is used to schedule items within a learning session, Memorylab also uses a mastery forgetting threshold (Braam, 2024). An item is mastered and removed from the lesson when the activation after 24 hours is predicted to be above this mastery forgetting threshold. A lesson is finished when all items are mastered. The mastery threshold can be set to "high" or "low". The likelihood of recalling items from a lesson 24 hours later is greater for the high threshold compared to the low threshold. Mastering items under the high threshold requires more successful repetitions and faster reaction times compared to the low threshold. As a result, the speed at which items are mastered and removed, and new items are introduced is slower under the high threshold compared to the low threshold. Consequentially, completing lessons under the high threshold requires more time and effort from users compared to the low threshold. The mastery threshold level represents the cut-off score employed in a lesson.

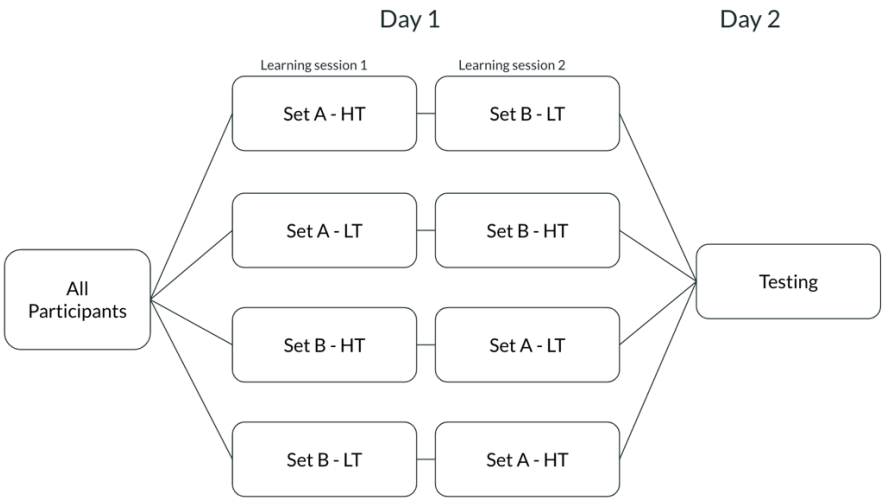
Procedure

This study utilized a 2x2 within-subject design (see Figure 1). In two learning sessions, participants studied two sets of facts (set A and B) based on two mastery forgetting thresholds

(high and low), which influenced the scheduling of item presentation (see the previous section). Participants studied each set under a different threshold. The order of fact sets and thresholds employed, were counterbalanced to control for order effects. In total there were four unique conditions. The experiment was conducted over two consecutive days. On the first day, participants engaged in the learning sessions. On the second day participants were tested on all items from both fact sets.

Figure 1

Overview of Experiment Design



Note. LT stand for low threshold, HT stands for high threshold.

The first part of the experiment lasted approximately 30 minutes. Participants would come to the lab, where they were seated behind computers within private cubicles. Participants first received the study information and were asked to give consent. Next, participants were randomly assigned to one of the four experimental conditions and redirected to the first learning

session. Within the study session new study items were introduced by displaying both the cue and the corresponding answer (see Figure 2). During subsequent presentations (rehearsal trials) of the item, participants were presented with only the cue, and they had to type in the corresponding answer (see Figure 3). Whenever participants answered incorrectly, the system would notify the user about this and provide the participant with the correct answer. The learning session lasted 12 minutes.

Figure 2

Example of Initial Presentations of Items

Figure 3

Example of Subsequent Presentation of Items

After the first learning session was finished, participants were redirected to the second learning session. The second learning session was the same as the first learning session, with the exception that a different set of facts was studied, and a different threshold was employed. After finishing the second learning session, participants were asked how much effort they invested while studying the facts and if they had previously encountered any of the practiced facts. After this, the first part of the experiment was finished.

On the next day participants came back to the lab and completed a recall test for all items within both fact sets. On the test, cues were given for all items and participants had to type in the corresponding answer. For all items on the test, reaction times were recorded. At the end of the

test, participants were asked how much effort they invested in completing the test and whether they had any conditional comments. Completing the test took no more than five minutes.

Results

Poll

Through the poll we aimed to get insight into what cut-off scores teachers would deem acceptable. Teachers were asked the following question: "How much of the study materials should students master when learning vocabulary at home, under the assumption that they would never be tested?". Responses to this question were analyzed by quantifying the votes for each answer option using LinkedIn's built-in polling features. A total of 23 participants voted that students should achieve an 80% mastery level ("good but not perfect"), three participants voted for a 100% mastery level ("perfect"), and one participant voted that a mastery level of 50% or less ("this is not what it's about") is sufficient. Notably, no participants chose a 60% ("a sufficient grade") mastery level (see Table 2).

Table 2

Responses Questionnaire Cut-Off Scores

Percentage	Votes
100%	3
80%	23
60%	0
<50%	1
Total	27

Questionnaire

Analysis of Cut-Off Scores.

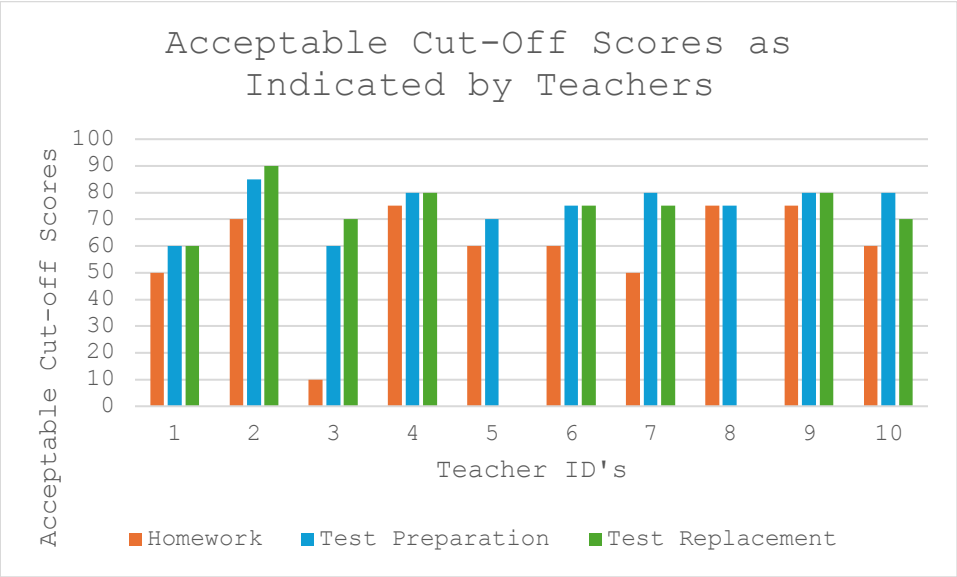
Through the questionnaire we aimed to get insight into what cut-off scores teachers would deem acceptable, for different learning tasks. The first question on the questionnaire inquired about what cut-off scores teachers would deem acceptable when the learning is part of the homework. On average teachers chose a score of 58.50%, with a relatively wide range of 10% to 75%. When asked what scores would be acceptable for test preparation, teachers showed more uniform and strict expectations, with an average score of 74.50% and a narrower range of 60% to 85%. The overall tendency to employ stricter scores for tests was also reflected by the fact that all individual participants but one chose higher scores for tests compared to homework (see Figure 4).

In addition, participants were asked if they were open to replacing some traditional tests with ALSs and, if so, what scores students would need to reach. In total, only eight participants responded to this question. Their responses averaged to a score of 75%, ranging from 60% to 90%, closely aligning with the scores that teachers indicated for test preparation.

Notably, no participant chose cut-off scores of 100% for neither homework, tests nor test replacements. In fact, only participant 2 indicated scores of 85% for test and 90% for tests replacements, while other participants chose no scores higher than 80%.

Figure 4

Acceptable Cut-Off Scores indicated on Questionnaire for Different Learning Tasks



Note. Each teacher is represented by a unique teacher ID. A missing bar for "Test Replacement" indicates that a teacher did not respond to this question.

Analysis Differentiation.

For both homework and test preparation, teachers were asked whether they would differentiate between students when setting cut-off scores. In addition, teachers were asked how they would differentiate: which score would be set for which student. Responses to the questions about differentiation are summarized in Table 3. Several teachers indicated they would be willing to differentiate based on individual student characteristics such as subject proficiency, difficult home situations, or whether a student is a deadline worker or a planner. The participants indicated that those students who struggle with the materials due to these personal challenges might be allowed to learn based on lowered cut-off scores, compared to students who do not face such challenges.

Besides differentiating based on individual students' characteristics, some participants also indicated they would differentiate based on group characteristics. Examples of group characteristics that were mentioned by teachers are: lower secondary classes vs upper secondary classes and educational levels (VMBO, HAVO, VWO). Participants did not indicate how they would differentiate based on these group characteristics. It is therefore inferred that upper secondary and higher educational classes are expected to study based on higher cut-off scores compared to lower secondary and lower educational level classes.

Notably, in total two participants indicated that they would not differentiate when they set scores for homework. This number increased to four when the question was applied to learning as preparation of a test. These results suggest a stricter assessment approach for test preparation, compared to homework, aligning with the differences in cut-off scores setting, discussed in the previous section.

Table 3

Overview of Participants' Responses to Questions About Differentiation

Teacher ID	Differentiation homework	Differentiation Test
1	Yes, based on individual student characteristics	Yes, based on individual student characteristics
2	Yes, based on class characteristics	Yes, based on class characteristics
3	Yes, based on individual student characteristics	Yes, based on individual student characteristics
4	Yes, based on individual student characteristics	Yes, based on individual student characteristics

Teacher ID	Differentiation homework	Differentiation Test
5	No	No
6	Yes, based on individual student characteristics	No
7	Yes, based on class characteristics	Yes, based on class characteristics
8	No	No
9	Yes, based on individual student characteristics	No
10	Yes, based on individual student characteristics	Yes, based on individual student characteristics

Note. Each teacher is represented by a unique teacher ID

In summary, the questionnaire results reveal that teachers prefer higher cut-off scores for test preparation or test replacement, compared to homework. In general, teachers were willing to adjust scores based on individual or group characteristics, but less for tests than for homework.

Interviews

The interview data were processed using thematic analysis. The first step of this analysis was transcribing the data. Next, I used codes to categorize participant's answers for every question separately. Codes were then compared, adjusted if necessary, and further organized into overarching themes for every question separately. Finally, these themes were used for the presentation of the interview data.

In this chapter, key findings that emerged from the thematic analysis are presented. It will be described which cut-off scores teachers consider acceptable for different situations. Whenever relevant, explanations of their choices will be illustrated using quotes or paraphrases. Note that for this presentation, the interview results were translated from Dutch to English, with the help ofChatGPT.4 (OpenAI, 2023). To access the results based on the original interview in Dutch, see Appendix D.

School Success and Assessment

The first main question of the interview addressed what teachers understand by "school success" and the role tests and grades play in evaluating school success. The teachers' responses are summarized below.

Personal Development and Well-Being. Three teachers mentioned characteristics of personal development as indicators of school success. For instance, teacher 7 emphasized that it is especially important for students to feel comfortable at school and to discover their place in society:

> I do think it's important that it happens on multiple levels. I don't think it's just about the didactic part: so, it doesn't have to be just about grades. I also think it's very important that they feel comfortable and […] that they know what their place in society is and will be in the community, and how they can actively participate in it.
>
> (Teacher 7)

Teacher 4 also mentioned that it is important for students to feel comfortable at school and that grades are not the only indicator of school success:

> I think it's always nice for a student when things are going well at school, in terms of grades, but I think it's even more important that a student is happy in the classroom, apart from the grades, that they have a good relationship with classmates, enjoy their time, and feel at home. And I think teachers play a big role in this. If a student feels comfortable with teachers, that can also contribute a lot.
>
> (Teacher 4)

Teacher 1 mentioned the importance of growing up and developing the social skills that come with it: "It is more about becoming an adult: learning to plan, dealing with people, showing respect, and developing social skills."

Process Focused Evaluation. Two teachers indicated that a student's learning the learning process is an important indicator of school success:

> At one point, we had a training on the concept of the 'growth mindset'. It was about how students are sometimes too focused on grades, while as teachers, we would want them to develop a growth mindset so that they assess themselves based on how much they have improved since the last time and what they have done to learn in more effective ways than before. I think that if you as a teacher can contribute to that mindset, then the highest achievement has been reached. This doesn't mean that grades are unimportant, but they might be of a somewhat secondary importance. […] Grades from tests can still help to compare a student with themselves.
>
> (Teacher 2)

According to teacher 3, sometimes the focus lays too much on tests, which results in less attention to the learning process, even though it can be more indicative of school success. Tests, according to teacher 3, can help identifying points that require attention:

> I also think more about the process, which is also more Montessori. Instead of just being busy with tests. I notice this when teaching the German language too: you're only busy with that test, it rushes through, while students may not have understood it yet, but then you already have to move on to the next topic while you should actually focus more on the learning process of: where does it go wrong, have you mastered it now? […] In such a process, of course, you must achieve something with tests. You have to work with a goal, otherwise you are obviously doing it for nothing, but when you focus more on the process, you can then shape the product, instead of it being loose components.
>
> (Teacher 3)

Functional Competence. Two teachers mentioned functional competence as an important indicator of school success. These teachers emphasized the importance of being able to work with the knowledge acquired during one's education:

> Personally, I think it's more about the student's own experience. At the educational institution where I worked, we used portfolios, and you had to demonstrate that you understood certain topics and whether you had mastered them or not. We also did this through interviews, and then you had to show what you knew. I believe that if you can prove that you can process and understand the theory covered in class into a story, then you have been quite successful.
>
> (Teacher 6)

Teacher 5 also considered it important for students to be able to apply what they have learned. Teacher 5 emphasized learning objectives within vocational education, focusing on preparing for professional practice:

> In vocational education, it's really about training students so they can manage in the professional world, with the foreign language I teach. So, I would say if they can professionally handle the English they have learned, then it seems to me that they have been successful in completing their education and preparing for professional practice. When they can manage the language professionally, they are usually at about a B1 level or so.
>
> (Teacher 5)

When I asked whether this teacher uses exams to determine if students are at the B1 level, they indicated: Yes, several at least. They must take 5 generic exams: reading, listening, speaking, conducting conversations, and writing. For those exams, we indeed test a specific aspect of English each period, which ranges from basic to increasingly advanced.

Summary School Success and the Role of Tests. Teachers approach the concept of school success more broadly than just academic performance and emphasize personal well-being and the learning of functional skills. The role of tests is seen as a tool for measuring individual progress and practical competence, and not as an all-encompassing indicator of school success.

Cut-off Scores for Homework

The second main question was what cut-off score teachers would maintain when the learning is part of the homework. I suggested the teachers to imagine that the learning material

consisted of a vocabulary list, and they could decide what percentage of this list students should know. Every teacher indicated that scores below 100% were acceptable. The difference between the highest (80%) and lowest (50%) mentioned score was 30%. The average score was 71%. See Table 4 for an overview of scores indicated per teacher.

Table 4

Acceptable Cut-off Score, as Indicated by Teachers, for Homework and Tests

Teacher	Cut-off Score Homework	Cut-off Score Test Preparation
1	80%	80%
2	75%	75%
3	65%	75%
4	50%	60-70%
5	80%	100%
6	70%	80-85%
7	75%	Preference Student

I asked the teachers to explain why they chose the scores presented above and why not a higher or lower score. Themes in their answers are presented below:

Students' Learning Experience: Motivation, Stimulation, and Enjoyment of Learning. Four teachers mentioned the learning experience of students as a reason for choosing the cut-off scores. For instance, two of them indicated that using higher cut-off scores than those chosen could lead to a loss of motivation or enjoyment in learning among students:

> I think higher percentages (than 70%) create more performance pressure, which can be counterproductive because you are more often confronted with the fact that you are not there yet. Success should be achievable more quickly, so that learning remains fun.
>
> (Teacher 6)

Another teacher also mentioned that applying a higher cut-off score could lead to a loss of enjoyment in learning:

> I think some students find learning vocabulary difficult while they are good at other aspects of the language. A vocabulary test does not accurately reflect how good they are in general. 75% is then a compromise, otherwise you punish students too harshly who are good in English but less good at cramming words. […] This can lead to some students who are good in English, but not in vocabulary, no longer finding the learning enjoyable even though they initially did.
>
> (Teacher 2)

In contrast to the arguments mentioned above, teacher 4 pointed out that using cut-off scores that are too low can have a negative effect on a student's learning experience. This teacher maintained a base cut-off score of 50%, which was already lower than that of other teachers, but gave a clear reason for not going lower:

> Look, that 50% will still increase to an acceptable percentage in class. And you should have some expectations of the student and if the bar is set too low, then you don’t really encourage them to keep trying their best.

(Teacher 4)

Room for Improvement During the Class. Three teachers mentioned choosing the aforementioned scores (and not higher) because there is time during class to increase this percentage. For example, participant 3, who chose a score of 65%, stated: "I normally don't give homework and they can make up the rest in class." As previously described, teacher 4 also mentioned that a score of 50% is acceptable, as it can be increased to a higher percentage during the class.

Learning from "Mistakes". Two teachers indicated choosing the aforementioned scores (and not higher ones) because allowing students to make mistakes can improve the learning process:

> If you are allowed to make mistakes, you have the opportunity to ask the teacher or classmates 'hey why is this the case', which helps them broaden their knowledge. If they go up to 100% and they are done with it, then in their head they are also done with it and think 'Oh good, done. Next topic.' […] But if they learn up to 80% and they have some errors and they then work with multiple grammatical aspects, then you can discuss this in class and [..] then you can really examine: why was this wrong? So that's why I would not choose 100%, because if it is 100%, they will leave it behind. They think they know it all.

(Teacher 1)

Minimal Workable Knowledge Level for Test, Class, and Practice. Several teachers indicated choosing the aforementioned cut-off scores because they assume these scores represent

a minimal workable level of knowledge. Two teachers mentioned that the score they mentioned is the minimum score needed to pass the test:

> When I look at tests, that's 4 x 8 words. I think a good score is 1-2 errors per task, so 24 words out of 32 correct, which equals 75% correct. So that must also apply to the homework. Because the test is a summation of everything for the homework, and if students achieve this then they will also likely pass the test. And you often can tell whether a student is well-prepared or not, and being well-prepared means that a student also is more likely to perform well on a test.
>
> (Teacher 2)

In line with teacher 2, teacher 5 expressed a concern that learning to a lower score (than 80%) increases the risk that students will not pass the test:

> Tests consist partly of vocabulary and partly grammar, but if you don't know the vocabulary, you often don't know the grammar either; usually, this means they have not studied well. But mostly, if you don't know the vocabulary, you have too little to compensate with. It's too big of a risk.
>
> (Teacher 5)

Besides the ability to pass a test, two teachers emphasized the importance of being able to keep up during the class. Teacher 3 said the following: "A foundation must be laid on which they can build during the class. This base of 65% is a foundation with which you can apply or enrich the course materials."

Finally, the ability to apply the acquired knowledge in practice was also mentioned as a reason for choosing a specific score:

> If you learn up to 70% then you have mastered the majority, and then you can deduce the meaning of other words from the context. […] I think it's very nice if you manage to learn more, but I personally don't believe that you always need to know all the words. When it comes to English, I think you should be able to communicate with someone, and that can also be achieved with a slightly smaller percentage. If you can communicate, then you are already well on your way.
>
> (Teacher 6)

Summary Cut-off Scores for Homework. In summary, teachers choose scores that they assume, would result in a balance between maintaining motivation and sufficient mastery of the material. Optimal scores allow room for improvement during the class, help in preparing for tests, and are sufficient for the practical application of the coursework. Teachers want to protect the motivation and enjoyment of learning of the students by not choosing very high cut-off scores. At the same time, they would not choose scores that are too low because they fear that such scores cannot ensure a workable level of knowledge or provide students with sufficient challenge.

Cut-off Scores for Tests

All teachers, except for one, indicated that scores below 100% are acceptable when it comes to learning in preparation for a test (see Table 4). The difference between the highest

(100%) and lowest (60%) mentioned score was 40%. The average score was 80%. About half of the teachers gave a higher score for learning for a test compared to learning as part of homework.

Again, I asked the teachers to explain why they chose these scores and not a higher or lower score. The answers given by the teachers can be summarized in two main themes: test norms and student motivation:

Test Norms. Five teachers indicated that their choice for scores was based on the norms applicable to the test. For example, three teachers chose the percentage that often leads to a passing grade (around 60% for HAVO and 70-75% for VWO):

> What we do now is that they must achieve three crowns for Slimstampen, which is 3 times 100%. And then they get a test and that is then 60 or 70 percent. But actually, 70% should be enough because that also applies to the test.
>
> (Teacher 3)

Teacher 1 also considered the norm-setting that applies to tests but chose a slightly higher score than the minimum percentage needed for a pass. Teacher 1 indicated that test stress sometimes prevents students from remembering everything they have learned. Therefore, Teacher 1 chose a score that could compensate for this stress:

> You also have to consider that there's more stress involved with a test. […] You need to score 70% to get a 5.5. During the test, you experience stress and can't really look things up. My mind goes back to 80, 90%. Maybe even a bit higher, so that you feel more confident.
>
> (Teacher 1)

Teacher 7 based their choice on existing norms. However, instead of requiring students to learn up to a specific percentage, this teacher would discuss with the students the grade they want to achieve. Based on these discussions, the teacher would then determine how much the students needed to learn, in line with the test norms.

Motivation. Two teachers indicated that they chose the score to protect or increase student motivation. For example, teacher 4 mentioned that using a higher score than 60-70% is not achievable for everyone and can discourage students. Teacher 6 also mentioned that allowing students to make a mistake can motivate students to do better next time. Additionally, this teacher mentioned that high cut-off scores that are too high could led to students believing they have mastered the material so thoroughly that further study is unnecessary: "I also think that when you've already achieved 90-100%, then you get kids who think 'oh, I can already do it all.' And then they will lean back during classes".

Protecting the Student. Two teachers indicated they chose the aforementioned scores because higher scores can negatively affect the well-being of students. For example, teacher 6 mentioned that students can learn from making mistakes and that in today's society it is important to learn that making mistakes occasionally is okay. In addition, teacher 1 specifically expressed fear of fueling burn-out symptoms when using scores of 100%:

> If students always have to understand 100%, that's really a lot and it really leads to burn-out symptoms. You see this now in all students with test anxiety and such things who study so much and always try to score 100%. And when they then take a test, they get a blackout if they've forgotten one thing. So, I wouldn't say 100%.
>
> (Teacher 1)

Homework vs. Test. As seen in Table 4, teacher 6 chose a score of 100%. The teacher's explanation was that they had chosen a score of 80% when it comes to learning for homework, and would want to increase the score for a test.

Summary Test Replacement. Overall, teachers consider scores below 100% acceptable when students learn for a test. The choices for these scores are based on test norms and student motivation. Three teachers chose scores that would result in a passing grade on tests (60-75%). Various teachers also emphasized the importance of setting realistic goals to prevent demotivation and burn-out symptoms in students. Generally, teachers opt for a balance between being able to pass the test and protecting the well-being and motivation of students.

Differentiation

The third main question was whether teachers are open to making distinctions between students when determining the scores they need to achieve. When posing this question, I mentioned that not every student learns at the same pace when using the learning software, and some students may take a long time to achieve high scores. The teachers' responses are briefly summarized in Table 5.

Table 5

Teachers' Preferences Regarding Differentiation When Setting Cut-off Scores

Teacher ID	Answer
1	Mainly when students want to learn more than the score set by the teacher

2	Mainly when students want to learn more than the score set by the teacher
3	Would differentiate both ways
4	Would differentiate both ways
5	Would differentiate both ways
6	Would offer different learning materials, instead of differen cut-off scores
7	Would differentiate both ways (preference student)

The explanations given by the teachers for their answers are displayed below, according to themes:

You Must Expect a Minimum Level. Teacher 1 expressed openness to differentiation, especially by tailoring the content for students who already understand the material well. For students who find the material challenging, this teacher would only make distinctions for a limited duration:

> Students who have mastered the material may skip a number of assignments. […] If you look more from the perspective of someone who finds it more difficult, I would really maintain that line. We're talking about an educational level and if they can't handle it for English, French, German, or something similar, then tutoring is probably something you can apply. But you shouldn't go lower because they can't handle it, I think. Maybe initially to build confidence, but if it happens every time, then those grades stay low. I believe the level should really be at a sufficient level. Someone who is in HAVO should

> not be constantly getting 4s, barely 5s or just 5.5s while they are really trying, so I don't think it should be lower.
>
> (Teacher 1)

Differentiation is Confronting. Two teachers expressed difficulty with differentiating downwards, as it could be confronting for students who find the material challenging:

> You quickly enter a difficult area: from the student's perspective, you're performing less well compared to someone else, and therefore you only need to learn to a lower percentage. […] Maybe you can present it a bit differently: leave the slower learner as is, but a quick learner gets a check next to their name, a kind of bonus or recognition that you're doing well.
>
> (Teacher 2)

Teacher 6 also mentioned that making distinctions based on different percentages can be confronting for students who struggle more with the material. Therefore, this teacher would prefer to differentiate by having students study materials at different language levels while completing the same percentages.

Four teachers expressed a willingness to adjust scores for both students who master the material and those who find it challenging. The arguments they provided are summarized below:

Student Learning Experience: Frustration and Motivation. Teacher 5 was open to differentiating in both directions, because they assumed it would be beneficial for student motivation: 'Yes, I think it's important that they stay motivated, so it has to be achievable, you know.' Teacher 3 also wanted to make distinctions because learning with such a system is not

equally easy or quick for every student. For some, using the system could even be frustrating, and for these students, it is important that they can learn until they reach a lower score. I asked this teacher how making distinctions affects motivation, and they responded as follows:

> I think this is good for motivation because you then look more at the student and they realize, 'okay, I find this difficult' but the teacher still makes an adjustment for me to make it feasible to deal with Slimstampen'. This is preferrable over just completing the list because then you get the bonus, even though for that student it's just an obstacle.
>
> (Teacher 3)

Customized Education. Teacher 4 wanted to differentiate because it would result in education that better meets the needs of students, providing more appropriate teaching: "You can provide much more suitable education: instead of always aiming for the average, you can actually teach everyone much more appropriately and thus better meet their needs. Yes, that's ideal."

(Teacher 4)

Student Preference. Teacher 7 was open to differentiation. This teacher would let the choice of a score be determined by the student.

Summary Differentiation. Most teachers were open to differentiation and the majority was willing to adjust scores both downward and upward. According to the teachers, differentiation helps maintain or increase motivation and reduce frustration. Differentiation is also seen as a way to provide personalized education that offers both challenge and support, depending on the individual needs of each student. The main reason teachers did not want to

make distinctions, or only wanted to do so to a limited extent, was the confrontational aspect of differentiation for students who find the material challenging.

Student Involvement

This question concerned whether the teacher would be open to students having a say in the score they aim to achieve. The teachers' responses are summarized in Table 6:

Table 6

Teachers Preferences Regarding Student Involvement When Setting Cut-off Scores

Docent	Answer
1	Only if they want to learn more than the standard percentage.
2	Both ways, but within a certain range.
3	Both ways, but within a certain range.
4	Both ways, but in consultation with the teacher.
5	Both ways, but in consultation with the teacher.
6	Both ways, but within a certain range.
7	Both ways, see previous answers.

Six out of seven teachers would allow students to have a say in the score they aim for, whether it is more or less than the standard percentage. The teachers mentioned that this should be done in consultation with the teacher or within a certain range. This way, the choice for the

score remains well-motivated, and students learn up to at least a minimum percentage. Below are some explanations for the teachers' responses:

Student Learning Experience: Motivation, Enjoyment in Learning, and Ownership. The main argument teachers gave for allowing students to have a say in the score-setting, was the positive influence they assumed it would have on the learning experience. Teacher 4 said: "It's actually nice if they have a bit of a say in this, they probably feel more ownership of the whole learning process." Teacher 2 also mentioned that being allowed to influence the score setting, enhances feelings of autonomy, which positively affects students' motivation:

> I think you can put some of the responsibility on the student themselves. I think they would feel more autonomy over the learning process. It's a bit of a feeling issue. You shouldn't give them too much freedom but give them a bit of a feeling that they make the choice themselves, I think that motivates them.
>
> (Teacher 2)

Teacher 6 indicated that being allowed to influence the score-setting, could lead to more pleasure in learning:

> Oh, I find that very interesting, I would like to experiment with that, see what happens, what they think, what they would do. I would be open to a pilot, and let's see what the effect is. And if it makes it more fun, because I think school isn't always very fun and you have to approach it in a fun way, then they get motivated and maybe this is the way.
>
> (Teacher 6)

Summary of Student Involvement. Most teachers are open to students having a say in the score they aim to achieve. One of the main arguments mentioned by teachers is the positive effect this involvement has on students' motivation and learning pleasure. According to the teachers, student involvement in their learning process leads to a greater sense of autonomy and ownership over their learning. However, teachers emphasize that this should happen within certain limits and in consultation to ensure the choices remain responsible. Teachers are willing to give students freedom in the score they aim for, provided they can guide the students in this process.

Replacing Tests

Finally, I asked whether teachers were open to the idea of the learning system replacing (some) traditional knowledge tests. I also asked if they agreed with students studying until they reach a score equal to a 6. The teachers' responses are summarized in Table 7:

Tabel 7

Teachers' Preferences Regarding Test-Replacement

Docent	Toets vervangen	6 voldoende
1	Yes	Yes
2	Yes	Yes
3	Yes	Yes
4	Yes	Yes
5	Yes	No
6	Yes	Yes
7	Yes	Yes

Note. Teachers indicated that to score a 6 on a traditional test, students usually would need to achieve scores of 60% correct for HAVO-tests and 75% for VWO-tests.

As seen in Table 7, all teachers agreed with the idea of the learning system replacing tests. Their arguments are summarized below:

Efficiency and Workload. Teacher 2 mentioned the desire to use such a system, as conducting and grading tests takes a lot of time. When tests are conducted with the help of a learning system and when they are automatically graded, teachers have more time for other important tasks and lesson preparation:

> Yes, very nice, I've thought about that too. Grading tests takes a lot of work. And a large part of that is just checking words and that feels inefficient. And if you don't have to worry about that as a teacher, you can focus on more important and useful things.
>
> (Teacher 2)

Keeping Up with Innovations. According to teacher 6, replacing traditional tests with a learning system is a good idea. Conducting only traditional tests can be considered outdated. Replacing such tests with an adaptive learning system means education keeps up with innovations. As long as students also find this way of working pleasant, teacher 6 saw no reason not to do it.

Test Score. When asked if students were allowed to learn until they score a 6, when a learning system replaces a test, six teachers agreed. Teacher 5 suggested aiming for a slightly higher score:

> No, that's strange, I think it should be higher. But why? Because a 6 is also just a 6 on a test? You'd hope it's presented in such a fun way that they'd want to score higher than a 6, but basically, a 6 is a 6. I think I'd want it to be higher, but not much. Around a 7.
>
> (Teacher 5)

Summary of Replacing Tests. All teachers are open to the idea of replacing traditional knowledge tests with learning through a system. Teachers mentioned reduced workload, increased efficiency, and this change being a sign of education keeping up with innovations. All teachers except one found it acceptable for students to aim for a score of 6. The reason one teacher wanted students to aim slightly higher than a 6 was that learning with such a system would be more enjoyable, motivating students to learn more.

Experiment

The experimental data were preprocessed and analyzed in R (version 4.2.2), using the *dplyr* (Wickham et al, 2022) and *ggplot2* (Wickham, 2016) packages. Preprocessing the data involved merging data from Qualtrics and Memorylab. The small sample size within this study prevented properly checking the assumptions for the statistical tests. Results of the statistical tests will therefore be interpreted with caution.

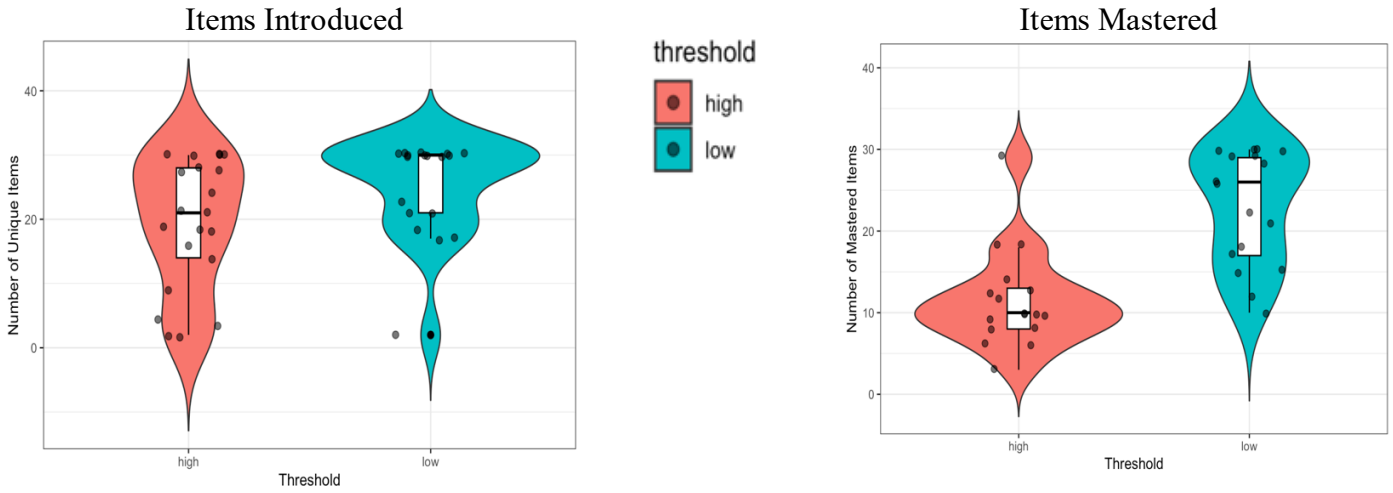
Learning Phase

One of the aims of this study was to gain insight in learning effects associated with different cut-off scores. In the method section we explained that the cut-off score employed, can influence the pace at which facts are introduced and mastered. Namely, the lower the cut-off score, the higher the expected pace of fact introduction and mastery. Analysis revealed that, on average, participants were indeed introduced to numerically more items in the low cut-off score

condition (M = 24.94, SD = 7.73) compared to the high cut-off score condition (M = 19.24, SD = 10.14) (see Figure 5a). A paired t-test indicated that this difference was statistically significant ($t(15) = 3.92$, $p < .01$). Additionally, on average more items were mastered in the low cut-off score condition (M = 22.82, SD = 7.047) compared to the high cut-off score condition (M = 11.52, SD = 5.98) (t (15) = -7.83, p < .01) (see Figure 5b).

Figure 5

Mean Number of Items Introduced (a) and Mastered (b) During the Learning Phase, Grouped by Threshold.



(a) (b)

Note. Every dot represents the mean number of items introduced for one participant in (b) or mean number of items mastered in (b). A black dot represents an outlier. The whiskers extend to the furthest data points within 1.5 times the interquartile range from the boxplot.

Reaction Times during the Learning Phase. Other measures indicative of learning effects are the mean reaction times associated with both cut-off scores. Namely, the faster the

reaction time, the better a fact is known. During the learning phase, mean reaction time across all practiced items was 5780.87 ms (*SD* = 4040.19) for the low threshold condition and 4145.76 ms (*SD* = 2300.081) for the high threshold condition. Results of a paired t-test indicated that the difference in reaction times between the two thresholds was non-significant ($t(15) = .18$, $p = .86$). To access the plots displaying the distribution of mean reaction times for both cut-off score conditions, see Appendix E.

Accuracy During Learning Phase. Accuracy scores are also an important indicator of how well facts are known. During the learning phase, mean accuracy across all practiced items was 82% (*SD* = 11) and in the low threshold condition this was 76% (*SD* = 16). A paired t-test indicated that the differences in mean accuracy scores were non-significant ($t = .56$, $p = .58$). To examine distributions of accuracy scores for both threshold conditions, see Appendix F.

Summary Learning Phase. In summary, the results from the learning phase provide no evidence that either mean reaction times or mean accuracy scores differed between threshold conditions in the learning phase. We did find evidence that the low cut-off score condition is associated with a higher pace of introduction and mastery of items, compared to the high cut-off score condition.

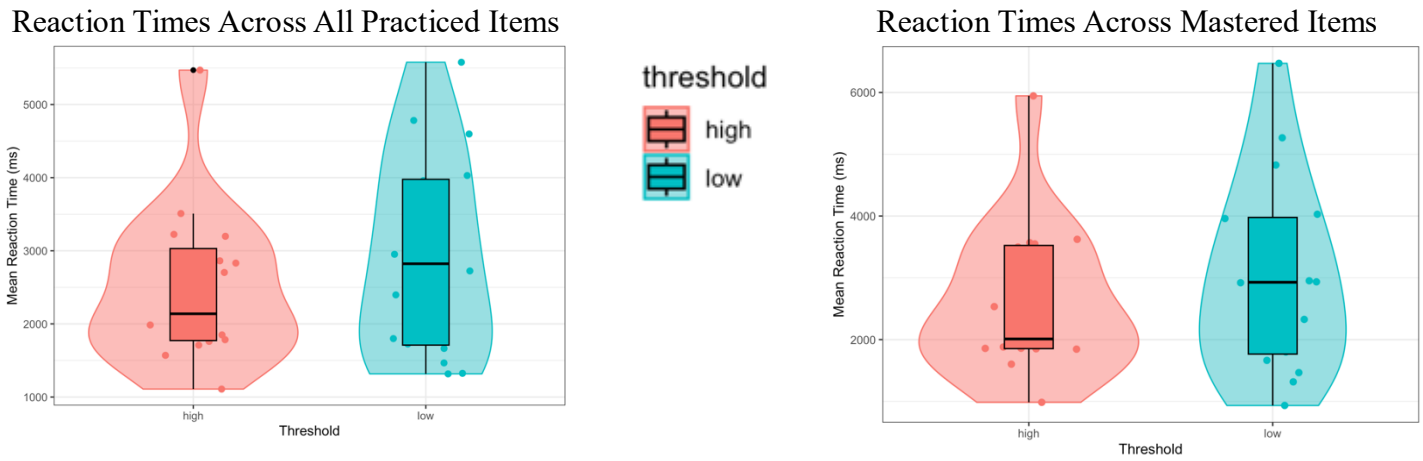
Test Phase

Reaction Time: Practiced Items. For the test data, reaction times were analyzed based on two different data subsets. The first subset included all items that had been practiced during the learning phase. Visualization of this data subset revealed that, variability in reaction times was greater for items practiced within the low threshold, compared to items practiced within the high threshold condition (see figure 6a). The mean reaction time for the low threshold was

2921.66 ms (*SD* = 1361.73) and for the high threshold 2513.86 ms (*SD* = 1083.67). Both a paired t-test and mixed linear effect model (MLE) indicated that the differences in mean reaction times were non-significant ($t(15) = 1.36$, $p = .097$; $\beta = 273.40$, $SE = 236.04$, $t = 1.16$, $p = .25$). Access Appendix G to examine all statistics and specifications of the MLE.

Figure 6

Mean Reaction Time During the Test Phase Across all Practiced Items (a) and Mastered Items (b), Grouped by Threshold.



(a) (b)

Note. Every dot represents the mean reaction time of one participant. A black dot represents an outlier. The whiskers extend to the furthest data points within 1.5 times the interquartile range from the boxplot.

Reaction Time: Mastered Items. To analyze reaction times during the test phase, we also created a data subset that included only those items that had been mastered during the

learning phase. Within this data subset, the median score of reaction times was higher in the low threshold condition, compared to the high threshold condition. Additionally, variability in reaction times was slightly higher in the low threshold condition (see Figure 6b). The mean reaction times in the low threshold condition was 4030.46 (*SD* = 2059.17) and for the high threshold condition this was 2602.41 ms (*SD* = 1073.28). Both a paired t-test and MLE-analysis indicated that the differences in mean reaction times were statistically significant ($t(15) = 2.73$, $p = .0076$; $\beta = 1307.44$, $SE = 519.82$, $t = 2.52$, $p = .012$). Access Appendix H to examine all statistics and specifications of the MLE.

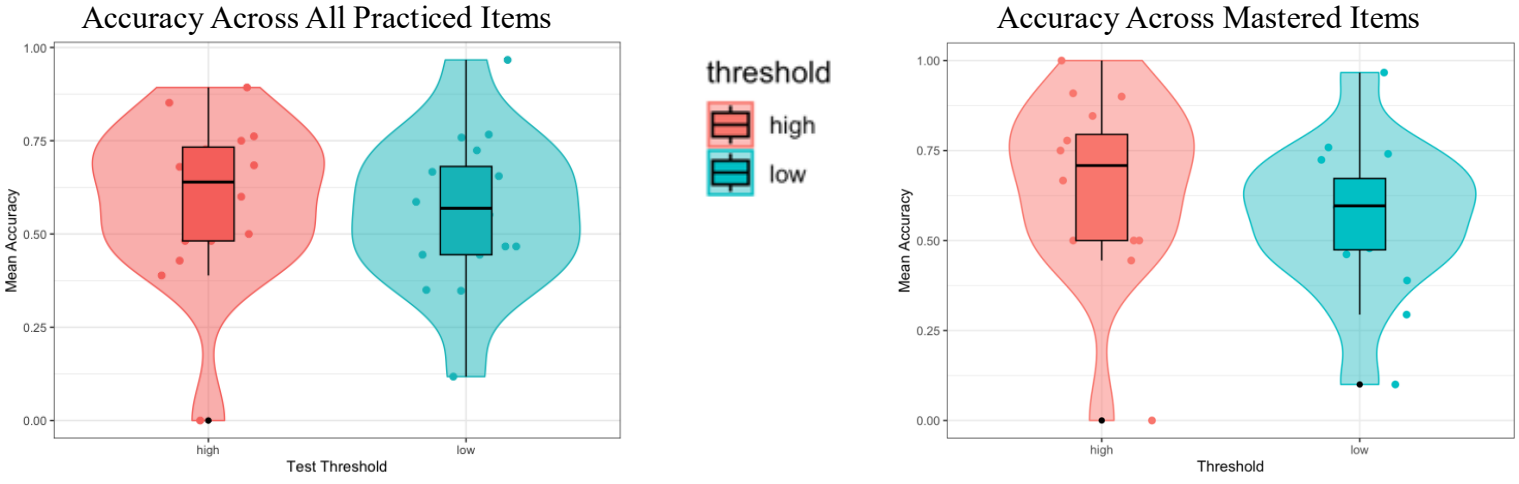
Summary Reaction Times Test Phase. In summary these results provide evidence that, when items are mastered, the high cut-off score condition is associated with lower reaction times, compared to the low cut-off score condition. This indicates that facts were remembered better when mastered in the high threshold, compared to the low threshold. This effect was not observed when items were merely practiced.

Accuracy Scores: Practiced items. To compare accuracy scores during the test phase, we first analyzed test data across all items that were practiced during the learning phase. Variability in accuracy scores across all practiced items was similar in both threshold conditions and the median score was slightly higher for the high threshold condition (see Figure 7a). Mean accuracy was 59% (*SD* = .22) in the high threshold condition and 56% (*SD* = .20) in the low threshold condition. A paired t-test indicated that the difference in accuracy scores between thresholds was non-significant ($t(15) = .94$, $p = .18$). Additionally, an MLE-analysis indicated that the difference in log-odds of answering correctly did not significantly differ between thresholds ($\beta = -.18$, $SE = .18$, $z = -1.022$, $p = .31$). Access Appendix I to examine all statistics and specifications of the MLE.

Accuracy Scores: Mastered items. In addition, we compared accuracy scores during the test phase across only those items that were mastered during the learning phase. Within this data subset, the median and variability were greater for accuracy scores within the high threshold condition compared to the low threshold condition (see figure 7b). The mean accuracy was higher in the high threshold condition (M = .66, SD = .24) compared to the low threshold condition (M = .56, SD = .20). A paired t-test indicated that this difference in mean accuracy scores was significant ($t(15) = 1.93$, $p = .036$). Additionally, an MLE-analysis indicated that the log-odds of answering correctly, significantly decreases when an item was mastered in the low threshold condition, compared to the high threshold condition ($\beta = -.49$, $SE = 0.25$, $z = -1.035$, $p = .042$). Access Appendix J to access all statistics and specifications of the MLE.

Figure 7

Mean Accuracy During Test phase across all Practiced Items (a) and Mastered Items (b), Grouped by Threshold.



(a) (b)

Note. Every dot represents the mean accuracy score of one participant. A black dot indicates an outlier. The whiskers extend to the furthest data points within 1.5 times the interquartile range from the boxplot.

Summary Accuracy Scores Test Phase. In summary these results provide evidence that the higher threshold is associated with higher accuracy and increased odds of answering correctly compared to the lower threshold condition. In other words, these results indicate that long-term knowledge retention is greater when items are mastered under the high threshold, compared to the low threshold. This effect was not observed when items were merely practiced.

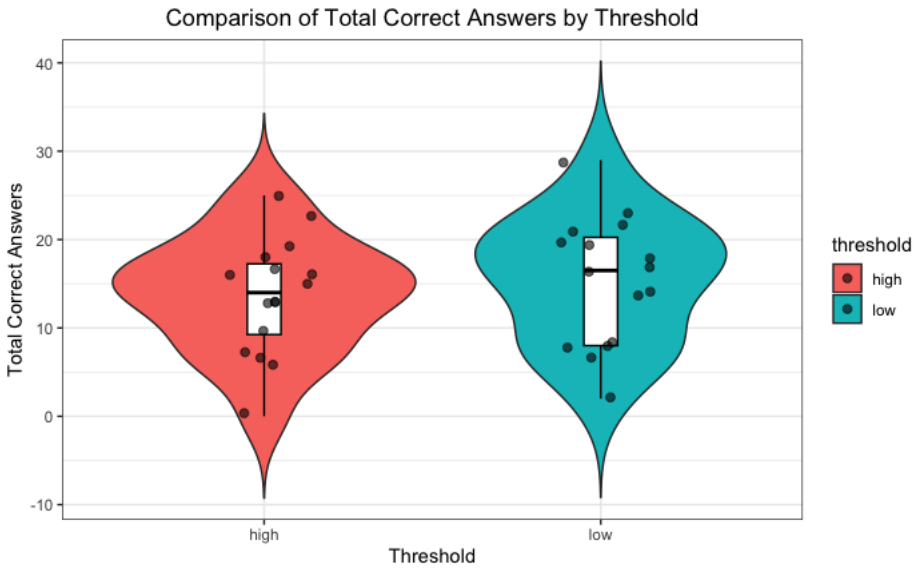
Learning Efficiency

In the previous sections accuracy scores were compared based on subsets of the test data. Namely, we only examined test data for items that were introduced or mastered during the learning session. In addition, we calculated accuracy scores by determining the proportion of correct answers within these subsets. However, these analyses lack some nuance, as they do not take into account that significantly more items were introduced and mastered in the low threshold condition compared to the high threshold condition. In order to provide a more comprehensive analysis, we also examined and compared total correct scores across all test items. This analysis revealed that, overall, the total correct scores were slightly higher in the low threshold condition, and variability of scores were similar across both conditions (see Figure 8). The mean number of correct answers was higher in the low threshold condition (M = 15.37, SD = 7.20), compared to the high threshold (M= 13.63, SD = 6.51). A paired t-test revealed that this difference was statistically significant ($t(15) = 1.86$, $p = .041$). These results likely reflect the

greater number of items practiced in the low threshold phase and indicate higher learning efficiency, compared to the high threshold.

Figure 8

Total Correct Scores for all Items on the Test, by Threshold



Note. Every dot represents the total correct score of one participant. The whiskers extend to the furthest data points within 1.5 times the interquartile range from the boxplot.

Discussion

The aim of this study was to investigate to what extend it is justifiable to consistently employ cut-off scores of 100% within ALSs and whether it would be desirable to employ more flexibility in these cut-off scores. To answer this question, we investigated teachers' opinions on this topic using a poll, questionnaire, and interviews. In addition, we conducted an experiment to analyze learning effects associated with different cut-off scores and compare these effects with teachers' preferences.

Main Findings

Teachers' Perspectives

In the poll, we found that only a small minority of the teachers would choose cut-off scores of 100%. The vast majority of participants indicated that they would prefer scores of 80%, representing the proficiency level of "good but not perfect".

The questionnaire results revealed no support among teachers for scores of 100%. Instead, teachers chose lower scores that were, on average, stricter for test preparation and test replacement compared to homework. In addition, the questionnaire revealed an openness of teachers to differentiation, but less so for test replacement than homework.

The interview data revealed that, according to teachers, school success is not only determined by academic achievements indicated by test results. Instead, teachers view the concept of school success more broadly and emphasize personal well-being, growth, as well as development of functional skills. The role of tests is seen as a tool for measuring individual progress and competence, and not as a comprehensive indicator of school success. This definition of school success is reflected by the cut-off scores the teachers judged as acceptable. Namely, among the teachers that were interviewed, all but one opposed the use of scores of 100%. Teachers explained that they would choose lower scores which they assumed would protect student motivation and well-being. At the same time, they want to ensure a minimum workable knowledge level and therefore they would maintain a minimum threshold when assigning scores. In line with the questionnaire results, the interview results reveal a tendency among teachers to set stricter scores for test preparation compared to homework. For test replacement, teachers indicated they would find it acceptable when students learn until they

reach a score that represents minimal passing grade on traditional tests. Most teachers were open to differentiation between students when setting scores and would allow students to have a say in the score setting.

In summary, all studies show that teachers generally opposed scores of 100%. In addition, according to teachers, one size does *not* fit all when it comes to score setting: what constitutes a suitable cut-off score depends on the learning task and characteristics of (groups of) student at hand.

Experiment

The results of our experiment provide evidence that mastering items based on different cut-off scores within Memorylab results in different performance levels. Namely, significant differences in accuracy scores and reaction times during the test phase indicate that learning based on the higher cut-off score was associated with higher levels of long-term knowledge retention. This finding suggests that Memorylab can effectively meet teachers' preference for differentiation in knowledge requirements when setting cut-off scores.

At the same time, learning efficiency was higher in the low cut-off score condition. Namely, after studying under both conditions for the same amount of time, on average, more items were mastered and introduced during the learning phase and the total number of items that was answered correctly on the test was higher for the low cut-off score condition compared to the high cut-off score condition.

In summary, the experimental results provide evidence that learning based on different cut-off scores is associated with main effects on both long-term knowledge retention and learning efficiency. These results demonstrate Memorylab can be applied in flexible ways,

thereby aligning with learning goals for various situations, as will be further exemplified in the sections below.

Choosing Cut-off Scores Lower than 100%

In this study, we found several reasons to adopt cut-off scores lower than 100%. To begin with, results from the poll, questionnaire and interview all revealed that teachers are generally opposed to scores of 100%. According to those teachers interviewed, setting lower scores is needed to protect student well-being and motivation. This means that, when setting cut-off scores, teachers are aware of and take into consideration, the possible negative effects of consistent employment of 100% scores, which were outlined in the introduction (Barbayannis, 2022; Dunford & Tamang, 2014; Pekrun, 2006). This overlap between the literature and teachers' preferences further validates and emphasizes the need for adopting cut-off scores below 100%.

Another reason to adopt lower cut-off scores is that our experiment provided evidence that lower cut-off scores are associated with higher learning efficiency. This indicates that learning based on lower cut-off scores allows for minimizing time spent on studying while maximizing learning gains. The time saved by learning based on a lower cut-off score could help to minimize stress and related issues among students, thereby supporting the goal of protecting student motivation and wellbeing. At the same time, due to its efficiency, the lower cut-off score could possibly still ensure that students understand most of the study materials. Therefore, learning based on a lower cut-off score could also ensure attainment another highly valued goal of teachers: acquirement of a minimum workable knowledge level.

Practical Recommendations

Teachers indicated varying expectations for acceptable cut-off scores under different circumstances, thereby complicating comparison of the experimental data with every score they indicated. The most often mentioned score is, however, 80%. The results of our experiment show that, while this score is not consistently achieved under either condition, the high cut-off score condition resulted in accuracy scores closer to 80% after mastering the material (see Figure 7b). Additional analysis on the distribution of scores confirmed that scores of 80% or above are more frequently achieved when items are mastered under the high cut-off score (see Table 8). This suggests that the high cut-off score would be more suitable, yet not sufficient, when teachers want students to learn until they master 80% of the materials.

Table 8

Number of Accuracy Scores Within a Specific Range, During the Test Phase and for Both Thresholds

	Accuracy Scores			
	0-60%	60-75%	75-80%	80-100%
Low Threshold	8	6	1	1
High threshold	6	2	2	6

Note. This table shows the total number of participants scoring within specific accuracy ranges for both low and high thresholds. Each participant is included twice: once for scores under the low threshold and once for scores under the high threshold.

When replacing traditional tests, teachers that were interviewed indicated that they would find it acceptable when students study until they score a minimal passing grade. Teachers explained that for HAVO, this means students would need to study until they would reach scores of around 60% and for VWO this would be around 75%. Previous analyses revealed that on the test, accuracy scores for mastered items center around 60% for the low cut-off score condition, and around 75% for the high cut-off score condition (Figure 7b). In fact, for the low cut-off score condition, exactly half of the accuracy scores were below 60% and half above. In the high cut-off score condition, exactly half of the mean accuracy scores were above 75% and half below (see Table 8). These results thus indicate that mastering items based on the low cut-off score could be suitable for test replacement in HAVO classes, while mastering items based on the high cut-off score could be suitable for test replacement in VWO classes.

Efficiency vs High Scores. Importantly, when choosing a cut-off score, it can be worth to consider the trade-off between efficiency and long-term knowledge retention associated with learning under both cut-off scores. For example, for those instances when teachers want to prioritize protecting student well-being or motivation, the low cut-off score might be more suitable due to its greater learning efficiency. Based on the research among teachers, examples of such situations might be students who struggle with the materials or students who are dealing with a difficult home situation. In addition, the low cut-off score might be more suited for learning tasks where close to perfect knowledge is not required and thus extensive time investment is not justified. An example of such a learning task would be homework, as teachers chose lower scores for homework because, for example, understanding of the materials could still be improved in class.

The high cut-off score might be more suitable for those situations when time or effort spend on the task is of secondary concern and maximized long term knowledge retention is preferred. In addition, the high cut-off score might be more suitable when teachers want to challenge students. Examples of such situations are when cut-off scores must be set for students with high subject proficiency. Another example is that of learning tasks with higher expected knowledge acquisition, like test preparation or replacement.

Strengths and Limitations

This study had several strengths. To begin with, to our knowledge, this was the first study that addressed the question whether it is justifiable to consistently employ cut-off scores of 100%. Therefore, with this study we have provided unique insights and perspectives regarding cut-off scores in adaptive learning systems.

Another main strength of this study was that we collected and connected data on teachers' perspectives and experimental data to answer our research questions. Namely, we used the data on teachers' perspectives to evaluate the experimental data. Based on this evaluation, we provided practical recommendations. This way, we allowed science and practice to communicate and exchange relevant information, thereby enhancing the practical relevance of our results and contributing to bridging the gap between science and practice.

This study also faced several limitations. To start with, to understand why teachers choose certain cut-off scores, we only interviewed language teachers. Our findings might have been different if we had also interviewed teachers from other subjects. For example, a language teacher may accept a student learning less than 100% of a vocabulary list because they believe lower scores can still result in a workable knowledge level. For instance, one teacher indicated

that a score of 70% would be sufficient because a student would be able to "deduce the meaning of other words from the context". It is possible that teachers from certain other subjects, view their learning materials highly interconnected and less suited for partial understanding. It is also possible that teachers from most other subjects accept scores below 100% but would provide different reasons to believe so. Therefore, the results of the interviews should be interpreted as primarily applicable to language teachers and generalizable only to a limited extent. To address this issue, future studies should conduct interviews among teachers of subjects other than foreign languages.

For the experiment, one limitation was that all participants were first-year university students. It is possible that the learning effects we found within this sample differed from learning effects that would be observed among other student population such as high school students. Therefore, the extent to which our experimental results can be generalized is also limited. To prevent this limitation, future studies should include samples from a different student populations.

Another limitation related to the experiment was the relatively small sample size (n = 16), which complicated checking assumptions for the statistical tests. Both the small sample size and running the statistical test without assumption checking might have resulted in over- or under estimation of the effect size, increased probability of type I and type II errors, or limited reproducibility of the results (Osbourne & Waters, 2002; Button, 2013). To strengthen reliability and validity of the experimental results, future studies should aim for a sufficiently large sample size.

Conclusion

The overall results of this study do not provide support for consistent employment of cut-off scores of 100%. Namely, teachers generally opposed to cut-off scores of 100%. Instead, teachers prefer lower cut-off scores that they assume would result in a balance between protection of students' well-being and motivation and a guarantee of minimum workable knowledge levels. What cut-off scores are deemed acceptable, depends on the learning task and student inquired about. The research among teachers thus not only highlights the need for implementation of scores below 100%, but also for considerable flexibility when setting these scores. The experiment's results further underscore the need, benefits and justification for adopting lower and more flexible cut-off scores. Namely, mastering items based on different cut-off scores was associated with significant differences in knowledge retention, indicating Memorylab can meet teachers' preference for differentiation in knowledge requirements for different learning tasks. In addition, learning based on lowered cut-off scores was associated with higher learning efficiency. Learning based on a lowered cut-off score can therefore help to reduce time spent on studying and reduce stress, while likely still ensuring a workable level of knowledge. In conclusion, this study provides substantial evidence that employing lower and flexible cut-off scores, as opposed to fixed 100% cut-off scores, would represent a significant advancement for educational practices.

Acknowledgements

While writing this paper I utilized ChatGPT.4 (OpenAI, 2023). I used it to generate scripts for R and to translate the interview guide and results, and the questionnaire from Dutch to English. In addition, I used it as a language tool to fine-tune my own writings. Namely, when in doubt, I would ask the model to confirm whether I used the English language correctly (with regards to spelling, grammar or interpunction). Sometimes, I would ask the model to provide me with the English equivalent for Dutch expressions or phrases and how I could prevent excessive repetitions of certain words or phrases. Also, I would use the model to provide feedback on the structure of paragraphs that I found challenging to write clearly and coherently.

References

Angoff, W. H. (1971). barba. In R. L. Thorndike (Ed.), Educa- tional measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Edu- cation.

Anki - powerful, intelligent flashcards. (n.d.-c). https://apps.ankiweb.net/

Barbayannis, G., Bandari, M., Zheng, X., Baquerizo, H., Pecor, K. W., & Ming, X. (2022). Academic Stress and Mental Well-Being in College Students: Correlations, affected groups, and COVID-19. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.886344

Braam, M. (2024, 12 april). *MemoryLab's new Model-Based Mastery algorithm helps students learn faster*. MemoryLab. https://www.memorylab.nl/en/blog/memorylabs-new-model-based-mastery-algorithm-helps-students-learn-faster/

Brandon, P. R. (2004). Conclusions about frequently studied modified ANGOFF Standard-Setting topics. *Applied Measurement in Education*, *17*(1), 59–88. https://doi.org/10.1207/s15324818ame1701_4

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, *1*(4), 309–330. https://doi.org/10.1007/bf01320097

Dunford. R., Su, Q., and Tamang, E. (2014) 'The Pareto Principle', The Plymouth Student Scientist, 7(1), p. 140-148.

Dutch grading system. (n.d.). Erasmus University Rotterdam. https://www.eur.nl/en/dutch-grading-system

Eckes, T. (2012). Examinee-Centered Standard Setting for Large-Scale Assessments: The Prototype Group Method. *Psychological Test And Assessment Modeling*, *54*(3), 257. http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2012_20120925/03_Eckes.pdf

Ennouamani, S., & Mahani, Z. (2017b). An overview of adaptive e-learning systems. International Conference on Intelligent Computing and Information Systems. https://doi.org/10.1109/intelcis.2017.8260060

Guterman, O. (2020). Academic success from an individual Perspective: A proposal for redefinition. International Review of Education, 67(3), 403–413. https://doi.org/10.1007/s11159-020-09874-7

Katsaris, I., &; Vidakis, N. (2021b). Adaptive e-learning Systems Through Learning Styles: A review of the literature. Advances in Mobile Learning Educational Research, 1(2), 124–145. https://doi.org/10.25082/amler.2021.02.007

McIntyre, S. H. & Munson, J. M. (2008). Exploring cramming: Student behaviors, beliefs, and learning retention in the principles of marketing course. Journal of Marketing Education, 30(3), 226–243.

Mohajerzad, H., Martin, A., Christ, J., & Widany, S. (2021). Bridging the gap between science and practice: research collaboration and the perception of research findings. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.790451

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Osbourne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, *8*(2), 2. https://doi.org/10.7275/r222-hv23

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review, 18(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Penn, P. (2019). *The Psychology of Effective Studying: How to succeed in your degree*. https://openlibrary.org/books/OL29441597M/Psychology_of_Effective_Studying

Qualtrics XM (2024). [Computer software]. Qualtrics and all other Qualtrics product or servicenames are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. https://www.qualtrics.com

Sense, F., Behrens, F., Meijer, R.R., Van Rijn, H., 2016. An individual's rate of forgetting is stable over time but differs across materials. Top. Cogn. Sci. 8 (1), 305–321.

Sense, F., Van Der Velde, M., & Van Rijn, H. (2021). Predicting university students' exam performance using a Model-Based Adaptive Fact-Learning system. *Journal of Learning Analytics*, *8*(3), 155–169. https://doi.org/10.18608/jla.2021.6590

Simon, D. A. & Bjork, R. A. (2001). Metacognition in motor learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(4), 907–912

Anki Manual. (n.d.). https://docs.ankiweb.net/studying.html

Van Rijn, H., Van Maanen, L., Van Woudenberg, M., 2009. Passing the test: Improving learning gains by balancing spacing and testing effects. In: Proceedings of the 9th International Conference of Cognitive Modeling, Vol. 2. pp. 7–6.

Wickham, H., François, R., Henry, L., & Müller, K. (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.9. https://CRAN.R-project.org/package=dplyr

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org

Appendix A

Questionnaire Content

Thank you for taking the time to fill out this questionnaire. While completing this questionnaire, you are asked to imagine the following: there exists a software program that allows you to view, control, and guide the learning process of students. With the help of this software, you can make students learn until they demonstrate a certain level of knowledge of the learning material. For example, you can make students learn until they know 15 out of 20 facts (75% of the specified material). Please indicate for the following two situations to what percentage (0 - 100%) you think students should learn:

When learning is part of the homework for the next lesson: …….%

When learning is in preparation for a test: ……..…%

For the software described above, achieving the last 10-20% can take relatively much effort and time. This is also known as the Pareto principle and applies especially to students for whom the material is challenging. Based on this information, would you:

3. Distinguish between students in how well they should learn the material when learning is part of the homework? Please circle your answer: yes/no. If yes, how would you make this distinction (which percentage would apply to which type of student):

……..……..……..……..……..……..……..……..……..……..……..……..…
……..……..……..……..……..……..……..……..……..……..……..……..

4. Distinguish between students in how well they should learn the material when they are learning in preparation for a test? Please circle your answer: yes/no. If yes, how would you make this distinction (which percentage would apply to which type of student):

……..……..……..……..……..……..……..……..……..……..……..…….
……..……..……..……..……..……..……..……..……..……..……..…

5. If you knew for sure that this software provides a reliable picture of a student's knowledge level, would you agree to replace (some) knowledge tests with learning through this software? Please circle your answer: yes/no. If yes, to what percentage should students learn in this situation: …….%

6. If you would like to provide an explanation for one or more of your answers or make another comment, you can do so here:

……..……..……..……..……..……..……..……..……..……..……..……..…
….……..……..……..……..……..……..……..……..……..……..……..…

…….
…….

Appendix B

Interview Content

******Intro*** ***

Introductory Questions:

1. How many years of experience do you have as a teacher:
   …….
2. In what year did you graduate from teacher training college?:
   ……..……..……
3. Which subject(s) do you
   teach?:…….…….…….……
4. Which classes do you
   teach?:…….…….……..
5. You teach at a Montessori school
   o Did you consciously choose to teach at a Montessori school?
   · Why?

Middle Section

*** Now I would like to address the main topic of the interview and discuss learning programs and related topics with you ***

6. First, I would like to briefly discuss the meaning of school success, as it presumably plays a significant role in designing learning programs. Various indicators/definitions of school success are described in the literature. Examples include grades, the process, and the student's own opinion. There is no consensus on the definition of school success. What is your definition of school success?
   ……
   ……

……
……

7. What is your view on testing?
   o Purpose?
   o What does a test look like, what is the content (insight/application)?
8. Have you ever used learning software?
   o How?

Main Questions – Cut-off Scores

9. I would like to hear your opinion on what optimal learning programs look like for a number of specific situations. I would like to present a hypothetical situation to you: Suppose you could, with the help of software, see, control, and direct how students learn. You could make students learn until they know, for example, 60, 80, or 100% of the material. [give example with 20 facts/words]

With this in mind, at what percentage would you say a student has learned enough: when can the student stop learning? I would like to ask you to base your answer on the following situations:

10. Learning for homework (in preparation for a lesson)?
To what percentage do you think students should learn?
……
……
……

Why do you choose this percentage?
……
……
……

Points to consider when probing further:
o Advantages/disadvantages of lowering cut-off scores
o Why (not) up to 100%?
o How do you think this affects students' motivation?

11. Learning as preparation for a test?
To what percentage do you think students should learn?
……
……
……
· Why do you choose this percentage?
……
……
……

Points to consider when probing further:
o Advantages/disadvantages of lowering cut-off scores
o Why (not) up to 100%?
§ (Why up to 100% when 100% is not necessary on the test itself?)
o How do you think this affects students' learning outcomes and motivation?

12. Learning for test and/or homework: if a score < 100% is mentioned:
And if students reach this percentage after 10 minutes? Can they stop learning then?
…….
……...…….. .

How long should a student learn at a minimum before they can stop learning?
……
……

13. Learning the last 10 to 20% can take students relatively much time [explain: Pareto principle/law of diminishing returns]. This applies especially to students for whom the material is challenging. Based on this, would you make a distinction between students regarding how well they should learn the material? If so, how?
……
……
……..…

Does this differ between test/homework?
……
……
……
· Why do you choose these percentages?
……
……
……

Points to consider when probing further:
o Advantages/disadvantages of distinguishing between students
o Influence on motivation and learning outcomes?
……
……
……..
o For homework/test, do you think students should be able to help decide up to what score they learn?
……
……
……
……

10. If you were certain that this software provides a reliable picture of a student's knowledge level, would you agree to replace (some) knowledge tests with learning through this software?
    ……
    ……
    ……
11. 8 And would you then find it acceptable for students to learn up to a 6?
    ……
    ……
    …… 8.2 A 6 is sufficient to pass a regular test. What makes you set different requirements when a student demonstrates knowledge using this software?
    ……
    ……
12. What would you think if such methods were implemented in your own lessons?

Closing Questions

11. Are there any matters I should come back to?
    Specific statement:
    ……
    ……..
    Specific story:
    …….
    ……..
    Contradictions:
    ……
    ……..
    check: do I have percentage answers everywhere?
12. Do you have any further questions or comments about learning goals (the percentages) or other topics we have discussed? A thought or opinion you want to share? An important point I did not ask about?
    ……
    ……
    ……
    ……

*** closing: ***

Appendix C

Content of Fact Set A and B

Set A:

cue	correct answer
Animals can become fearless after damage to the _____.	amygdala
Benzodiazepines are a treatment for: _____ disorder.	anxiety
Alcohol and benzodiazepines activate _____ receptors.	gaba
Frequent anxiety attacks are a symptom of: _____ disorder.	panic
Hormone that is produced when we experience stress: _____.	cortisol
Stress activates the sympathetic nervous system and the _____ axis.	hpa
Activation of the frontal and temporal lobes of the left hemisphere is associated with _____.	approach
The ability to recover from a traumatic experience: _____.	resilience
The physical (neural) representation of a learned association: _____.	engram
Traditionally, emotions were linked to activation of the _____ system.	limbic
Things that might cause illness evoke an emotion called _____.	disgust
The process of strengthening a memory representation: _____.	consolidation
Comparing the utilitarian and emotional consequences of a decision: _____ prefrontal cortex.	ventromedial
Episodic memory relies on the _____.	hippocampus
Androgen linked to aggression: _____.	testosterone
Procedural memory relies on the _____.	striatum
Inability to form new memories: _____grade amnesia.	antero
Inability to recall memories from past the past after brain injury: _____grade amnesia.	retro
Memory for factual information: _____ memory.	semantic
Influence of experience on behavior that can not be put into words: _____ memory.	implicit

Cells in the hippocampus that represent where we are in a certain environment: _____ cells.	place
A decrease in response to a repeated stimulus: _____.	habituation
A burst of intense stimulation of a post-synaptic cell causes: long-term _____.	potentiation
The formation of new synapses in the hippocampus relies on two kinds of _____ receptors.	glutamate
Most of the information that goes from one hemisphere to the other does so by passing through which brain structure?	corpus callosum
Which term refers to the functional asymmetries in the brain?	lateralization
The planum temporale is located in the _____ lobe	temporal
Children with which syndrome are characterized by good language skills despite low general intelligence?	Williams syndrome
Non-fluent aphasia, where the individual cannot speak fluently, is due to brain damage in the area of _____.	Broca
Fluent aphasia, where the individual has difficulty understanding language but can speak smoothly, is due to damage in the area of _____	Wernicke

Set B:

cue	correct answer
A brain proces that takes place regardless of environmental conditions: _____.	endogenous
A stimulus that can reset the circadian rhythm: _____.	zeitgeber
Neurons in the SCN control the circadian clock through the synthesis of _____.	proteins
In diurnal mamals like humans, melatonin _____ sleepiness.	increases

An animal that is awake during the night and at sleep during day: _____.	nocturnal
The reason why all species of animals sleep is it that helps conserve _____.	energy
The reason why we are not conscious during slow-wave sleep is the release of _____.	gaba
The EEG of someone who is relaxed and awake shows _____ waves.	alpha
A predictor of memory consolidation during sleep: sleep _____.	spindles
Narcolepsy is caused by a lack of cells that produce _____.	orexin
During REM sleep, the muscles of the body are _____.	relaxed
Acetylcholine causes _____ of the brain.	arousal
If we eat less, our metabolism eventually goes down and our weight stays the same: _____stasis.	homeo
The body and brain anticipate and try to prevent changes in bodily variables: _____stasis.	allo
The energy used to maintain a stable body temperature: _____ metabolism.	basal
An animal that generates its own body heat: _____thermic.	endo
Registration of temperature in the brain takes place in the _____ area.	preoptic
An increase in the set point for body temperature: _____.	fever
Hormone that increases blood pressure by constricting blood vessels: _____.	vasopressin
Thirst after eating something salty: _____thirst.	osmotic
The hormone that indicates the amount of fat cells in the body: _____.	leptin
The hormone that allows glucose to enter the cells of the body: _____.	insulin
Hunger can result from a hormone released by stomach contractions: _____.	ghrelin
Our feeding behaviour is controlled by a network of nuclei in the _____.	hypothalamus
SRY causes the undifferentiated gonads to develop into _____.	testes
The category of hormones that are more abundant in males: _____.	androgens

The category of hormones that are more abundant in females: _____.	estrogens
Long-lastig, structural effects of sex hormones are called _____ effects.	organizing
The menstrual cycle is controlled by hormones released by the _____ gland.	pituitary
Temporary effects of sex hormones are called _____ effects.	activating

Appendix D

Interview Results in Dutch

Schoolsucces en Toetsen

De eerste hoofdvraag betrof wat docenten verstaan onder "schoolsucces" en welke rol toetsen spelen bij het evalueren van schoolsucces. De antwoorden van docenten zijn hieronder samengevat.

Persoonlijke Ontwikkeling en Welzijn. Drie docenten benoemenden kenmerken van persoonlijke ontwikkeling als indicatoren van schoolsucces. Docent 7 benadrukte daarbij dat het belangrijk is dat studenten zich comfortabel voelen en school en hun plek in de maatschappij ontdekken:

> Ik vind het wel belangrijk dat het op meerdere vlakken gebeurt. Ik denk niet dat het alleen maar om het didactische gedeelte gaat: dus het hoeft niet alleen maar cijfers te zijn. Ik vind het ook heel erg belangrijk dat ze zich fijn voelen en […] dat ze weten wat hun plek in de maatschappij is en in de samenleving zal zijn en hoe ze daar actief aan kunnen deelnemen.
>
> (Docent 7)

Ook docent 4 benoemde dat het belangrijk is dat leerlingen zich prettig voelen op school en dat cijfers niet de enige indicator zijn van schoolsucces:

> Ik denk dat het voor een leerling altijd fijn is als het goed gaat op school, cijfermatig gezien, maar ik denk dat het nog veel belangrijker is dat een leerling gelukkig is in de klas, los van de cijfers, dat hij een goede band heeft met klasgenoten, het naar zijn zin

> heeft en zich op zijn plek voelt. En ik denk dat docenten daar wel een grote rol bij spelen. Als een leerling zich op zijn gemak voelt bij docenten, dan kan dat ook heel erg bijdragen.
>
> (Docent 4)

Docent 1 benoemde het belang van volwassen worden en het ontwikkelen van de sociale vaardigheden die daarbij horen: “Meer het volwassen worden: leren plannen, met mensen omgaan en respect tonen en sociale vaardigheden ontwikkelen.”

Procesgerichte Evaluatie. Twee docenten gaven aan dat het leerproces van een leerling een belangrijke indicator is van schoolsucces:

> Wij hebben op een bepaald moment een training gehad over het concept van de “growth mindset”. Dat ging erover dat leerlingen soms te veel gefocust zijn op cijfers, terwijl je als docent juist zou willen dat ze een growth mindset ontwikkelen, zodat ze zichzelf beoordelen op basis van hoeveel ze zijn verbeterd sinds de vorige keer en wat ze hebben gedaan om beter te leren dan voorheen. Ik denk dat als je als docent kunt bijdragen aan die mindset, dat dan het hoogst haalbare is bereikt. Dit betekent niet dat cijfers onbelangrijk zijn, maar ze zijn misschien wat ondergeschikt aan die mindset. […] Cijfers van toetsen kunnen wel helpen om een leerling te vergelijken met zichzelf.
>
> (Docent 2)

Volgens docent 3 wordt er soms te veel gefocust op toetsen, waardoor het proces minder aandacht krijgt, terwijl deze juist meer indicatief kan zijn voor schoolsucces. Toetsen kunnen volgens docent 3 wel helpen bij het identificeren van aandachtspunten:

> Ik denk ook meer aan het proces, dat is ook meer op Montessori. In plaats van dat je alleen maar bezig bent met toetsen. Dat merk ik bij Duits ook, dat je alleen maar bezig bent met die toets, dat die erdoorheen dendert, terwijl leerlingen het eigenlijk nog niet begrepen hebben, maar dat je dan alweer bezig moet met het volgende onderwerp terwijl je eigenlijk meer moet focussen op het leerproces van: waar gaat het fout, heb je het nu wel onder de knie? […] Bij zo'n proces moet je natuurlijk wel iets behalen met toetsen. Werken met een doel, anders ben je natuurlijk voor niets bezig, maar je kunt dan meer het product vormgeven in plaats van dat het losse componenten zijn.
>
> (Docent 3)

Functionele Bekwaamheid. Twee docenten benoemde functionele bekwaamheid als belangrijke indicator van schoolsucces. Deze docenten onderstreepten namelijk het belang van het kunnen werken met de kennis die je opdoet tijdens je opleiding:

> Persoonlijk vind ik dat het meer gaat over de ervaring van de leerling zelf. Bij de opleiding waar ik heb gewerkt, werkten we meer met portfolio's en dan moest je aantonen dat je begrip had van bepaalde onderwerpen en of je dat onder de knie had of niet. Dit deden we ook wel aan de hand van interviews, en dan moest je laten zien wat je wist. Ik vind dat als je kunt bewijzen dat de theorie die voorbij is gekomen in de les, en dat je die kunt verwerken in een verhaal en kunt begrijpen, dan ben je behoorlijk succesvol geweest.
>
> (Docent 6)

Ook docent 5 vond het belangrijk dat leerlingen het geleerde kunnen toepassen. Docent 5 legde daarbij de nadruk op leerdoelen binnen het MBO, waarbij de focus ligt op voorbereiden op de beroepspraktijk:

> In het mbo is het echt zo dat studenten getraind worden zodat ze zich kunnen redden in de beroepspraktijk, met de vreemde taal die ik doceer. Dus ik zou zeggen, als ze professioneel uit de voeten kunnen met het Engels dat ze hebben geleerd, dan lijkt mij dat ze succesvol zijn geweest in het afronden van hun opleiding en het klaarstomen voor de beroepspraktijk. Ze kunnen zich beroepsmatig redden, en dan zitten ze meestal wel op een B1 niveau ongeveer.

Toen ik vroeg of deze docent toetsen gebruikt om vast te stellen of leerlingen op niveau B1 zitten, gaf deze het volgende aan:

> Ja, meerdere in ieder geval, examens dus. Ze moeten 5 generieke examens doen: lezen, luisteren, spreken, gesprekken voeren en schrijven. Voor die examens toetsen we elke periode inderdaad gewoon een bepaald onderdeel van Engels en dat gaat van basis naar steeds meer gevorderd.

(Docent 5)

Samenvatting Schoolsucces en de Functie Toetsen. Docenten benaderen het concept van schoolsucces breder dan enkel academische prestaties en nadrukken persoonlijke en welzijn en het aanleren van functionele vaardigheden. De rol van toetsen wordt gezien als een hulpmiddel bij het meten van individuele vooruitgang en praktische bekwaamheid, en niet als een allesomvattende indicator van schoolsucces.

Cut-off Scores Huiswerk

De tweede hoofdvraag bedroeg welke cut-off score docenten zouden aanhouden voor leren als onderdeel van het huiswerk. Iedere docent gaf aan scores beneden de 100% acceptabel te vinden. Het verschil tussen de hoogst (80%) en laagst (50%) genoemde score is 30%. De gemiddelde score is 71%.

Tabel 3

Cut-off Scores Wanneer het Leren Onderdeel is van Huiswerk

Docent	Cut-off Score
1	80%
2	75%
3	65%
4	50%
5	80%
6	70%
7	75%

Ik vroeg de docenten uit te leggen waarom zij voor bovengenoemde scores kozen en waarom niet voor een hogere of lagere score. Thema's in antwoorden zijn hieronder weergegeven:

Leerervaring van de Leerling: Motivatie, Stimulatie en Leerplezier. Vier docenten noemden de leerervaring van leerlingen als argument voor het kiezen van de cut-off scores. Zo gaven twee van hen aan dat gebruik van hogere cut-off scores, dan door hen gekozen, zouden kunnen leiden tot een verlies van motivatie of leerplezier onder studenten:

> Ik denk dat hogere percentages (dan 70%) zorgen voor meer prestatiedruk, en dat kan averechts werken, omdat je dan vaker geconfronteerd wordt met het feit dat je er nog niet bent. Succes moet sneller behaald kunnen worden, zodat het leren leuker is en blijft.
>
> (Docent 6)

Ook een andere docent gaf aan dat het hanteren van een hogere cut-off score kan leiden tot verlies van leerplezier:

> Ik denk dat sommige leerlingen woordjes leren lastig vinden terwijl ze wel goed zijn in andere aspecten van de taal. Een woordjes toets geeft niet goed aan hoe goed ze zijn in algemene zin. 75% is dan een compromis, anders straf je leerlingen te hard die wel goed zijn in Engels maar minder goed in woordjes stampen. […] Dat kan ertoe leiden dat sommige leerlingen die wel goed zijn in Engels, maar niet in woordjes, het niet meer leuk vinden terwijl ze het in eerste instantie wel leuk vonden.
>
> (Docent 2)

In tegenstelling tot de hierboven genoemde argumenten, wees docent 4 erop dat juist het gebruik van te lage cut-off scores een negatief effect kan hebben op de leerervaring van een leerling. Deze docent hanteerde een basis cut-off score van 50%, die al lager was dan die van andere docenten, maar gaf een duidelijke reden om niet lager te gaan:

> Kijk, die 50 % gaat in de lessen nog wel omhoog naar een acceptabel percentage. En je mag best wat verwachtingen hebben van de leerling en als de lat dan zo laag wordt gelegd, dan stimuleer je ze ook niet echt om wel hun best te blijven doen.
>
> (Docent 4)

Ruimte voor Verbetering Tijdens de Les. Drie docenten gaven aan bovengenoemde score te kiezen (en niet hoger), omdat er tijd bestaat tijdens de les om dit percentage omhoog te brengen. Zo gaf participant 3, die een score van 65% koos, het volgende aan: "Ik geef normaal gesproken geen huiswerk en ze kunnen de rest in de klas behalen". Zoals eerder omschreven gaf ook docent 4 aan dat een score van 50% acceptabel is, omdat deze in de les omhoog kan worden getrokken naar een hoger percentage.

Leren van "Fouten". Twee docenten gaven aan voor bovengenoemde scores te kiezen (en geen hogere score), omdat fouten mogen maken, het leerproces juist kan verbeteren:

> Als je fouten mag maken, heb je de kans om bij de docent of om bij klasgenoten na te gaan van 'hey waarom is dit nou eigenlijk zo?', waardoor ze hun kennis wat meer verbreden. Als ze tot 100% gaan en ze zijn daar klaar mee dan zijn ze er in hun gedachten ook klaar mee en denken ze 'Oh mooi, klaar. Volgende onderwerp.' […] Maar als ze tot 80% leren en ze hebben een aantal fout en ze werken dan bijvoorbeeld met meerdere grammaticale onderdelen, dan kun jij In de les gaan kijken van oh kijk, jij vindt deze vorm juist lastig en dan kun je ook echt gaan kijken van nou, waarom was dit fout? Dus daarom niet 100%, want als het 100% is dan laten ze het liggen. Dan denken ze namelijk dat ze het allemaal weten.
>
> (Docent 1)

Minimaal Werkbaar Kennisniveau voor Toets, Les en Praktijk. Verschillende docenten gaven aan voor de bovengenoemde cut-off scores te kiezen, omdat deze scores een minimaal werkbaar kennisniveau representeren. Zo gaven twee docenten aan dat de door hen genoemde score minimaal nodig is voor het halen van de toets:

> Als ik kijk naar toetsen dan zijn dat 4 x 8 woordjes, dan vind ik een goede score 1-2 fouten per opdracht, dan zit je op 24 woorden van 32 goed, dus 75% goed. Dus dat moet ook voor het huiswerk. Want de toets is een optelsom van alles voor het huiswerk, en als leerlingen dit halen dan zullen ze ook een voldoende halen op de toets. En je ziet vaak wel of een leerling goed is voorbereid of niet, en goed voorbereid zijn houdt gewoon in dat de toets ook met een grotere waarschijnlijkheid goed komt.
>
> (Docent 2)

In lijn met docent 2, gaf docent 5 aan bang te zijn dat leren tot een lagere score (dan 80%) het risico vergroot dat leerlingen de toets niet halen:

> Toetsen bestaan voor een gedeelte uit woordjes en gedeelte grammatica, maar als je de woordjes niet kent dan ken je vaak de grammatica ook niet, meestal hebben ze dan sowieso niet goed geleerd. Maar meestal als je de woordjes niet kent, dan heb je te weinig om te compenseren. Het is een te groot risico.

Naast het kunnen halen van een toets, benoemden twee docenten het belang van mee kunnen komen tijdens de les. Docent 3 zei het volgende: “er moet een basis worden aangelegd waarop ze kunnen voortbouwen in de les. Deze basis van 65% is een basis waarmee je de lesstof kan toepassen of verrijken.”

Tot slot werd ook het kunnen toepassen van de opgedane kennis in de praktijk genoemd als reden om voor een specifieke score te kiezen:

> Als je tot 70% leert dan heb je het overgrote gedeelte onder de knie en dan kan je de betekenis van andere woorden wel uit de context halen. […] Ik vind het wel een heel mooi streven als je het wel redt om meer te leren, maar ik ben zelf niet van mening dat je

> altijd alle woorden moet kennen. Als het gaat om Engels dan vind ik dat je moet kunnen communiceren met iemand, en dat kan ook met een iets minder deel. Als je kan communiceren dan ben je al aardig op weg.
>
> (Docent 6)

Samenvatting Cut-off Scores voor Huiswerk. Samengevat kiezen docenten voor scores waarvan zij geloven dat deze resulteren in een balans tussen het behouden van motivatie en voldoende beheersing van het lesmateriaal. Optimale scores laten daarbij ruimte voor verbetering tijdens de les, helpen bij het voorbereiden op toetsen, en zijn voldoende voor praktische toepassing van de leerstof. Docenten willen de motivatie en het leerplezier van de studenten beschermen door niet te kiezen voor zeer hoge cut-off scores. Tegelijkertijd vermijden ze te lage scores om leerlingen uit te dagen en om een werkbaar kennisniveau te waarborgen.

Cut-off Scores Toets

Alle docenten, op één na, gaven aan scores beneden de 100% acceptabel te vinden, wanneer het gaat om leren als voorbereiding op een toets (zie tabel 4). Het verschil tussen de hoogst (100%) en laagst (60%) genoemde score is 40%. De gemiddelde score is 80%. Ongeveer de helft van de docenten gaf een hogere score op voor het leren bij een toets in vergelijking met leren als onderdeel van het huiswerk.

Tabel 4

Cut-off Scores bij Leren als Voorbereiding op een toets

Docent	Cut-off Score
1	80%

2	75%
3	75%
4	60-70%
5	100%
6	80-85%
7	Voorkeur leerling

Wederom vroeg ik de docenten om uit te leggen waarom zij voor bovengenoemde scores kozen en niet voor een hogere of lagere score. De antwoorden die docenten op deze vraag gaven, zijn samen te vatten in twee hoofdthema's: toets-normering en motivatie van leerlingen:

Normering Toets. Vijf docenten gaven aan hun keuze voor scores te baseren op normeringen die gelden bij de toets. Zo kozen drie docenten het percentage dat bij toetsen tot een voldoende leidt (vaak rond 60% voor havo en 70-75% voor vwo):

> Wat we nu doen is dat ze drie kroontjes moeten halen voor slimstampen, dat is 3 keer 100%. En dan krijgen ze daarna een toets en dat is dan weer 60 of 70 procent. Maar eigenlijk zou 70% dan voldoende moeten zijn want dat is ook zo op de toets
>
> (Docent 3)

Docent 1 hield ook rekening met de normering die geldt voor toetsen, maar koos een iets hogere score dan het minimale percentage dat nodig is voor een voldoende. Docent 1 gaf namelijk aan dat toets stress soms verhindert dat studenten zich alles herinneren wat zij hebben geleerd. Daarom koos docent 1 voor een score die deze stress zou kunnen compenseren:

> Je moet ook nagaan dat er bij een toets wat meer stress achter zit. […] Je moet 70% goed hebben om een 5,5 te halen. Bij de toets en dan ervaar je stress en kun je dingen niet echt nog opzoeken. Mijn hoofd gaat weer gewoon uit naar 80, 90%. Misschien wel iets hoger, zodat je wel wat zelfverzekerder in je schoenen staat.
>
> (Docent 1)

Ook docent 7 baseerde diens keuze op bestaande normeringen. Echter, in plaats van studenten te verplichten om tot een specifiek percentage te leren, zou deze docent met de studenten in gesprek gaan over het cijfer dat zij willen behalen. Aan de hand van deze gesprekken zou de docent dan bepalen tot welk percentage de studenten moesten leren, op basis van de normering.

Motivatie. Twee docente gaven aan de score te kiezen vanwege de motivatie van studenten. Zo gaf docent 4 aan dat het gebruik maken van een hogere score dan 60-70% niet voor iedereen haalbaar is en leerlingen kan ontmoedigen. Docent 6 gaf ook aan dat het mogen maken van een foutje, kan motiveren om het de volgende keer beter te doen. Daarnaast benoemde deze docent dat leren op basis van te hoge cut-off scores leerlingen ten onrechte het gevoel kan geven dat zij de stof al zo goed beheersen, dat verder leren niet meer nodig is:

> Ik denk ook dat wanneer je dus al 90-100% hebt gehaald, dan krijg je van die kids die denken “oh ik kan het allemaal al wel”. En dan gaan ze in de les achteroverleunen, en die denken dan laat maar zitten.
>
> (Docent 6)

Beschermen van de Leerling. Twee docenten gaven aan bovengenoemde scores te kiezen, omdat hogere scores een negatief effect kunnen hebben op het welzijn van leerlingen. Zo

gaf docent 6 aan dat het maken van fouten ook leerzaam kan zijn en dat het in de huidige maatschappij belangrijk is dat leerlingen leren dat af en toe fouten maken mag. Docent 1 gaf daarbij specifiek bang aan te zijn voor het aanwakkeren van burn-out klachten wanneer deze scores van bijvoorbeeld 100% zou hanteren:

> Als leerlingen 100% moeten snappen dan is dat echt heel veel en dat zorgt echt voor burn-out klachten. Dat zie je nou wel bij allemaal leerlingen terug met faalangst en dat soort dingen die studeren zoveel en proberen altijd 100% te halen. En wanneer ze dan een toets maken krijgen ze een black out als ze één ding dan even zijn vergeten. Dus ik zou niet zeggen 100%.
>
> (Docent 1)

Huiswerk vs. Toets. Zoals te zien in tabel 4, koos docent 6 voor een score van 100%. De toelichting van de docent was dat deze voor een score van 80% had gekozen wanneer het gaat om leren voor huiswerk, en bij een toets de score zou willen verhogen.

Samenvatting Toets Vervanging. Over het algemeen achten docenten scores van beneden de 100% acceptabel wanneer leerlingen leren voor een toets. De keuzes voor deze scores zijn gebaseerd op toets normeringen en de motivatie van leerlingen. Drie docenten kozen scores die op reguliere toetsen resulteren in een voldoende gebruikelijke voldoende percentages (60-75%). Daarbij bedandrukten verschillende docenten het belang van realistische doelen om demotivatie en burn-out klachten bij leerlingen te voorkomen. Docenten kiezen over het algemeen dus dus voor een balans tussen het kunnen halen van de toets maar ook het beschermen van het welzijn en motivatie van leerlingen.

Differentiatie

De derde hoofdvraag betrof of docenten open staan voor onderscheid maken tussen leerlingen, bij het bepalen tot welke score zij moeten leren. Bij het stellen van deze vraag gaf ik aan dat niet elke leerling even snel leert wanneer zij de leersoftware gebruiken en dat sommige leerlingen daardoor erg lang kunnen doen over het behalen van hoge scores. De antwoorden van de docenten zijn kort samengevat in tabel 5.

Table 5

Titel

Docent	Antwoord
1	Voornamelijk wanneer leerlingen meer willen leren dan de opgegeven score
2	Enkel wanneer leerlingen meer willen leren dan de opgegeven score
3	Beide kanten op differentiëren
4	Beide kanten op differentiëren
5	Beide kanten op differentiëren
6	Indelen op ander niveau (biedt andere lesstof aan in plaats van andere score)
7	Beide kanten op differentiëren (voorkeur leerling)

De toelichtingen die docenten gaven voor hun antwoorden zijn hieronder aan de hand van thema's weergegeven.

Je Mag een Minimaal Niveau Verwachten. Docent 1 gaf aan open te staan voor onderscheid, maar met name door de lesstof aan te passen voor leerlingen die de stof al goed beheersen. Voor leerlingen voor wie de stof uitdagend is zou deze docent enkel voor een beperkte tijdsduur onderscheid maken:

> Leerlingen die de stof machtig zijn, mogen en aantal opdrachten skippen. [..] Als je meer kijkt vanuit het perspectief van iemand die het wat moeilijker vindt, zou ik wel gewoon echt die lijn aanhouden. Je hebt het wel over een niveau en als ze dat niet aankunnen voor Engels, Frans, Duits of iets dergelijks, dan is bijles waarschijnlijk iets wat je kan toepassen. Maar je moet niet laag gaan omdat ze het niet aankunnen, denk ik. Misschien in het begin wel om zelfvertrouwen te kweken, maar als dat elke keer gebeurt, dan blijven die cijfers laag. Ik ben wel iemand die vindt dat het niveau moet wel op het niveau liggen. Iemand die havo doet moet niet elke keer 4tjes, net 5jes net 5,5 gaan halen. Terwijl die echt zijn best doet, zeg maar, dus Ik vind niet dat dat lager moet.
>
> (Docent 1)

Onderscheid is Confronterend. Twee docenten gaven aan moeite te hebben met naar beneden differentiëren. Dit zou namelijk confronterend kunnen zijn voor de leerlingen voor wie de stof uitdagend is:

> Dan kom je snel in een moeilijk gebied: vanuit de leerling gezien doe je het dus, vergeleken met iemand anders, minder goed en daarom hoef je maar tot een lager percentage te leren. […] Misschien kan je het dan net een beetje anders presenteren: de langzame leerling laat je gewoon zo, maar een snelle leerling krijgt een vinkje bij de naam, een soort van bonus of erkenning dat je goed bezig bent.

(Docent 2)

Ook docent 6 gaf aan dat het maken van onderscheid op basis van verschillende percentages confronterend kan zijn voor de leerlingen die meer moeite hebben met de stof. Daarom zou deze docent liever onderscheid maken door leerlingen op verschillende taalniveaus opdrachten te laten maken, waarbij zij leren tot dezelfde percentages.

Vier docenten gaven aan scores aan te willen passen voor zowel leerlingen die de stof goed beheersen als leerlingen voor wie de stof uitdagend is. De argumenten die zij hiervoor gaven zijn hieronder samengevat:

Leerervaring van de Leerling: Frustratie en Motivatie. Docent 5 gaf aan beide kanten op te willen differentiëren, omdat dit goed is voor de motivatie van studenten: "Ja ik denk dat het wel belangrijk is dat ze ook gemotiveerd blijven, dus ja het moet wel haalbaar zijn zeg maar." Ook docent 3 gaf aan onderscheid te willen maken, omdat het leren met een dergelijk systeem niet voor elke leerling even makkelijk of snel gaat. Voor sommigen zou het gebruik van het systeem zelfs frustrerend kunnen zijn, en voor deze leerlingen is het belangrijk dat zij tot een lagere score kunnen leren. Ik vroeg aan deze docent hoe onderscheid maken motivatie beïnvloedt, en deze antwoorde het volgende:

> Ik denk dat dit goed is voor de motivatie, omdat je dan meer naar de leerling kijkt en zij ook wel door hebben van 'oké ik vind het lastig' maar de docent maakt toch een aanpassing voor mij om het haalbaar te maken om met Slimstampen om te gaan, in plaats van gewoon die lijst af te maken omdat je dan pas de bonus krijgt, terwijl het voor die leerling gewoon een obstakel is.

(Docent 3)

Onderwijs op Maat. Docent 4 gaf aan te willen differentiëren omdat dit zal resulteren in onderwijs dat beter aansluit op de behoeftes van leerlingen, wat zorgt voor passender onderwijs: “Je kan veel passender onderwijs geven: in plaats van dat je altijd op die middenmoot gaat zitten, kan je eigenlijk iedereen veel passender lesgeven en dus beter voorzien in wat diegene nodig heeft. Ja, dat is wel ideaal.”

Voorkeur Leerling. Docent 7 gaf aan open te staan voor differentiatie. De docent zou daarbij de keuze voor een score door de leerling laten bepalen.

Samenvatting Differentiatie. De meeste docenten staan open voor differentiatie. Daarbij staan de meeste docenten ervoor open om de scores naar beneden of naar boven aan te passen. Volgens de docenten helpt differentiatie motivatie te behouden of verhogen en frustratie te verminderen. Differentiatie wordt daarnaast gezien als een middel om gepersonaliseerd onderwijs te bieden dat zowel uitdaging als ondersteuning biedt, afhankelijk van de individuele behoeften van elke leerling. De voornaamste reden dat docenten niet of slechts beperkt onderscheid wilden maken, was het confronterende aspect van differentiatie voor leerlingen voor wie de stof uitdagend is.

Inspraak Leerling

Deze vraag betrof of de docent ervoor open zou staan dat leerlingen meebepalen tot welke score zij leren. De antwoorden van docenten zijn kort samengevat in tabel 6:

Table 6

Titel

Docent	Antwoord

1	Alleen wanneer zij meer willen leren dan het standaard percentage.
2	Beide kanten op, maar binnen een bepaalde range
3	Beide kanten op, maar binnen een bepaalde range
4	Beide kanten op, maar in overleg met de docent
5	Beide kanten op, maar in overleg met de docent
6	Beide kanten op, maar binnen een bepaalde range
7	Beide kanten op, zie vorige antwoorden.

Zes van de zeven docenten zouden leerlingen laten meebepalen in de score tot waaraan zij leren, zowel wanneer leerlingen meer en minder willen leren dan het standaard percentage. Docenten noemden daarbij de voorwaarde dat de score werd gekozen in overleg met de docent of binnen een bepaalde range. Op deze manier zou de keuze voor de score goed gemotiveerd blijven en leren leerlingen in ieder geval tot een minimum percentage. Hieronder zijn enkele toelichtingen op antwoorden van docenten omschreven:

Leerervaring van de Leerling: Motivatie, Leerplezier en Eigenaarschap. Het voornaamste argument dat docenten gaven voor het mee laten bepalen van leerlingen, was de positieve invloed die dit zou hebben op de leerervaring. Docent 4 zei het volgende: “Het is juist mooi als ze daar zelf een beetje inspraak in hebben, dan voelen ze zich waarschijnlijk iets meer eigenaar van het hele leerproces.”

Ook volgens docent 2 zou het mogen meebepalen het gevoel van autonomie versterken, wat een positieve invloed heeft op de motivatie van leerlingen:

> Ik denk dat je wel een deel van de verantwoordelijkheid bij de leerling zelf kan leggen. Ik denk dat ze dan wel meer autonomie voelen over het leerproces. Het is wel een beetje een gevoelskwestie. Je moet ze niet te veel vrijheid geven, maar wel een beetje het gevoel aan ze geven dat ze zelf die keuze maken, ik denk dat dat wel motiveert.
>
> (Docent 2)

Docent 6 gaf aan dat mee mogen bepalen kan leiden tot meer leerplezier:

> Oh dat vind ik wel een hele interessante, ik zou daar wel mee willen experimenteren, kijken wat er dan gebeurt, wat zij zou vinden, wat die zou doen. Ik zou er wel voor openstaan als je een soort van pilot kunt draaien, en laten we dan eens kijken wat daar het effect van is. En of het daardoor leuker wordt, want ik denk dat school soms niet heel leuk is en je moet het op een leuke manier aanpakken, dan raken ze gemotiveerd en misschien dat dit de manier is.
>
> (Docent 6)

Samenvatting Inspraak Leerlingen. De meerderheid van de docenten staat ervoor open dat leerlingen meebepalen tot welke score zij leren. Een van de hoofdargumenten die docenten daarbij noemen is het positieve effect van deze inspraak op de motivatie en het leerplezier van de leerlingen. Volgens de docenten leidt betrokkenheid van leerlingen bij hun eigen leerproces tot een groter gevoel van autonomie en eigenaarschap over het leren. Docenten benadrukken wel dat dit binnen bepaalde grenzen en in overleg moet gebeuren, zodat de keuzes verantwoord blijven. Docenten zijn dus bereid leerlingen vrijheid te geven in de score tot waaraan zij leren, op voorwaarde dat zij de leerling daarin mogen begeleiden.

Vervangen Toets

Tot slot vroeg ik of de docenten het akkoord vinden wanneer het leersysteem (sommige) reguliere kennistoetsen zou vervangen. Ook vroeg ik hen of zij dan akkoord gaan wanneer leerlingen leren tot een 6. De antwoorden van de docenten zijn samengevat in tabel 7.

Tabel 7

Titel

Docent	Toets vervangen	6 voldoende
1	Ja	Ja
2	Ja	Ja
3	Ja	Ja
4	Ja	Ja
5	Ja	Nee
6	Ja	Ja
7	Ja	Ja

Zoals te zien in tabel 7, gaven alle docenten aan dat zij het akkoord vinden wanneer het leersysteem toetsen zou vervangen. De argumenten die hiervoor gegeven werden, zijn hieronder samengevat.

Efficiëntie en Werklast. Docent 2 gaf aan (graag) van een dergelijk systeem gebruik te maken. Het afnemen en beoordelen van toetsen neemt namelijk veel tijd in beslag. Wanneer toetsen worden afgenomen met behulp van een leersysteem en wanneer deze ook automatisch antwoorden nakijkt, dan houden docenten meer tijd over voor andere belangrijke zaken en het voorbereiden van de les:

> Ja heel erg fijn, daar heb ik ook wel over nagedacht. Toetsen nakijken en zo kost veel werk. En een groot deel daarvan is gewoon woordjes nakijken en dat voelt inefficiënt. En als je je als docent daar niet mee bezig hoeft te houden, dan kan je je bezighouden met belangrijkere en nuttigere zaken.
>
> (Docent 2)

Meegaan met de tijd. Volgens docent 6 is het vervangen van traditionele toetsen met een leersysteem een goed idee. Het afnemen van enkel traditionele toetsen is namelijk ouderwets te noemen. Het vervangen van dit soort toetsen met een adaptief leersysteem betekent dat het onderwijs meegaat met haar tijd. Zolang leerlingen deze manier van werken ook prettig vinden, ziet docent 6 geen reden om dit niet te doen.

Toets cijfer. Bij de vraag of leerlingen mogen leren tot een 6 wanneer een leersysteem een toets vervangt, gaven vijf docenten aan dat zij dit akkoord vinden. Docent 5 gaf aan misschien toch voor een wat hoger cijfer te gaan:

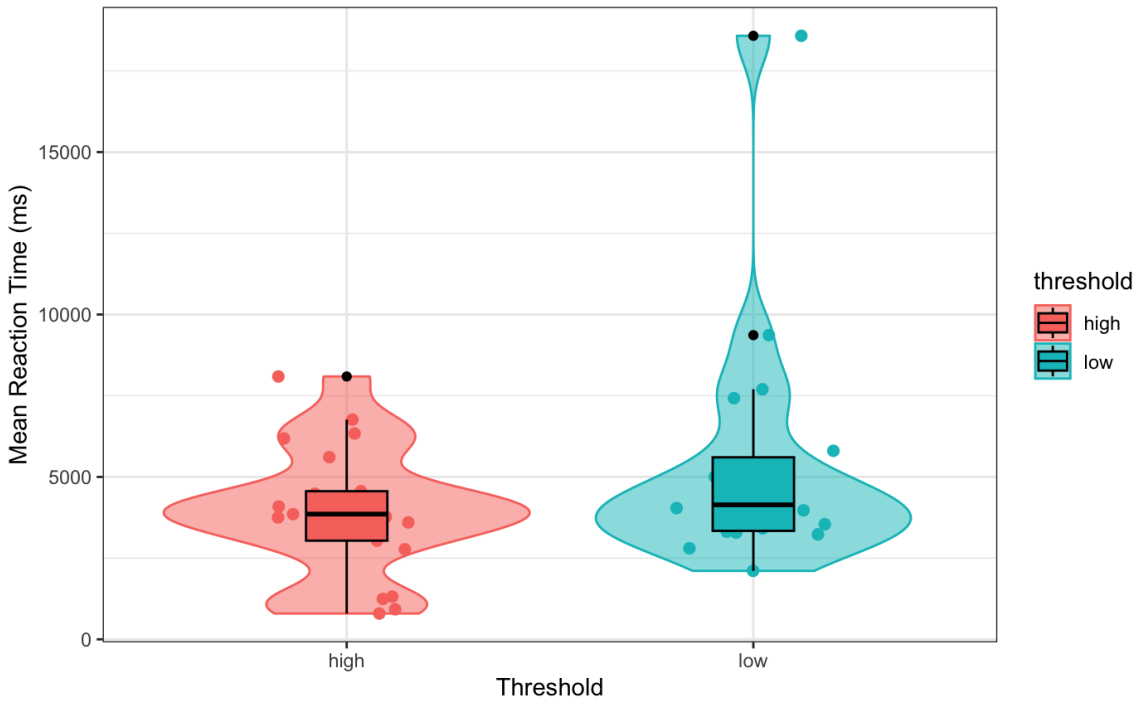
> Nee dat is gek, ik denk dan meteen van nou nee, dan moet het ook hoger zijn. Maar ja waarom? Want ja bij een toets is een 6 ook gewoon een 6? Je zou hopen dat het dan op zo'n leuke manier gepresenteerd wordt, dat ze er zo door de smaak door te pakken krijgen dat ze dan ook zelf wel wat hoger zouden willen scoren dan een 6, maar in principe is een 6 een 6. Ik denk dat ik wat hoger zou willen, maar dan niet heel veel. Een 7 ongeveer.
>
> (Docent 5)

Samenvatting Toetsen Vervangen. Alle docenten staan open voor het idee om traditionele kennistoetsen te vervangen door leren met een leersystem. Docenten gaven als reden

op dat dit werklast vermindert, efficiëntie verhoogt en dat deze verandering een teken is dat het onderwijs meegaat met haar tijd. Alle docenten behalve één vonden het acceptabel wanneer leerlingen leren tot een 6. De reden dat één docent zou wille dat leerlingen tot iets hoger leren dan een 6, was dat het leren met een dergelijk systeem leuker zou zijn, wat leerlingen zou motiveren om meer te leren.

Appendix E

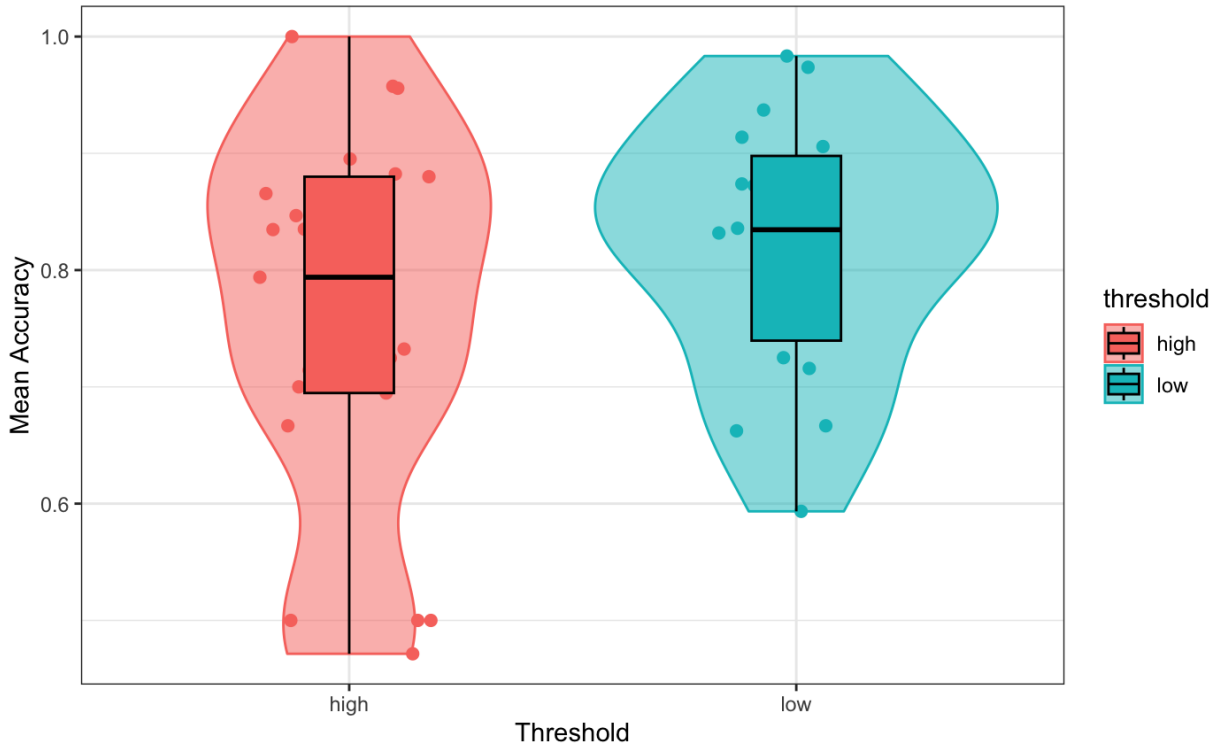
Mean Reaction Time Across all Practiced Items During Learning Phase, by Threshold.



Note. Every dot represents the mean reaction time of one participant. A black dot represents an outlier. The whiskers extend to the furthest data points within 1.5 times the interquartile range from the boxplot.

Appendix F

Mean Accuracy Across all Practiced Items During Learning Phase, by Threshold.



Note. Every dot represents the mean accuracy score of one participant. The whiskers extend to the furthest data points within 1.5 times the interquartile range from the boxplot.

Appendix G

Model for Comparison Reaction Times During Test Phase, for Both Thresholds and Across all Practiced Items

Fixed effects	Effect	Estimate	*SE*	*t*-value	*p*-value
	Intercept	2551.44	303.43	8.41	< .01
	Threshold low	273.40	236.040	1.16	.25
Random Effects	Group	Variance	*SD*		
	Answer (intercept)	548163	740.40		
	Participant (intercept)	772133	878.70		
	Residual	6104332	2470.70		
	Number of obs.	464			
Model Metrics	REML Criterion at Convergence				
	8591.70				

Appendix H

Model for Comparison Reaction Times During Test Phase, for Both Thresholds and Across all Mastered Items

Fixed effects	Effect	Estimate	*SE*	*t*-value	*p*-value
	Intercept	2621.28	509.74	5.14	< .01
	Threshold low	1307.44	519.82	2.52	.012
Random Effects	Group	Variance	*SD*		
	Answer (intercept)	541939.00	736.20		
	Participant (intercept)	1023742.00	1011.80		
	Residual	31108034	5577.5		
	Number of obs.	547			
Model Metrics	REML Criterion at Convergence				
	10980.2				

Appendix I

Model for Comparison Accuracy During Test Phase, for Both Thresholds and Across all Practiced Items

Fixed effects	Effect	Estimate (log-odds)	*SE*	*z*-value	*p*-value
	Intercept	0.55	0.34	1.60	.11
	Threshold low	-0.18	0.18	-1.022	.31
Random Effects	Group	Variance	*SD*		
	Answer (intercept)	1.40	1.18		
	Participant (intercept)	1.19	1.093		
Model Fit	AIC	BIC	Log Likelihood	Deviance	Residual df
	902.80	921.40	-447.40	894.80	763

Appendix J

Model for Comparison Accuracy During Test Phase, for Both Thresholds and Across all Mastered Items

Fixed effects	Effect	Estimate (log-odds)	*SE*	*z*-value	*p*-value
	Intercept	0.96	0.35	2.73	< .01
	Threshold low	-0.50	0.25	-2.035	0.042
Random Effects	Group	Variance	*SD*		
	Answer (intercept)	1.37	1.17		
	Participant (intercept)	0.90	0.95		
Model Fit	AIC	BIC	Log Likelihood	Deviance	Residual df
	640.10	657.30	- 316.00	632.10	543