The effect of computer assisted formative assessment on behavioral student engagement and learning outcomes in a high school class on digital tooling

SL3502 The Science Education Thesis

by

Justin Rademaker

Supervisor:

Rachel Baan

A Dissertation

Submitted to Applied Sciences faculty

Delft University of Technology,

In Partial Fulfillment of the Requirements

For the Master of Science Education and Communication

May 6, 2025

Abstract

Formative assessment has been shown to improve student engagement and learning outcomes across several subject domains in K-12 education. However, its effectiveness within the subject domain of digital tooling remains understudied. This research investigated the effect of computer assisted formative assessment on learning outcomes and behavioral student engagement, with the latter as a potential mediating variable, within the subject domain of digital tooling.

This research conducted a quasi-experiment with 122 second-grade students from a Havo/Vwo high school during a course on Google Spreadsheets. The experimental group had access to a button that let the computer instantly check their homework and provide feedback, while the control group did not. At the end of the course, students in the experimental group completed a questionnaire to share their experiences.

This research found no significant effects of the intervention on behavioral student engagement and learning outcomes, nor was a mediated relationship established. However, students did report several cognitive and metacognitive benefits. Namely: enhanced motivation, enhanced self-regulated learning, enhanced autonomy, feedback that helped them move forward and an improved understanding of the material. On the other hand, they reported some downsides. Namely: increased time consumption, a lack of added value and a lack of feedback quality.

Since this research did not find effects of computer assisted formative assessment similar to those observed in other subject areas, it raises the question of whether the subject domain of digital tooling interacts differently with this type of intervention. Further research with a greater sample size and statistical power is needed to explore this possibility. Future research could investigate which characteristics make this subject domain distinct, in which contexts within digital tooling this type of intervention is most effective and which benefits it should aim to provide.

Table of contents

Glossary

Term	Definition
Cohen's d (d = …)	A measure for effect size. It measures how different the averages between two groups are and is expressed in standard deviations.
Direct effect	A term used within mediation analyses. It is the causal effect between the independent and dependent variable while controlling for the mediator variable(s).
Ecological validity	A subtype of external validity that is concerned with the generalizability of the findings to real-world settings.
Effect size	Indicates the practical significance of a research outcome.
Embedded mixed methods approach	Combines quantitative and qualitative research in order to answer the research question. 'Embedded' means that one type of data is secondary to the other.
External validity	The extent to which findings can be generalized to a broader context (different situations, people, etc.).
Havo	The Dutch name for 'Senior General Secondary Education' which is a five year program that prepares students for universities of applied sciences.
Indirect effect	A term used within mediation analyses. It is the causal effect between the independent and dependent variable, through the mediator variable, while controlling for the direct effect.
Internal validity	The extent to which a cause-and-effect relationship is established purely by the independent variable(s).
K-12 education	Covers all grades starting from kindergarten up until (and including) secondary education.
Mediation analysis	An analysis that establishes the extent to which some causal variable influences an outcome through one or more mediator variables.
Multiple linear regression	A linear regression model that estimates the relationship between two or more independent variables and one dependent variable.
Pearson Correlation Coefficient (r = …)	A parametric measure for correlation. It measures the strength and direction of the relationship between two variables.
Population validity	A subtype of external validity that is concerned with the generalizability of the findings to the targeted population.
Spearman's Rank-Order Correlation (r_s)	A non-parametric statistical test that calculates a correlation coefficient.
Total effect	A term used within mediation analyses. It is the sum of the direct effect and the indirect effect.
Vmbo	The Dutch name for 'Preparatory Vocational Secondary Education'. It's a four year program that prepares students for 'secondary vocational education' (Dutch: MBO), which prepares students for a specific (type of) job.
Vwo	The Dutch name for 'University Preparatory Education' which is a six year program that prepares students for research universities.

1. Introduction

Intuition tells us that providing students with formative assessment and feedback improves their learning outcomes. Yet, in their 1996 review on feedback interventions, Kluger and Denisi found that 231 of the 607 effect sizes they calculated were actually negative, meaning that students would have performed better without the feedback. These results indicate that although feedback has great potential, it should be studied in different forms and contexts to determine when it really thrives. Bennett (2011) made a similar argument for formative assessment.

Many later studies have done exactly that. Meta-analyses have calculated specific effect sizes for feedback and formative assessment in different contexts, such as grade level or subject domain, as well as in different forms, such as with different information densities or by the use of a computer (Hattie & Zierer, 2019; Klute et al., 2017; Lee et al., 2020). Formative assessment has also been shown to be effective for different outcome measures, such as student engagement (Barana et al., 2019), which in turn has been shown to improve learning outcomes (Lei et al., 2018).

This research contributes to the body of knowledge by investigating a specific form of formative assessment in the context of an understudied subject domain. Namely, computer assisted formative assessment within the subject domain of digital tooling, which is a part of ‘Digital Literacy’, which in turn is a formal subject in Dutch primary and secondary education. This research uses a quasi-experimental design combined with a survey to investigate the effects of computer assisted formative assessment on students’ behavioral engagement, learning outcomes, and to what extent their learning outcomes are mediated by their behavioral engagement. While the contribution of this research is primarily theoretical, it also offers practical insights into the potential value of developing computer assisted formative assessment tools for the teaching of digital tools.

The research was conducted during a seven week course on ‘Google Spreadsheets’ at a Dutch havo/vwo high school. Six classes (n = 122) from the second grade participated in the study.

This paper starts off with a literature review which discusses the scientific history and theoretical debate around formative assessment, feedback and student engagement. This is followed by a theoretical framework, which states the theoretical assumptions that this research is based upon. With this theoretical foundation, the research questions are then formulated, including the hypotheses and the conceptual framework for this research. Next, the methodology section explains the sampling method, data collection and data analyses that are employed to answer the research questions. It also describes how reliability and validity have been taken into account. Then, the results section presents the research results, followed by the discussion which interprets these results and goes into potential directions for future researchers to expand upon this work. Finally, the paper concludes by answering the research question.

2. Literature review

2.1 Introduction

This section reviews the literature of the three relevant constructs within this research: formative assessment, feedback and student engagement. Each construct is introduced with its scientific history and effectiveness, followed by a description of the construct and a brief summary of the current theoretical debate. Finally, the literature review identifies a theoretical gap in the literature and shows how this research aims to help close it.

This literature review provides context for the upcoming theoretical framework, which describes the position of this research within the theory.

2.2 Review method

To get an overview of the theoretical landscape, Google Scholar was used; the construct names as mentioned above were used as keywords. Whenever the papers used relevant synonyms or related terms, these were also added to the set of keywords (e.g. ‘assessment for learning’ and ‘behavioral student engagement’).

As a first step, seminal works were identified, either by looking for papers with high citation counts or by reviewing those that were recognized as ‘seminal’ by other authors. To get an overview of the scientific history, no timeframes were filtered out. To get an overview of the current theoretical debate, findings before 2020 were cut off.

After reviewing the seminal papers, more papers were found by either selecting relevant sources from the reference list (backwards in time) or by selecting sources that cited the paper (forward in time). Throughout the review, a relatively high citation count remained a key criterion for selecting papers. During the review, extra attention was placed on identifying theoretical themes, debates and gaps.

2.3 Formative assessment

2.3.1 History and effectiveness of formative assessment

In 1971, Bloom et al. first used the term ‘formative assessment’ to describe a form of assessment that, in contrast to summative assessment, has the primary purpose to help improve learners in what they are doing (Black & Wiliam, 2003). The field truly gained traction in 1998, when Black and Wiliam reviewed 250 publications and presented their findings in two seminal papers (Black & Wiliam, 1998a; Black & Wiliam, 1998b). Here, they presented uncommonly large effect sizes of the formative assessment experiments on students’ learning outcomes, which were often between 0.4 to 0.7 standard deviations. These results were later criticized for overreliance on sources that are untraceable, flawed, dated or unpublished (Bennett, 2011; Dunn & Mulvenon, 2009; Kingston & Nash, 2011).

Since then, several meta-analyses have confirmed a (more modest) positive impact of formative assessment on learning outcomes. In 2011, Kingston and Nash found an average effect size of 0.2 standard deviations, with the second most effective strategy being computer-based formative assessment ($d = 0.28$) (the most effective was teacher professional development on formative assessment, $d = 0.30$). In 2017, Klute et al. found an

average effect size of 0.26 standard deviations. They found that it was more effective to let teachers or computers do the assessment (d = 0.29), than it was to let students assess their own or each other's work (0.2). In 2020, Lee et al. found an overall effect size of 0.29 standard deviations. In contrast to the findings of Klute et al., they found student-initiated self-assessment (d = 0.61) to be far more effective than interventions that promoted teacher's practices for formative assessment (d = 0.18). Furthermore they found an effect size of 0.21 standard deviations for computer based formative assignments.

Although research of formative assessment is largely devoted to the improvement of learning outcomes, other outcome measures have been studied as well. In 2020, Leenknecht et al. found that formative assessment improves students' feeling of autonomy, competence and relatedness and therewith their motivation. Another such study is that of Barana et al. (2019) who found that formative assessment improves student engagement.

Finally, as technology advances, so do the applications for formative assessment. As an example, Tobler (2024) used a large language model to automatically grade open questions. Although it showed promising results, the author stated that current AI technology is still limited and likely to make errors, which raises ethical concerns.

2.3.2 The theoretical debate

In another seminal paper, Black and Wiliam (2009) acknowledged their initial works to lack a theoretical basis and proposed a theoretical framework that unifies the diverse set of practices that have been described as formative (see Figure 1 in section 'Theoretical framework'). Although the framework has been widely adopted in research, an ongoing theoretical debate in the field has persisted. Researchers have been arguing about:

- **Terminology:** e.g. 'assessment for learning' vs 'formative assessment';
- **Conceptualization:** e.g. how it should relate to summative assessment;
- **Definitions:** e.g. to what extent should we talk about a process instead of a test?

See Dunn and Mulvenon (2009), Bennett (2011), Black and Wiliam (2018), and Wiliam (2018) for more details on the theoretical debate. Nevertheless, Brookhart (2018) emphasized that despite the various perspectives in the field, they all share the fundamental concept that information coming from assessment should serve as instructional feedback to enhance student learning.

In conclusion, despite the ongoing theoretical debate, the field of formative assessment has led to effective strategies to improve students' learning outcomes.

2.4 Feedback

2.4.1 History and effectiveness of feedback

The concept of feedback became of interest in the fields of psychology and education around the mid-20th-century. Early research explored the construct primarily through a behavioristic paradigm and investigated the effects of positive and negative reinforcement through feedback (Wiliam, 2018). Around the 80s, feedback research became increasingly influenced by cognitive and constructivist theories, shifting the focus toward how feedback

could help students process information and construct their knowledge (Lipnevich & Panadero, 2021).

In 1996, Kluger and DeNisi published a seminal paper reviewing the research on feedback up to that point in time. Although they found the overall effect size of feedback on student achievement to be positive (d = 0.4), they also found that over one third of the effect sizes they calculated were negative. This highlighted the importance of further research, as it showed that improperly administered feedback could actually lower students' achievement.

Part of the research has focused on the effectiveness of different types of feedback across different contexts. Typical moderators in meta-analyses include: research design, feedback type (e.g. feedback on correctness vs feedback on self-regulation), feedback direction (e.g teacher to student vs student to student), outcome measure (e.g. cognitive vs motivational) and learners' age (Li, 2010; Hattie & Zierer, 2019; Wisniewski et al., 2020). This helps us clarify the conditions under which feedback is most effective.

For example, Wisniewski et al. (2020) conducted a meta-analysis of 435 studies and found feedback to be effective for both cognitive outcomes (d = 0.51), which includes student achievement, retention and cognitive test performance, as for motivational outcomes (d = 0.33), which includes intrinsic motivation, locus of control, self-efficacy and persistence. They also found high-information feedback to be more effective (d = 0.99) than simply stating whether a students' performance is right or wrong (d = 0.46).

Part of the feedback research is specifically focused on computer-based environments to optimize students' learning. Feedback in this context has specific characteristics. Van Der Kleij et al. (2015) mentioned that the computer-based environment allows feedback to be provided immediately and can be tailored to students' individual needs. However, it is also more easily ignored by the students. Kuklick et al. (2023) found that feedback in computer-based environments that merely indicates whether something is right or wrong (which is less common in human-delivered feedback) can negatively impact student motivation, but only after incorrect responses.

Regarding the effectiveness of feedback in computer-based environments, Van der Kleij et al. (2015) performed a meta-analysis of 40 studies that showed that instructional feedback is more effective than feedback limited to correctness or simply providing the correct solution (Van Der Kleij et al., 2015). Furthermore, they found that higher information feedback leads to higher learning outcomes, aligning with the findings of Wisniewski et al. (2020). This applies to both lower level-learning (remembering, understanding and applying) and higher-level learning (analyzing, evaluating and creating), though higher-level learning benefits the most. See Anderson and Krathwohl (2001) for more details on lower- vs higher-level learning.

2.4.2 The theoretical debate

In the meantime, research has spawned numerous theoretical models and definitions for feedback. One seminal article was that of Nicol and Macfarlane-Dick (2006). They connected formative assessment with self-regulated learning, arguing that students generate internal feedback on the way to their goals. Another seminal article was by Hattie and Timperley (2007) in which they presented a feedback model distinguishing between different

types of feedback (see section ‘Theoretical framework’ for more details). Some of which proved to be more effective than others and through this model they provided a more structured view at the construct of feedback. Finally, as for a definition, Lipnevich and Panadero (2021) showed in their review that researchers’ proposed definitions seem to get more aligned in time, and include most of the following elements:

- **Information:** feedback consists of information that is exchanged;
- **Gap:** feedback intends to close the gap between the learner's current performance and desired performance;
- **Process:** the feedback process involves cognitive, affective and regulatory steps. Simpler put: the learner thinks, feels, and plans when receiving feedback;
- **Agents:** feedback can be provided by different educational agents like teachers, peers or computers;
- **Students’ active processing:** the learner should actively receive the feedback;
- **Internal feedback:** feedback can also be produced by the learner itself.

In conclusion, feedback research has received lots of attention and has made substantial progress. Although there is no complete theoretical consensus yet, the research has already provided guidelines on how to give effective feedback.

2.5 Student engagement

2.5.1 History and effectiveness of student engagement

One of the earliest efforts to formalize the construct of student engagement was the ‘participation model’ by Natriello in 1984 (Wong & Liem, 2021). The model focused on behavioral variables; student engagement was akin to participation in school activities. As research progressed, new perspectives on student engagement emerged. In 2004, Fredricks et al. proposed a seminal framework describing three types of student engagement:

- **Behavioral:** positive conduct, participation in school-related activities and involvement in learning and academic tasks;
- **Emotional:** students’ affective reactions in the classroom like boredom, happiness, anxiety, etc;
- **Cognitive:** psychological investment in learning, self-regulation and learning strategies.

This framework has been widely adopted by researchers and in 2018, Lei et al. conducted a meta-analysis encompassing 196,473 participants where they showed that the construct is moderately correlated with academic achievement (r = .269). The strongest correlation was found between behavioral engagement and academic achievement (r = .350), followed by cognitive engagement (r = .245) and emotional engagement (r = .216). A moderator analysis

showed that the correlation regarding behavioral engagement was lower for self-reported measures (r = .303) than for other types of measures (r = .428).

Researchers have since proposed expansions to the framework. For example, Reeve (2013) considered a fourth type, called ‘agentic engagement’, which refers to the amount of agency students apply in their own learning process and showed it to be a significant predictor of academic achievement even when controlling for other types of engagement. Another example is the ‘social engagement’ of Fredricks et al. (2016), which is about interacting and collaborating with others and was shown to be a unique predictor to academic achievement.

2.5.2 The theoretical debate

Although the field seems to progress gradually, Wong and Liem (2021) noted in their review that the field also suffers from conceptual haziness. specifically:

- **overgeneralization:** student engagement is sometimes used as a catch-all term for various concepts that increase student school success;

- **jingle-jangle fallacies:** Sometimes the term student engagement is used for concepts that mean different things and sometimes different terms are used to describe (dimensions of) student engagement;

- **object ambiguity:** research often lacks specificity about what exactly students are engaged in;

- **under-theorization:** there is no theory that represents the core of what student engagement is about. Instead, often concepts from motivational research are used.

In conclusion, the stimulation of student engagement seems like a worthwhile pursuit in order to increase academic achievement. Yet, researchers are still in debate on how to theoretically formalize the construct and thus, new research should clearly specify what they mean by student engagement.

2.6 Theoretical gaps

So far we have seen how formative assessment, feedback and student engagement improve academic success. Furthermore there has been evidence of effective use of computer assisted formative assessment. These are interesting findings since computer assisted formative assessment could both alleviate teachers’ workload and contribute to personalized learning for the students.

In 2011, Bennett argued that, in order for formative assessment to be maximally effective, it should be considered in the context of specific subject domains. Much research has indeed focused on specific domains, which often show a considerable variation in effect sizes. Examples of researched domains are: mathematics, science (e.g. physics, chemistry and biology), languages, arts, reading, writing and social sciences (Kingston & Nash, 2011; Klute et al., 2017; Van Der Kleij et al., 2015; Lee et al., 2020).

In 2021, Wong and Liem made a similar argument about student engagement, proposing that the construct should be explored through specific subject domains.

So far, little research has been devoted to the subject domain of Digital Literacy. In Dutch K-12 education, Digital Literacy aims to equip students with digital skills across four domains[1]: practical IT skills (usage of digital tools), media awareness (responsible media use), digital information skill (collecting, evaluating, processing and sharing digital information) and computational thinking (strategies to formulate problems such that computers can solve them) (SLO, 2024).

This research aims to help fill this gap and help verify the existing theories through the subject domain of digital tools (i.e. practical IT skills). Although this domain is relevant throughout all K-12 education, this research solely aims for generalizable findings for the second grade of secondary education. Ideally, a future meta-analysis will incorporate this research as part of a moderator analysis.

[1] At the time of writing, new domains have been proposed, which are conceptual for now. For the purposes of this paper, they are similar enough to not elaborate on.

3. Theoretical framework

To ground the research in theory, this section discusses the theoretical frameworks for ‘formative assessment’, ‘feedback’ and ‘student engagement’ that are used in this research.

3.1 Formative assessment

The literature review section mentioned the widely adopted theoretical framework of Black and Wiliam (2009) on formative assessment in Figure 1. The model shows three phases of formative assessment (horizontal) and emphasizes the responsibility of not only the teachers, but also of the students and their peers (vertical).

<table>
<tr><th></th><th>Where the learner is going</th><th>Where the learner is right now</th><th>How to get there</th></tr>
<tr><td>Teacher</td><td rowspan="2">1 Clarifying learning intentions and criteria for success</td><td>2 Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding</td><td>3 Providing feedback that moves learners forward</td></tr>
<tr><td>Peer</td><td colspan="2" rowspan="2">4 Activating students as instructional resources for one another</td></tr>
<tr><td>Learner</td><td>Understanding learning intentions and criteria for success</td></tr>
<tr><td></td><td></td><td colspan="2">5 Activating students as the owners of their own learning</td></tr>
</table>

Figure 1: Aspects of formative assessment (Black & Wiliam, 2009)

The elaborate nature of the framework sometimes leads authors to focus on a specific subset of it. For example, Irons and Elkington (2021) devoted their book on element 2, 3 and the learners perspective of element 1. The research in this paper does the same; while all the elements will be somewhat relevant, the emphasis will be on element 2 (via assignments) and 3 (via computer assisted feedback).

3.2 Feedback

The research in this paper uses computer assisted feedback in an attempt to improve students’ behavioral engagement and learning outcomes. This section discusses the literature that is used to construct the feedback in an effective way.

In 2021, Lipnevich and Panadero reviewed the 14 most prominent models and theories on feedback, showing the vibrant nature of the field. They noted that not all models are supported by empirical evidence. Moreover, the models often have different aims and focus and are best suited within specific contexts. However, one of the models that is relatively general, is rooted in empirical evidence, and has been highly influential, was proposed by Hattie and Timperley (2007).

In their paper, Hattie and Timperley (2007) argued that feedback should fit the learning context. For example, if the feedback does not match students’ prior knowledge, it will be unhelpful and the student might even feel threatened. On the other hand, when a teacher

uses proper cues to point the student in the right direction, this can be very effective. Cues may be preferred over revealing an answer since students learn from retrieving their knowledge, which is called ‘retrieval practice’ as was shown by Roediger and Butler (2011).

In their model Hattie and Timperley (2007) differentiate between four types of feedback (p. 87):

- **Task level:** “how well tasks are understood/performed”
- **Process level:** “the main process needed to understand/perform tasks”
- **Self-regulation level:** “self-monitoring, directing and regulating of actions”
- **Self level:** “personal evaluations and affect (usually positive) about the learner”

In 2020, Wisniewski et al. replicated and expanded the research of Hattie and Timperley. They investigated the impact of different types of feedback and distinguished:

- **Reinforcement/punishment:** focused on applying (un)desirable consequences, using minimal information on task, process or self-regulation level.

- **Corrective feedback:** focused on the task and process level. E.g. whether an answer is correct, what a correct answer would be, how the student performed a skill and how the student could improve in that regard.

- **High-information feedback:** the same as corrective feedback, but additionally contains information on self-regulation (e.g. monitoring attention, emotions or motivation).

They found that high-information feedback (d = 0.99) was most effective, followed by corrective feedback (d = 0.46) and then reinforcement/punishment (d = 0.24).

The methodology section will show in detail how this research applied these theoretical findings by using high-information feedback that fits the learning context and makes use of cues.

3.3 Student engagement

The research of this paper aims to make inferences about student engagement by analysing how much time students devoted to their exercises (see methodology). The construct of behavioral engagement as described by Fredricks et al. (2004) seems useful at first sight (see literature review). However, since it also encompasses engagement in school activities (e.g. the school dance), the scope is too broad.

A better suited model is that of Wong and Liem (2021), who proposed a theoretically underpinned framework that allows researchers to refine their scope in detail (see Figure 2). The framework distinguishes ‘learning engagement’ (related to learning tasks) and ‘school engagement’ (related to school activities). The model asserts that learning engagement should be investigated within the context of specific subject domains and timescales.

Within this research, behavioral engagement is defined as it is modeled in Figure 2. It is investigated in the subject domain of digital tooling and a timescale of seven weeks.

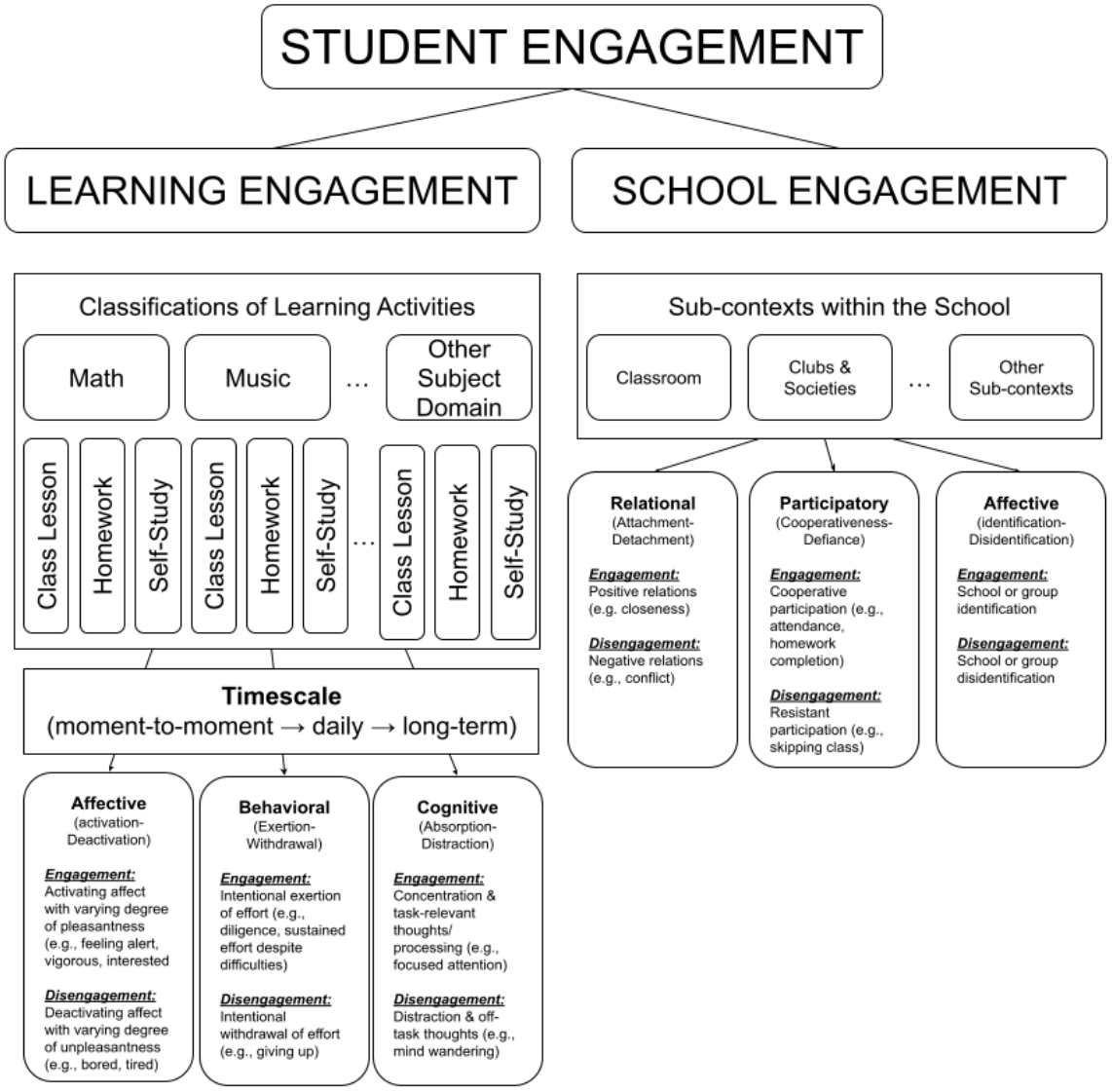

Figure 2: Dual Component Framework of Student Engagement (Wong & Liem, 2021).

3.4 Research questions and objectives

Following the theoretical overview, this section formulates the research questions of this research, followed by a conceptual framework to clarify which relationships are investigated. The section ends by formulating the hypotheses that are tested by this research.

3.4.1 The research questions

This research aims to answer the following research question:

> How does computer assisted formative assessment affect students' behavioral engagement and learning outcomes in the second grade of Dutch secondary education within the subject domain of digital tools?

The research question is broken into the following subquestions:

- RQ 1: To what extent does formative assessment affect students' learning outcomes?
- RQ 2: To what extent does formative assessment affect behavioral student engagement?
- RQ 3: To what extent does behavioral student engagement mediate the relationship between the formative assessment and learning outcomes?
- RQ 4: What are students' experiences regarding the relationships between computer assisted formative assessment, their behavioral engagement and their learning outcomes?

This research describes the relevant concepts as follows:

- **Formative assessment:** this research adheres to the elements shown in Figure 1.
- **Behavioral student engagement:** this research adheres to definition given in Figure 2.
- **Digital tool:** any digital program that is generally regarded as useful for academic or practical goals. Examples are: word-processors, game-engines and simulators.
- **Learning outcomes:** measurable skills or knowledge, as described by the learning goals, that are obtained by the student.

3.4.2 Conceptual framework

This section demonstrates how the research questions are related through a conceptual framework. The labels in the figures (e.g. 'a' and 'prt_c') will be used throughout this paper to denote the relations that are defined by this section. 'prt' stands for 'pre-test score' and the letter after the underscore indicates the relationship that the covariate is controlling for. All effects in this research will be calculated via linear regression analyses.

Research question 1

First, Figure 3 shows that the effect of computer assisted formative assessment on the learning outcomes will be calculated. This relation is denoted by *c*. Note that relation *c* does not control for behavioral student engagement. However, it does control for the pre-test score, which will be explained in the methodology (section 4.2.4).

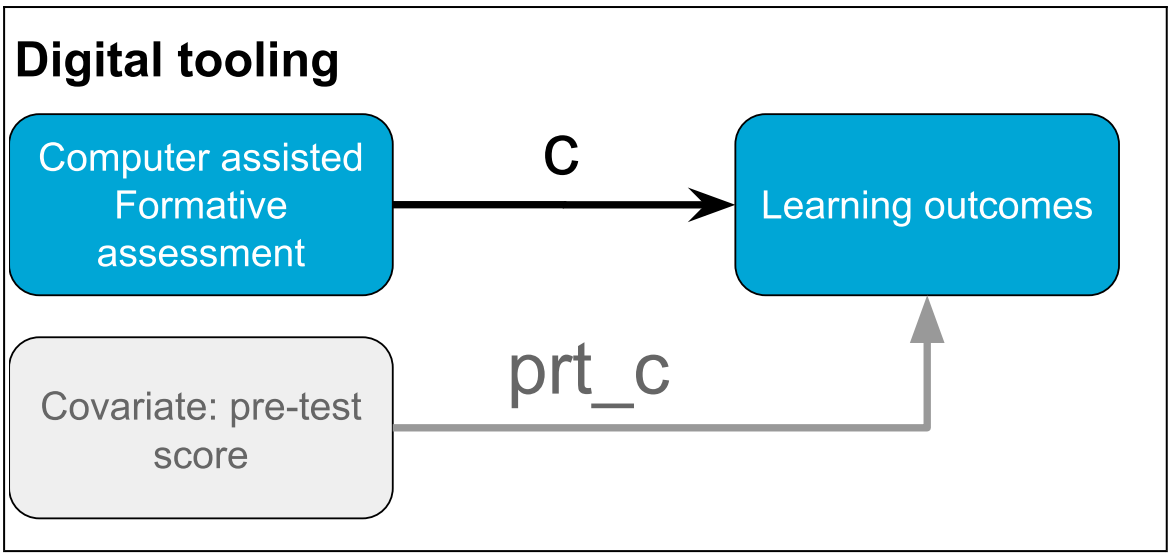

Figure 3: Part 1 of the conceptual framework (RQ 1).

Research question 2 and 3
Then, a mediation analysis will be conducted. Inspired by the theoretical findings, this research investigates the relation as shown in Figure 4. Note that multiple arrows mean that their effects are calculated while controlling the other variables. Furthermore, $c' = c - a \cdot b$. See the methodology for more details (section: The mediation analysis).

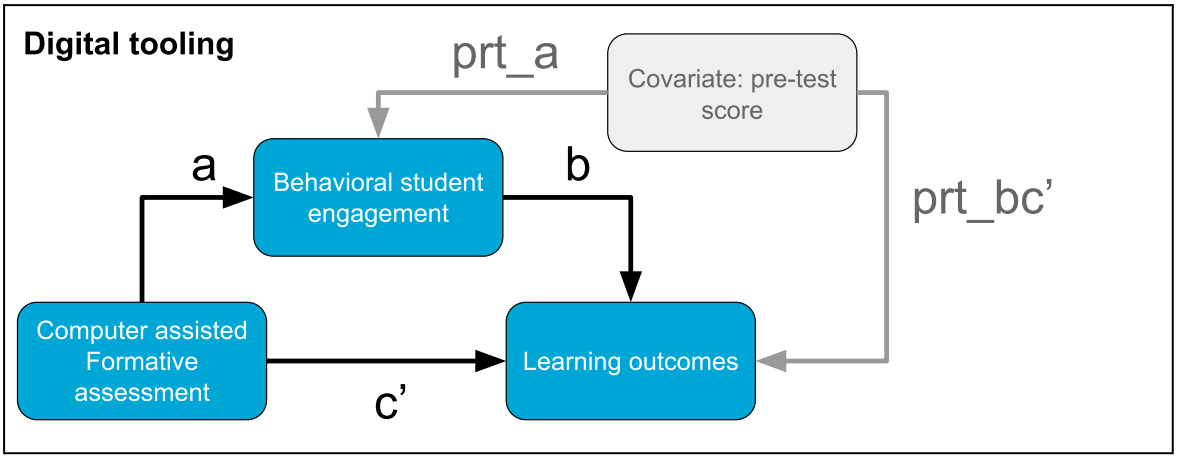

Figure 4: Part 2 of the conceptual framework (RQ's 2 and 3).

Research question 4
Finally, in this qualitative part of the research, students express their opinions on relations *a*, *b*, and *c*. This will support the quantitative findings.

3.4.3 Hypotheses
Regarding RQ1, the following hypothesis will be tested:

- H_{10}: Second grade students in Dutch secondary education that are following a Digital Literacy course on digital tooling **will not differ in their final test scores** when computer assisted formative assessment is offered compared to when it is not offered.

- H_{1a}: Second grade students in Dutch secondary education that are following a Digital Literacy course on digital tooling **will differ in their final test scores** when computer assisted formative assessment is offered compared to when it is not offered.

Regarding RQ2, the following hypothesis will be tested:

- H_{20}: Second grade students in Dutch secondary education that are following a Digital Literacy course on digital tooling **will not differ in their behavioral engagement** when computer assisted formative assessment is offered compared to when it is not offered.
- H_{2a}: Second grade students in Dutch secondary education that are following a Digital Literacy course on digital tooling **will differ in their behavioral engagement** when computer assisted formative assessment is offered compared to when it is not offered.

Regarding RQ3, the following hypothesis will be tested:

- H_{30}: In the second grade of Dutch secondary education where students follow a Digital Literacy course on digital tooling, **behavioral student engagement does not mediate** the relationship between computer assisted formative assessment and learning outcomes.
- H_{3a}: In the second grade of Dutch secondary education where students follow a Digital Literacy course on digital tooling, **behavioral student engagement does mediate** the relationship between computer assisted formative assessment and learning outcomes.

4. Methodology

This research employed an embedded mixed methods approach. RQ's 1 to 3 extend the existing theory about formative assessment through deductive reasoning by using a quantitative approach. It aims to verify a causal path between formative assessment, behavioral student engagement and learning outcomes in the understudied subject domain of digital tooling. On the other hand, RQ 4 aims to provide deeper insights into the results of RQ's 1 to 3 using a qualitative approach.

4.1 The sampling method

When choosing the sampling method, a trade-off was made between internal validity and population validity. The research has been conducted by a teacher (who was also the researcher) at a single high school, who was responsible for six out of eight second-grade classes at the school. These six classes formed the convenience sample of the population (after giving informed consent and excluding those who repeated the grade). The position of the researcher provided the opportunity to control many extraneous variables and therewith achieve a relatively strong internal validity. Furthermore, the realistic real-life setting was beneficial for ecological validity. However, since the population of the research entails all second grade students in Dutch secondary education, the population validity is limited.

To better understand how the results may be generalized, here follows a description of the school. The school is relatively large, consisting of circa 2.000 students and 200 employees. The school offers 'Senior General Secondary Education' (Dutch: Havo) and 'University Preparatory Education' (Dutch: Vwo), but no 'Preparatory Vocational Secondary Education' (Dutch: Vmbo). The school embraces the paradigm of 'Ignatian Pedagogy', which stimulates versatile personal growth. This is expressed in the large variety of courses that the students can take, like Chinese and Philosophy. Courses in Digital Literacy and Computer Science are well established at the school, consisting of a team of 5 teachers. Students take 20 hours of Digital Literacy in the first year and 40 hours in the second. If they choose Computer Science as an elective, they take it for 80 hours per year from the fourth year on.

4.2 Data collection

4.2.1 The experiment

RQ 1 to RQ 3 required a controlled environment from which causal relations could be inferred. Therefore the research employed a between-subjects nonequivalent groups pretest-posttest quasi-experimental design, which will be explained in this section.

Variables and operationalization

The independent variable in this experiment is 'computer assisted formative assessment'. It is treated as a dichotomous variable; either students receive it or they don't. The dependent variables are students' behavioral engagement and learning outcomes. The former is also investigated in the role of a mediator variable. Both are treated as discrete quantitative variables. Furthermore, 'prior achievement' and 'type of schooling' (Havo or Vwo) have been identified as control variables and were used for the assignment of the treatment groups. They will be explained and operationalized in the section 'assignment to treatment groups'.

As shown in Figure 2, this research defines behavioral student engagement as ‘intentional exertion of effort’. Furthermore, the literature review mentioned that self-reported measures correlate less with academic achievement than other types of measures (Lei et al., 2018). With this in mind, the chosen measure for behavioral student engagement is the amount of minutes that students spent on their exercises (see section: ‘Data collection for behavioral student engagement (RQ 2)’).

The chosen measure for learning outcomes is students’ final test scores. Prior knowledge is measured and controlled for (see section: ‘Data collection for learning outcomes (RQ 1) and covariates (RQ 1 to RQ 3)’).

Course content
The experiment was conducted during a course on Google Spreadsheets, which is part of the standard curriculum of the students. They learned skills like setting formulas, formatting cells and creating graphs.

The course consisted of 27 instructional videos, each about a minute in length. Additionally it offered 15 practical assignments (together encompassing 70 sub-assignments) where students applied the theory in an actual spreadsheet. See Figure 5 for an example.

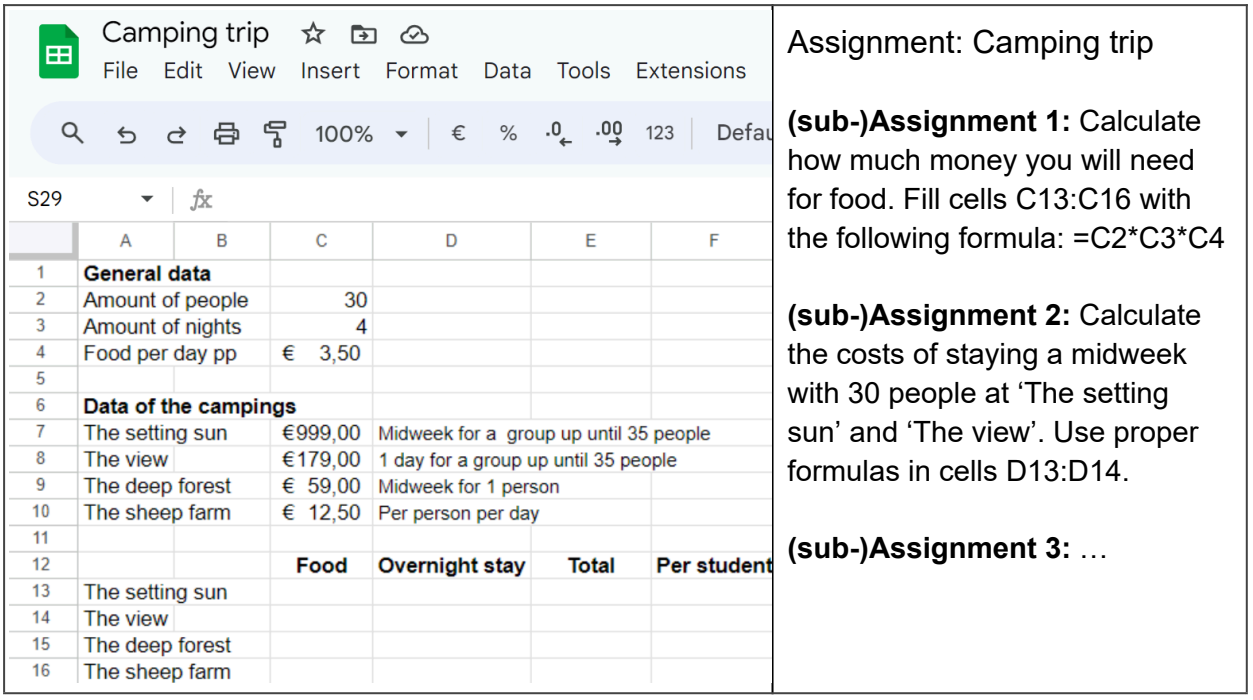

Figure 5: An example of a Google Spreadsheets practical assignment (translated to English)

The exercises and final exam mostly targeted lower-level thinking skills (See Anderson & Krathwohl, 2001). Students needed to understand how to do certain actions and apply this knowledge in new spreadsheets. Students were rarely asked to analyse a situation or to come up with their own strategies to solve a problem.

The role of the teacher
Each lesson the teacher began a plenary introduction of about five minutes where he discussed practical matters and introduced the new topics of Google Spreadsheets. The next 52 minutes, students would autonomously watch the videos and work on the

assignments. The teacher repeatedly walked a fixed route around the classroom to help any student that requested it. The students were also encouraged to ask each other for help. During the final three minutes of each class, the teacher concluded the lecture and assigned homework. The teacher never proactively checked any student homework. This mimics the realistic setting where teachers do not have the time to do so (benefitting ecological validity).

The control group and experimental group

The experiment consisted of a control and an experimental group. Apart from the experimental treatment, the groups have been treated as similarly as possible:

- Both groups received the same number of lessons.
- All students received the same instructional videos, assignments and course information.
- The teacher strictly maintained the same structure and teaching style for all lessons. The teacher was aware of potential performance bias and researcher bias and aimed to minimize them;

The experimental treatment

In contrast to the control group, the experimental group was offered an extra button in their Google Spreadsheet, labeled ‘Check homework’ (see Figure 6). At the beginning of the course, all classes in the experimental group received an explanation of how to use the button.

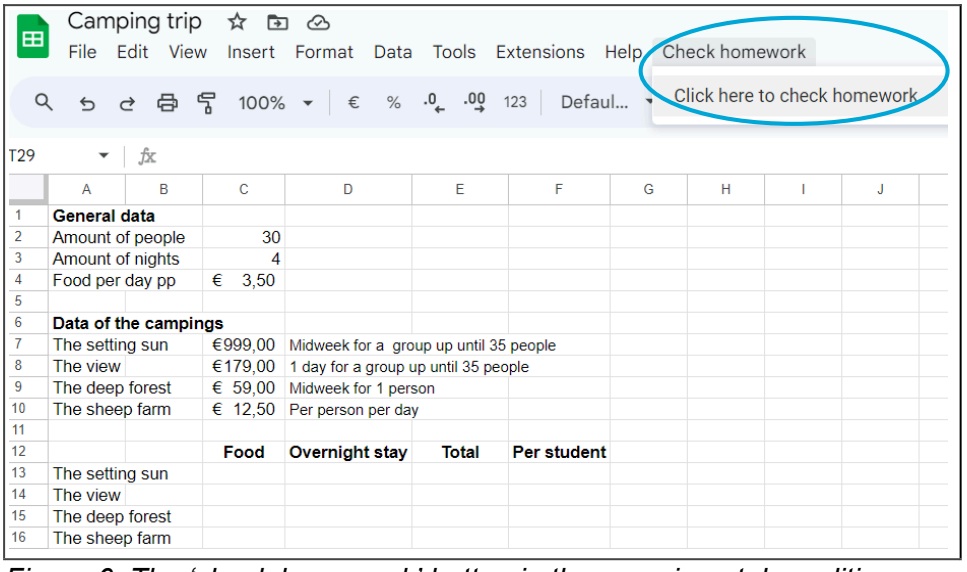

Figure 6: The ‘check homework’ button in the experimental condition (translated to English)

Upon pressing this button, students’ homework is programmatically checked and a new sheet ‘Evaluation’ is added. This sheet contains a feedback table, tailored to the student's work (see Figure 7). The feedback adheres to the theory outlined in the theoretical framework:

- All feedback is specific to the sub-assignment at hand and therefore fits the learning context. The feedback occasionally refers back to previous teachings.

- When a sub-assignment is correctly made, the feedback repeats the learning goal, rather than solely giving feedback at the self-level (e.g. ‘well done!’).

- The feedback can be classified as ‘high-information feedback’. It always starts at the task level by stating what is correct and incorrect. Additionally, it often provides ‘tips’ that offer cues or advice on how to approach the problem (feedback on the process level). When some sub-assignments remain incorrect, the final evaluation gives advice regarding self-regulation. Namely, to reread the feedback or to ask their classmates, family or teacher.

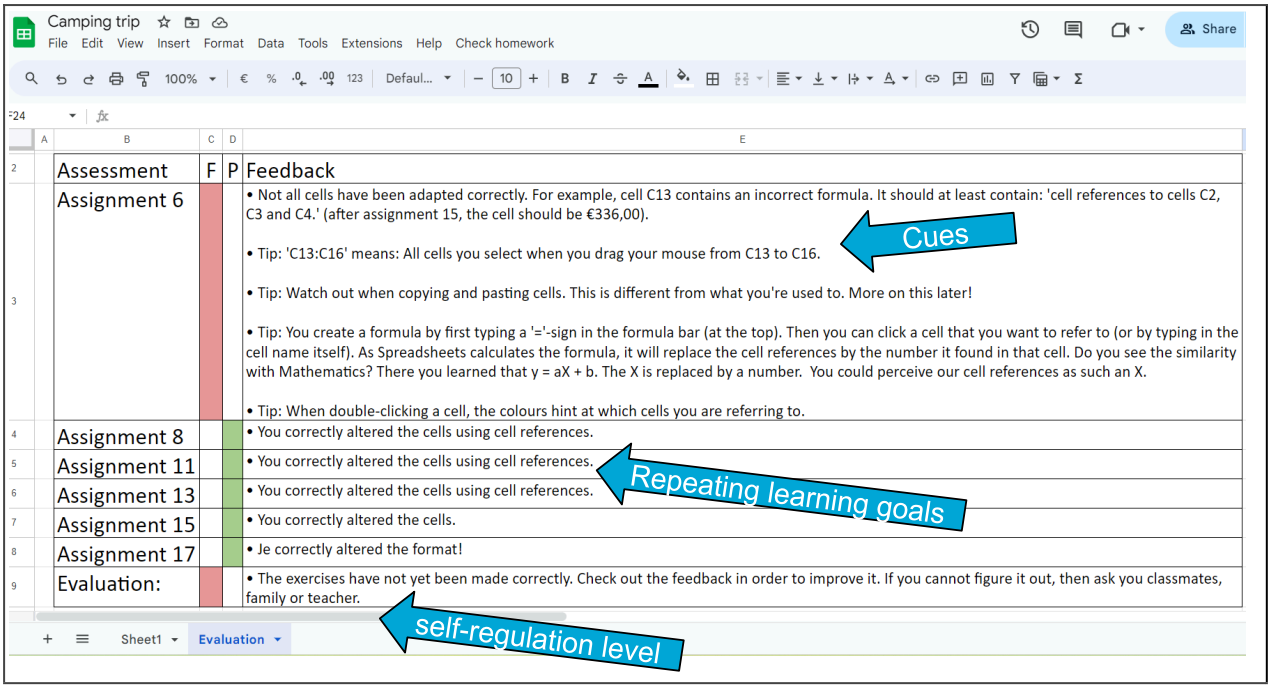

Figure 7: An example of automated feedback provided to students (translated into English). The blue arrows are not part of the feedback but serve as annotations, showing how the theory from the theoretical framework has been integrated.

Furthermore, the experimental condition can be classified as formative assessment according to the theoretical framework in Figure 1:

- The learning goals are stated at the start of each chapter (1);
- The assignments elicit evidence for student understanding (2);
- The students are provided computer assisted feedback that move them forward (3);
- Students are encouraged to help each other, with or without using the feedback (4);
- The feedback on self-regulation level activates students as owners of their own learning (5).

Finally, some notes on the application. The front- and backend of the application have been coded using Google Apps Script and no machine learning tools (like generative AI) have been used. Assignments about formulas and text are checked using regex expressions that often allow small errors. Formatting is usually checked by obtaining booleans from the api (e.g. text has either been made bold or it has not). The code has been manually written for each assignment rather than being generated. However, functions for most assignments have been generalized for reusability. Appendix 1 provides a detailed high-level explanation of how the application works.

4.2.2 Assignment to treatment groups

The experiment aims to minimize interference with the participants' natural environment, which is beneficial for ecological validity. To this end, the assignment of participants to the control or experimental group was not conducted randomly. The reason is that students are part of a predetermined group formation and random assignment would mean that some students in the same classroom would get the experimental treatment and some would not. Such an explicit difference between students could induce all sorts of biases related to feelings of unfairness, perceived expectations and other social dynamics.

Instead, a nonequivalent group design has been conducted where the six existing group formations were kept intact. In an attempt to make the control and experimental group as similar as possible, different variables were investigated in the literature as potential control variables.

An important variable according to literature is ‘prior achievement’. Hattie (2023) demonstrated prior ability in a similar subject to be a powerful predictor for future student achievement and Splett et al. (2018) found academic performance to be a significant predictor of student behavioral and emotional risk. The latter one is particularly interesting in the context of second grade classrooms since students heavily influence each other's learning environment, giving some classes an advantage over others. The variable of prior achievement has been used in two ways.

To control for prior achievement in a similar subject, a pre-test has been conducted. During the first lesson, the students took a test that covered all learning goals of the course.

To use prior achievement to control for differences in classes regarding behavioral and emotional risk, students’ average scores of all courses in their first quarter have been collected. The usefulness of this statistic was demonstrated by a quantitative analysis on data from previous year’s students at this school (n = 124) which showed a significant correlation between their performance in the first quarter and their Spreadsheets scores ($r_s(122) = .546$, $p < .001$).

Finally, the experiment controlled for the type of schooling (Havo or Vwo). An independent-samples Mann-Whitney U test showed that Vwo in the previous year had a significantly higher average Spreadsheet score (7.9) than Havo (6.9), U = 2339 , p = .002.

The literature shows more potential control variables like: socioeconomic status (Sirin, 2005), Critical thinking (Orhan, 2022) and Ethnicity (Splett et al., 2018). These were not included due to both ethical concerns and to avoid overburdening students with questionnaires.

4.2.3 Statistics on the treatment groups

Table 1 shows the statistics on control variables of the treatment groups. The Mann-Whitney U tests imply that the groups are indeed far from different. Figure 8 shows the distribution in type of schooling and (for completeness) students’ gender across the treatment groups and shows their similarity.

	Control group	Experimental group	U-value	p value
Pre-test scores	mean: 4.3 SD: 1.6	mean: 4.5 SD: 1.4	1745	.560
Average first quarter scores	mean: 6.8 SD: 0.7	mean: 6.7 SD: 0.8	2009.5	.774

Table 1: Statistics on control variables of the treatment groups ('SD' stands for 'standard deviation'). An independent-samples Mann-Whitney U test has been conducted to compare the control and experimental groups.

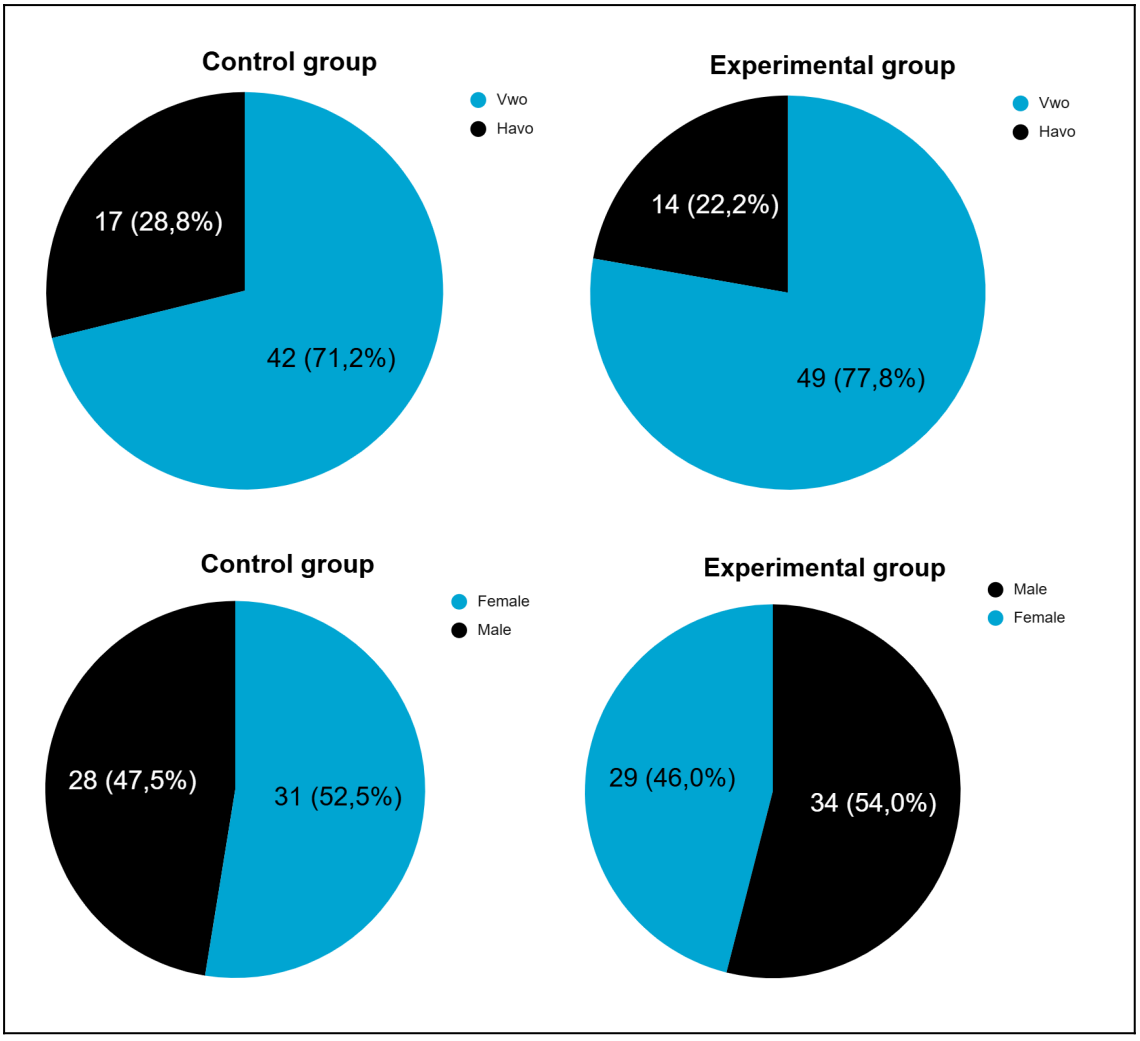

Figure 8: Level of schooling and gender distribution of the treatment groups

4.2.4 Data collection for learning outcomes (RQ 1) and covariates (RQ 1 to 3)

To measure the learning outcomes, the students took a pre-test and post-test. The tests consisted of 15 assignments covering all learning goals. The students had 60 minutes to complete each test individually, without access to any study materials, and were supervised to enforce this. See appendix 2 for the standardized instructions that the students received for the pre-test. Only the post-test was part of the official exam week and academic record.

The teacher graded the tests. Grading Google Spreadsheets is partially subjective which introduces the risk of various biases. For one, it could lead to performance and researcher bias since the grader is also the researcher. Even apart from bias related to the content, Malouff and Thorsteinsson (2016) showed that irrelevant information about the students, such as educational deficiency or ethnicity, may also lead to biased results.

To mitigate the risk of bias, the tests were programmatically graded by the application developed with Google Apps Script, meaning that all students have been assessed by the same objective criteria.

Finally, data of the covariates ‘average first-quarter scores’ and ‘type of schooling’ were obtained directly from the educational administration system ‘Magister’.

4.2.5 Data collection for behavioral student engagement (RQ 2)

The amount of minutes spent on exercises was collected by analyzing the timestamps in the version history of all students’ spreadsheets. Consecutive timestamps with a difference of less than 6 minutes were considered continuous work. Otherwise it was considered a break and excluded from the total amount of minutes spent.

Towards the end of the course, some students completed all the exercises ahead of schedule, posing the risk of a ceiling effect. This risk has been mitigated by adding 5 extra assignments, comprising 23 sub-assignments, that provided extra practice but did not introduce new material.

4.2.6 Data collection for the mediation analysis (RQ 3)

The input for the mediation analysis was the same data that was collected for RQ’s 1 and 2.

4.2.7 Data collection for students’ experiences (RQ 4)

In order to provide deeper insights on relations *a*, *b* and *c*, students in the experimental group were asked to fill in a questionnaire to share their experiences (see appendix 3). In addition to questions about the stated relationships, students have also been asked about the extent to which they use the feedback, why they use it, how they use it and what their general opinions about it are. These questions were aimed to elicit practical information that could support the data of interest.

The students were informed that their responses would remain anonymous to the outside world, but not to the teacher. The questionnaire consisted exclusively of open questions (except for the first question) and the students were encouraged to actually think before they typed out their answers. The students got 20 minutes to fill it in but they all finished sooner.

4.3 Data analysis

4.3.1 The mediation analysis (RQ 1 to RQ 3)

The mediation analysis has been conducted by using Hayes’ (2022) tool called PROCESS (using ‘model 4’). The author thoroughly describes the procedure in his book.

Assumptions of linear regression
The procedure involves several linear regressions. Appendix 4 provided a detailed explanation of how the data were tested for the following assumptions:

- Linearity
- Normal distribution of residuals
- Homoscedasticity
- Independence (i.e. no autocorrelation)
- multicollinearity

The data were deemed suitable for the linear regressions.

Variables
Computer assisted formative assessment served as the independent variable, encoded as 0 for the control group and 1 for the experimental group. The number of minutes spent on exercises was used as the mediator variable, while final test scores represented the dependent variable.

Additionally, the pre-test scores, which were used during the assignment to treatment groups, were also used as a statistical control to provide an additional layer of control. As an extra measure, the analysis was also conducted with the first-quarter scores and type of schooling (Havo encoded as 0 and Vwo as 1) as additional covariates of which the results are presented in appendix 5.

Unstandardized and (partially) standardized effects
The analysis estimates all relations shown in the conceptual framework, addressing RQ 1 to RQ 3. It distinguishes between a total, direct and indirect effect. The total effect is equivalent to relation *c*, the direct effect to relation *c'* and the indirect effect to the product of relations *a* and *b*.

The unstandardized effects (expressed by the regression coefficient) shows how much a dependent variable changes after the dependent variable changes by one unit. They require some domain knowledge to interpret.

The standardized effect expresses how many standard deviations the dependent variable changes after the independent variable changes by one standard deviation. Since standardized effects are easier to compare to other research results, they are more suitable as a measure for effect size.

However, standardized effects make less sense when an independent variable is dichotomous. More suitable are partially standardized effects, which express how many standard deviations the dependent variable changes after the independent variable changes by one unit.

This research presents unstandardized, partially standardized and (in appendix 5) standardized effects.

Inference
In order to avoid the assumptions of normality, a 95% bootstrap confidence interval method was employed for inference. The data were resampled with replacement 50,000 times and the statistics of interest were calculated for each sample to form a distribution. The 2.5th and 97.5th percentiles of this distribution defined the lower limit confidence interval (LLCI) and the upper limit confidence interval (ULCI) respectively. If the 95% bootstrap confidence interval did not contain zero, the result was considered significant.

4.3.2 Thematic analysis (RQ 4)
The qualitative data from the questionnaire were analyzed using a thematic analysis approach, as described by Braun and Clarke (2006). An inductive and semantic approach was adopted, allowing the themes to emerge from the data while analyzing the explicit content of the data (in contrast to reading into subtext).

First, all data was skimmed through to get familiar with it. Sentences or paragraphs were assigned a single label, or ‘code’, that reflects their meaning. After that, the different codes were checked for connections and patterns, after which they are combined into ‘themes’. These themes are described in the results section. The whole process has been iterative; codes and themes have been reviewed and changed multiple times.

5. Results

This section outlines the key findings from the data collection and analysis, as described in the methodology. First, the results of the mediation analyses are presented, addressing RQ's 1 to 3. Then, the themes emerging from the thematic analysis are described. They provide insights into students' experiences with the experimental condition.

Appendix 5 provides a supplementary results section that goes into results that are insightful but not directly relevant to answering the research questions. It starts with descriptive results to provide intuition about the data (e.g. means and standard deviations). It then provides additional information about specific regression analyses while discussing significant, but irrelevant effects found. These results are presented for both the mediation analysis shown in this section, as for a mediation analysis including 'first quarter scores' and 'level of schooling' as covariates.

5.1 Results of the mediation analysis (RQ's 1 to 3)

First, Figures 9 and 10 visually present the results of the mediation analysis by integrating them into the conceptual framework shown in Figures 3 and 4. For clarity, variable names have been replaced with their operationalized counterparts. The arrows represent the regression coefficients and indicate how many points or minutes the dependent variable is expected to change when the independent variable increases by one unit, while controlling for other incoming arrows. For example, Figure 10 shows that the use of automated feedback predicts a (non-significant) decrease of 11.322 minutes spent on exercises while controlling for the pre-test scores.

One important observation is that the data of the 'pre-test scores' are mostly concentrated within the range of 3.3–5.2. Since regression models can only reliably predict outcomes within the range of the data used to create them, the regression coefficient should be interpreted with care. See appendix 5 for more details on the data distribution.

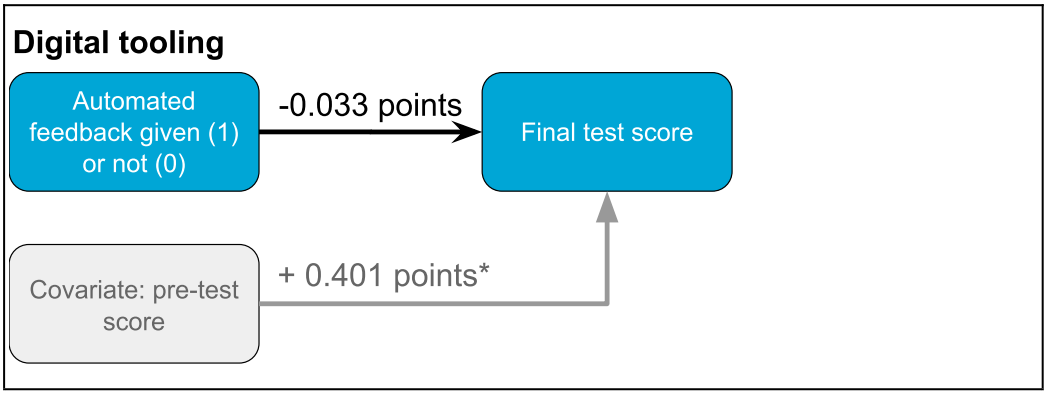

Figure 9: The results of the mediation analysis integrated in part 1 of the conceptual framework as presented in Figure 3. Significant results on a 95% bootstrap confidence interval are denoted by an asterisk (*)

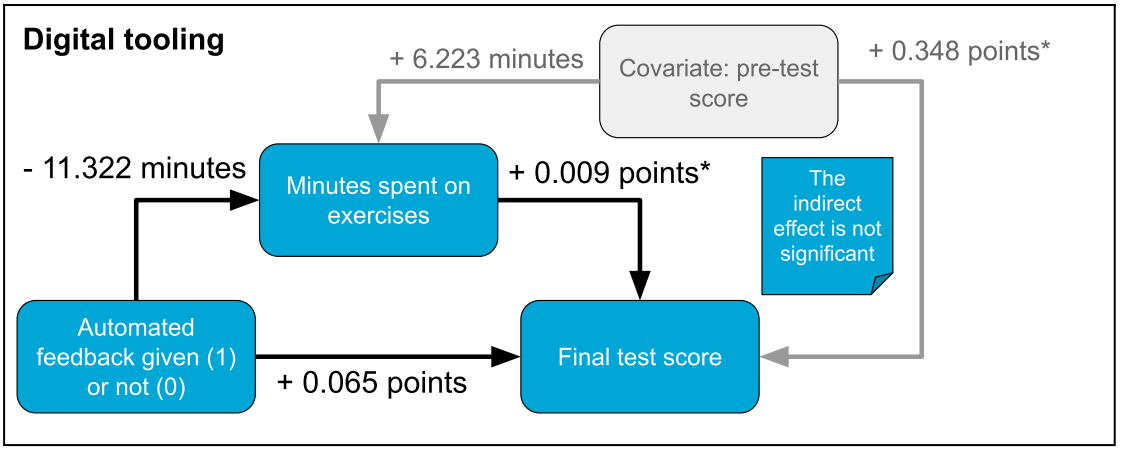

Figure 10: The results of the mediation analysis integrated in part 2 of the conceptual framework as presented in Figure 4. Significant results on a 95% bootstrap confidence interval are denoted by an asterisk (*).

The three regression analyses were all statistically significant:

- For the total effect: R^2 = .160, F(2, 119) = 11.341, p < .001.
- For behavioral student engagement as the outcome variable: R^2 = .025, F(2, 119) = 1.556, p = .215.
- For learning outcomes as the outcome variable: R^2 = .302, F(3, 118) = 16.991, p < .001

Table 2 presents the results regarding the direct, indirect and total effect and relation *a*. Just like in figures 9 and 10, the coefficients should be interpreted as the change in the dependent variable when the computer assisted formative assessment is applied compared to when it is not.

Total, direct and indirect effects and relation *a*				
Type of effect	**Coefficient (*B*)**	**Partially standardized coefficient (β)**	**LLCI**	**ULCI**
Relation *a*	-11.322	-0.170	-35.154	12.173
Direct effect (c')	0.065	0.043	-0.376	0.503
Indirect effect (ab)	-0.098	-0.065	-0.321	0.105
Total effect (c)	-0.033	-0.022	-0.534	0.469

Table 2: The direct, indirect and total effect resulting from the mediation analysis

Table 2 shows that neither the total effect, direct effect, nor indirect effect is significant on a 95% bootstrap confidence interval, providing insufficient evidence to reject the null hypotheses H_{10} and H_{30}. The same goes for the relation between computer assisted formative assessment and behavioral student engagement (relation *a*), providing insufficient evidence to reject the null hypothesis H_{20}.

5.2 Students' experiences (RQ 4)

This section presents the experiences of students in the experimental group regarding relations *a*, *b* and *c,* as shown in figures 3 and 4, by describing the themes that arose from the thematic analysis. The analysis is based on the responses of 62 students. It starts with brief contextual information about students' usage of the feedback.

Then the section presents the themes related to how the feedback influenced students' engagement and learning outcomes (i.e. relations *a* & *c*). It first discusses themes with a positive sentiment, followed by themes with a non-positive sentiment (neutral or negative).

Finally, the section provides a brief overview of the themes related to how the assignments influenced students' learning outcomes (i.e. relation *b*). This data is presented concisely as its purpose is mostly to support the other findings and the findings were largely consistent across responses.

Occasionally, this section uses the words 'few', 'some', 'many' and 'most' to indicate the prevalence of certain responses, with the following meaning:

- Few: two to four responses
- Some: four to ten responses
- Many: ten to 31 responses.
- Most: 31 or more responses.

Note that all student quotes in this section have been translated to English by the researcher while remaining as faithful as possible to their original phrasing and meaning.

5.2.1 Contextual information about the usage of the feedback

Figure 11 illustrates that most students use the 'check homework' button regularly. Students give various reasons for using or not using it and most of them are related to the themes discussed in the rest of this section. Three additional reasons to not use it are that they didn't know how it worked, often forgot, or they simply didn't feel like using it.

With respect to *how* they used it, many students reported checking their work after completing an entire assignment and a few checked themselves after each sub-assignment. One student emphasized that he used the feedback to support an iterative learning process of doing, checking and improving.

Illustrative quotes are:

- "I don't use it much because I didn't really think about using it"
- "Sometimes I check in between to see if I did something right or wrong, and sometimes only at the end (when I'm done)"
- "When I'm ready I press the button and see whether I got it correct. If I didn't get it correct then I try to improve it and press the button again to see if I got it correct that time around"

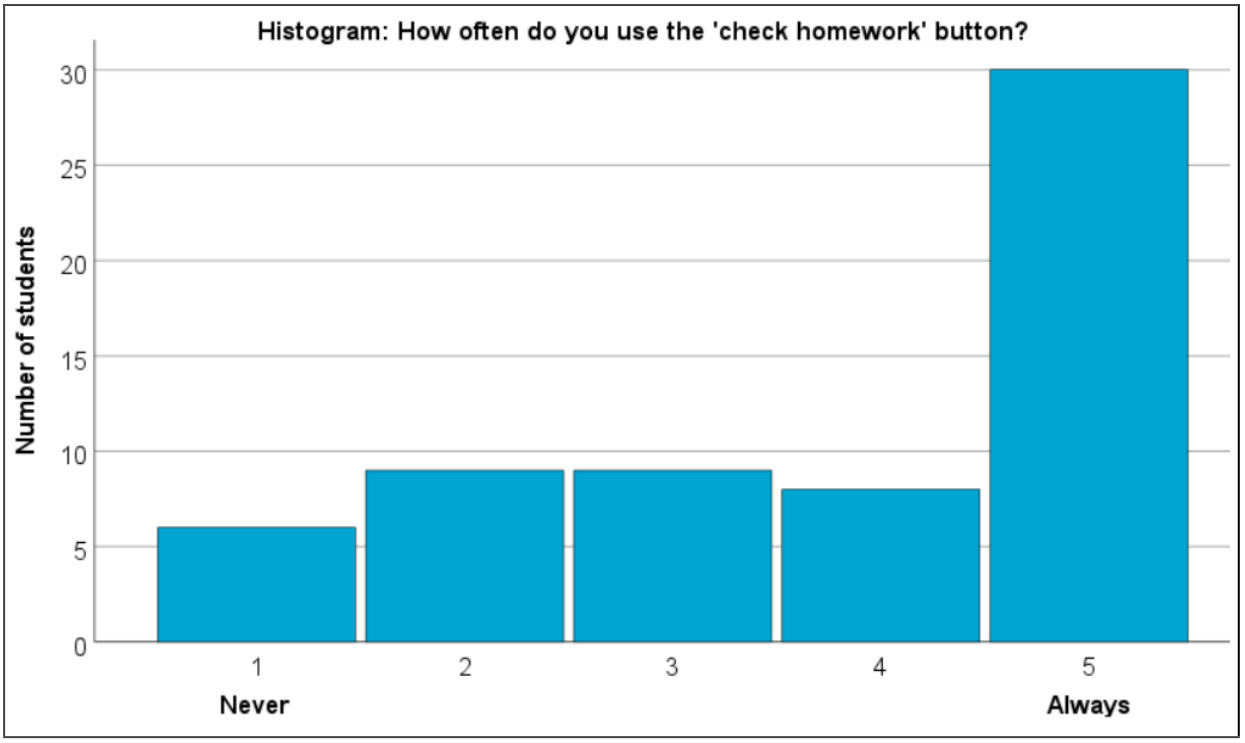

Figure 11: Responses of students in the experimental group on how often they use the 'check homework' button (n = 62)

5.2.2 Positive themes related to relations *a* and *c*

Enhanced motivation
Students have indicated the feedback to increase their motivation in several ways and therewith their perceived engagement and learning outcomes. Note that 'motivation' refers to the inclination and drive to learning whereas 'engagement' refers to the behaviors that reflect this inclination (Martin et al., 2017).

Some of the students described an increase in extrinsic motivation, noting that the teacher would now be able to see their progress and stating that they wanted to avoid repercussions. Another external stimulus was for the feedback to help them increase their final test score.

Other students described an increase in intrinsic motivation. Some of them indicated that they were more motivated to do the exercises now that they had feedback to look forward to. Some students felt an internal drive to get everything correct or "green", purely for the sake of it.

Illustrative quotes are:

- "I'm now extra curious to see how to do it correctly".
- "It's a bit more like a game now".
- "I do it because the teacher makes me, otherwise there will be consequences"[2].

Enhanced self-regulated learning
When asked about engagement and learning outcomes, many students' responses indicated an increase in their self-regulated learning. Self-regulated learning is about how students

[2] Note that this was the students' experience even though the teacher didn't check any assignments.

activate their cognitions, affects and behaviors to reach their personal learning goals (Zimmerman & Schunk, 2011).

Many students appreciated the improved insight into what they had done right or wrong, what they did or did not understand, or which exercises they had or had not completed. This way they could act accordingly. A few students said that they revised the material as a consequence of the feedback. Some students stated that they would go back to improve or redo exercises that they had not perfected yet.

Illustrative quotes are:

- “I often feel like I’m doing well but then it turns out to be wrong. It gives me the feeling that I have to check everything twice”.
- “When I do something wrong I revise the material”.
- “I do more of the assignments because I can see which ones I haven’t done yet”.
- “I like that you can check your own homework and to know what you did wrong or don’t understand properly”.

Enhanced student autonomy
Some students appreciated that they could get the feedback when- and wherever they wanted. They preferred to be slightly less dependent on the teacher in this way and indicated that it improved their engagement and learning outcomes.

Illustrative quotes are:

- “I like it because the teacher cannot help everyone at the same time so in that case you can also use the feedback”.
- “I definitely like it when I’m not at school but at home. It is definitely good to have”.

Feedback helps students move forward
Many students appreciated the feedback and assignment-specific tips after they got an exercise wrong; it helped them move forward and improve the quality of their work. This both increased their perceived engagement and learning outcomes.

Illustrative quotes are:

- “It makes me do more of the assignments. Because of the tips I know what to do and then I do them”
- “The feedback mostly makes me do better on the assignments [as opposed to doing more of them]”.

Improved student understanding
Many students indicated that the feedback helped to improve their learning process and their understanding of Google Spreadsheets (relation *c*). Students have stated that they learn both from their mistakes and from what they did well.

One student noted that “it helps me know whether I misunderstood something”, implying that the feedback helps in dealing with misconceptions. Another student said: “You learn more and you better understand the assignments”.

5.2.3 non-positive themes related to relations *a* and *c*

Increased time consumption
Some students indicated that reading and applying the feedback means that the assignments take more time to finish. For some, this results in completing fewer assignments, while for a few, it leads to disregarding some of the feedback altogether.

Illustrative quotes are:

- “On the one hand you learn more but on the other hand I just want to be done with it so I usually just rush it”.
- “I barely use it because I’m slow at doing homework so I don’t often get the chance to check it”.
- “I finish less assignments because by improving the feedback you don’t get to do another task right away”.

Lack of added value
Many students indicated that the feedback didn’t change their engagement, since they would have to do the assignments anyway. A few students found the feedback to be redundant, since it mostly repeats the initial instruction. Finally, a few students found feedback from a computer to be inferior to the feedback of a teacher.

Illustrative quotes are:

- “You don’t learn more or less per se, because the things mentioned by the button/feedback are often also mentioned by the teacher I think”.
- “If the button would not be there then I would still just do the assignments. So for me it doesn’t really make a difference”.
- “When you’re working on the assignment then a teacher can help you way better than a computer”.

Lack of feedback quality
Some students found the feedback to be rather unhelpful. Several described it to be vague, unclear or non-specific to their problem. Some noted that the feedback is mostly focused on what is wrong, and would have preferred more instruction on how to correct the assignment. One student noted that, since the feedback was mostly negative in sentiment, it could lower students’ self-esteem. Finally, a few students mentioned that the feedback was unnecessarily strict and unforgiving to mistakes that they perceived to be negligible.

Illustrative quotes are:

- “It’s nice to know what you did wrong so you can practice it more but it doesn’t really clearly say what the correct way of doing it is and that is annoying”.
- “I check out the feedback but often don’t understand what it is trying to tell me”.
- “Usually it’s just about careless mistakes”.
- “I would prefer an orange color instead of red when you get it partially correct because that is more encouraging”.

5.2.4 themes related to relation *b*
Most of the students found the practical exercises beneficial for their learning outcomes. They saw value in the opportunity to practice and to apply the theory in actual spreadsheets.

A few students specifically mentioned that the practical assignments stimulated them to think more deeply about the material and one student said that the practical exercises made the learning more fun.

On the other hand there were some students who found the assignments unclear and a few argued that there were too many of them.

Illustrative quotes are:

- “It makes me think more carefully about the assignment and how it works”.
- “Practicing things helps me to use it”.
- “With the practical assignments I don’t only see what I have to do but I can also apply it myself and see where I have to click and what happens when I click on a certain option”.

- “It’s a bit much and I don’t learn much from it”
- “Usually I don’t understand it which simply makes me randomly guess when doing the assignments”.

6. Discussion

This research investigated the effects of computer assisted formative assessment on behavioral student engagement and learning outcomes, as well as the potential mediating role of behavioral student engagement, in the subject domain of digital tooling. Objectively, no significant effects were found in these relationships. However, subjectively, students did perceive several benefits from the computer assisted formative assessment. These included cognitive benefits, like better understanding of the material, and metacognitive benefits, including motivation, self-regulated learning and autonomy. Students also perceived some downsides like a greater time consumption, a lack of added value and a lack of feedback quality.

This section explains the meaning and relevance of this research's results. It begins by interpreting the findings and discussing their implications in relation to existing theory, separately addressing learning outcomes, behavioral student engagement, the mediation model, and students' experiences. Finally it discusses limitations of this research and provides recommendations for future research.

6.1 Learning outcomes

Contrary to the expectation of this research, the findings do not provide evidence that the provision of computer assisted formative assessment in the subject domain of digital tooling influences student learning outcomes any more than not providing it. Since the literature review identified multiple studies reporting effect sizes for computer assisted formative assessment, ranging from 0.21 to 0.29 standard deviations, the findings of this research challenge the idea that computer assisted formative assessment is an universally effective intervention.

Since the effect sizes from the literature were based on data from subject domains other than digital tooling, the absence of effect in this research may be explained by the distinct nature of this subject domain. A possible explanation is that, unlike other subject domains, the learning of digital tooling requires little memorization (as in languages), creativity (as in the arts) or conceptual understanding (as in mathematics). Instead, students mostly need to find and click the right buttons, which can be reinforced through practice until it becomes second nature, or may be attempted on the fly by trial and error while utilizing investigative skills. This may be more relevant in this study as the course primarily required lower-level thinking skills; as noted in the literature review, while students should still benefit from feedback in this case, its effectiveness is likely lower compared to when higher-level thinking skills are required.

6.2 Behavioral student engagement

Similarly, the findings did not indicate a change in behavioral student engagement with the provision of computer assisted formative assessment in the subject domain of digital tooling, compared to when it was not provided. This sheds extra light on the findings of Barana et al. (2019) (see literature review) who did find such an effect within the subject domain of mathematics. This raises the question whether their results are generalizable to other subject domains like digital tooling.

A possible explanation is that exercises in the subject domain of digital tooling often already provide a feedback loop without explicit computer assisted formative assessment. The exercises request specific actions of students while often providing an example of what the end result should look like. Students in the experimental group mentioned that ‘getting everything correct’ is a motivational factor but this same factor may apply to students in the control group who insisted on finding the right buttons or replicating the provided examples. This may be one of the reasons that both treatment groups have invested a similar amount of time in their tasks. An additional explanation may be that the sample group consisted solely of Havo and Vwo students (with the majority from Vwo) who are typically already driven to do well in school. Therefore, the computer assisted formative assessment may have made little difference.

6.3 Behavioral student engagement as a mediator variable

The literature review identified a link between formative assessment and student engagement, as well as between student engagement and learning outcomes. Based on this, it was expected that student engagement would mediate the relationship between formative assessment and learning outcomes. However, the findings of this research did not support such a relationship. The only significant relationship observed was that an increase in student engagement correlated to an increase in learning outcomes, aligning with the findings of Lei et al. (2018). Possible explanations and implications are similar to the ones discussed in the previous section on ‘behavioral student engagement’.

6.4 Students’ experiences

Although no objective improvements in behavioral student engagement or learning outcomes were found, students’ experiences were largely positive, suggesting that the intervention may have yielded benefits that were not captured by this research.

In the literature review, Leenknecht et al. (2020) highlighted several benefits of formative assessment, such as increased feelings of autonomy, competence, and relatedness, which in turn enhance motivation. In this research, students reported all of these benefits except for relatedness. This suggests that these literature findings may also be applicable to the subject domain of digital tooling.

Furthermore, the enhanced autonomy, together with the enhanced self-regulated learning seems to indicate an increase in ‘agentic engagement’ as described in the literature review.

Moreover the students reported that the feedback helped them move forward and improved their understanding. These findings provide some evidence for this research’s construct validity, as they align with how formative assessment is defined in Figure 1.

On the other hand, a minority of students reported downsides. An interesting one is the lack of feedback quality. Feedback not being specific or clear enough seems like a typical characteristic of feedback that has been programmed beforehand, since it has to remain general enough to cover all possible scenarios. Further research using generative AI, such as Tobler’s (2024) research, referenced in the literature review, may help solve this problem.

Another downside is that some students expressed resistance to getting feedback from a computer and preferred human feedback. This implies that computer assisted formative assessment should preferably be used as an supplementary tool in the classroom rather than fully replacing teacher-provided feedback. Finally, some students noted that the assignments took longer to complete, potentially reducing the number of exercises they could finish. If this proves to be a significant issue, a possible solution would be to allow students the choice of whether or not to use the automated feedback.

6.5 Limitations and recommendations for future research

This section lists key limitations of this research, followed by recommendation for future research to address them. Note that these limitations do not undermine the validity of this research but rather clarify what can and cannot be concluded from it. Most of the limitations arise from unmeasured variables.

For one, even though student experiences suggest several benefits of computer assisted formative assessment, this research cannot make definitive inferences about them. To verify these benefits as significant in the domain of digital tooling, further research is needed to quantify their impact.

Furthermore there may be extraneous variables present that this research did not take into account. For example, the final test in this research provided ample time for students to complete it, which may have contributed to the high average scores shown in appendix 5. Similarly, students may have considered the final test to be of low difficulty. Future research could explore the effectiveness of computer assisted formative assessment in the subject domain of digital tooling under varying time constraints, difficulty levels and levels of required thinking skills.

Moreover, an interesting argument by Kirschner et al. (2006) is that final test scores as measured in this research may not be an appropriate measure of learning, since it is focused on short-term achievement. According to the authors, instruction should be aimed at changing long-term memory. Future research could address this by conducting follow-up tests at later moments in time to determine to what extent computer assisted formative assessment contributes to long-term retention.

Another potential limitation relates to the data collection of behavioral student engagement. Since this research relied solely on version history, it only captured students' activity within the spreadsheets, without accounting for other forms of engagement, such as watching instructional videos, thinking about the assignments, or discussing them with classmates. Future research could adopt more comprehensive methods, such as direct observation, possibly combined with self-reported measures, to gain a more complete picture of the behavioral student engagement. However, in high school settings, differences in behavioral engagement are often evident in how much of the exercises students complete, as students regularly tend to skip (parts of) entire assignments. Since this pattern is effectively captured through version history, it is likely that this research still provides an accurate representation of behavioral student engagement.

Finally, the relatively small sample size of 122 students has limited this research's statistical power, potentially contributing to the lack of significant findings. Especially considering the high variability of the data (see appendix 5) and small effect sizes found. Future research could verify the findings by conducting studies with larger sample sizes. However, given the small effect sizes observed, achieving statistical significance may have limited practical relevance.

7. Conclusion

This research investigated how computer assisted formative assessment affects students' behavioral engagement and learning outcomes, including the potential mediating role of behavioral student engagement, in the second grade of Dutch secondary education. As literature has shown these effects to be significant across several subject domains, this research investigated to what extent they hold up within the subject domain of digital tooling.

A quasi-experiment divided six second-grade classes of a Havo/Vwo high school into equivalent groups for a course on Google Spreadsheets, with only one group having unlimited access to a button that let the computer instantly check their homework and provide feedback. At the end of the course, both groups spent an equivalent amount of time on the exercises and achieved equivalent scores on the final tests. Consequently, no significant effects on behavioral student engagement or learning outcomes were found, nor was a mediated relationship established.

Additionally, the students with access to the button completed a questionnaire to share their experiences. Most responses were positive, with reports of both cognitive and metacognitive benefits. Namely: enhanced motivation, enhanced self-regulated learning, enhanced autonomy, feedback that helped them move forward and an improved understanding of the material. However, this was not the case for all students, as some reported a lack of benefits, or even negative consequences of the intervention. Namely: increased time consumption, a lack of added value and a lack of feedback quality.

This research raises the question of whether the subject domain of digital tooling interacts differently with computer-assisted formative assessment than other subject domains in secondary education, thereby diminishing the benefits for student engagement and learning outcomes. Further research with a greater sample size and statistical power is needed to explore this possibility. Future research could explore the distinct characteristics of this subject domain and investigate whether and how computer assisted formative assessment can be employed to benefit students. It could focus both on different contexts (e.g. difficulty levels within the subject domain) and what benefits should be targeted.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives : complete edition/*. Pearson Education. https://eduq.info/xmlui/handle/11515/18824

Barana, A., Marchisio, M., & Rabellino, S. (2019). *Empowering Engagement through Automatic Formative Assessment* (Vol. 1). IEEE. https://doi.org/10.1109/compsac.2019.00040

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education Principles Policy And Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594x.2010.513678

Black, P., & Wiliam, D. (1998a). Inside the Black Box: raising standards through classroom assessment. *School Of Education, King's College*.

Black, P., & Wiliam, D. (1998b). Assessment and Classroom Learning. *Assessment in Education Principles Policy And Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal*, *29*(5), 623–637. https://doi.org/10.1080/0141192032000133721

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation And Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education Principles Policy And Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594x.2018.1441807

Bloom, B. S., Hastings, T., & Madaus, G. F. (1971). *Handbook On Formative and Summative Evaluation of Student Learning*. https://www.jstor.org/stable/pdfplus/1434189.pdf

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brookhart, S. M. (2018). Summative and Formative Feedback. In *Cambridge University Press eBooks* (pp. 52–78). https://doi.org/10.1017/9781316832134.005

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment, Research, And Evaluation*, *14*(1), 7.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review Of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Fredricks, J. A., Wang, M., Linn, J. S., Hofkens, T. L., Sung, H., Parr, A., & Allerton, J. (2016). Using qualitative methods to develop a survey measure of math and science engagement. *Learning And Instruction*, *43*, 5–15. https://doi.org/10.1016/j.learninstruc.2016.01.009

Hattie, J. (2023). Prior achievement [Dataset]. In *Visible learning meta^x*. Visible Learning+. https://www.visiblelearningmetax.com/influences/view/prior_ability_&_intelligence

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review Of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hattie, J., & Zierer, K. (2019). Visible learning insights. In *Routledge eBooks*. https://doi.org/10.4324/9781351002226

Hayes, A. F. (2022). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. Guilford Publications.

Irons, A., & Elkington, S. (2021). *Enhancing Learning through Formative Assessment and Feedback*. https://doi.org/10.4324/9781138610514

Kingston, N., & Nash, B. (2011). Formative Assessment: A Meta-Analysis and a Call for Research. *Educational Measurement Issues And Practice*, *30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Kirschner, P., Clark, R., & Sweller, J. (2006). Work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). Formative Assessment and Elementary School Student Academic Achievement: A Review of the Evidence. REL 2017-259. *Regional Educational Laboratory Central*. https://eric.ed.gov/?id=ED572929

Kuklick, L., Greiff, S., & Lindner, M. A. (2023). Computer-based performance feedback: Effects of error message complexity on cognitive, metacognitive, and motivational outcomes. *Computers & Education*, *200*, 104785. https://doi.org/10.1016/j.compedu.2023.104785

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The Effectiveness and Features of Formative Assessment in US K-12 Education: A Systematic Review. *Applied Measurement in Education*, *33*(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

Leenknecht, M., Wijnia, L., Köhlen, M., Fryer, L., Rikers, R., & Loyens, S. (2020). Formative assessment as practice: the role of students' motivation. *Assessment & Evaluation in Higher Education*, *46*(2), 236–255. https://doi.org/10.1080/02602938.2020.1765228

Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior And Personality An International Journal*, *46*(3), 517–528. https://doi.org/10.2224/sbp.7054

Li, S. (2010). The Effectiveness of Corrective Feedback in SLA: A Meta-Analysis. *Language Learning*, *60*(2), 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Lipnevich, A. A., & Panadero, E. (2021). A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.720195

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal Of Education*, *60*(3), 245–256. https://doi.org/10.1177/0004944116664618

Martin, A. J., Ginns, P., & Papworth, B. (2017). Motivation and engagement: Same or different? Does it matter? *Learning And Individual Differences*, *55*, 150–162. https://doi.org/10.1016/j.lindif.2017.03.013

Natriello, G. (1984). Problems in the evaluation of students and student disengagement from secondary schools. *Journal Of Research And Development in Education*, *17*(4), 14–24.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Orhan, A. (2022). The Relationship between Critical Thinking and Academic Achievement: A Meta-Analysis Study. *Psycho-Educational Research Reviews*, *11*(1). https://doi.org/10.52963/perr_biruni_v11.n1.18

Reeve, J. (2013). How students create motivationally supportive learning environments for themselves: The concept of agentic engagement. *Journal Of Educational Psychology*, *105*(3), 579–595. https://doi.org/10.1037/a0032690

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review Of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

SLO. (2024). *digitale geletterdheid*. Geraadpleegd op 7 september 2024, van https://www.slo.nl/thema/meer/basisvaardigheden/digitale-geletterdheid/

Splett, J. W., Smith-Millman, M., Raborn, A., Brann, K. L., Flaspohler, P. D., & Maras, M. A. (2018). Student, teacher, and classroom predictors of between-teacher variance of students' teacher-rated behavior. *School Psychology Quarterly*, *33*(3), 460–468. https://doi.org/10.1037/spq0000241

Tobler, S. (2024). Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments. *MethodsX*, *12*, 102531. https://doi.org/10.1016/j.mex.2023.102531

University of Notre Dame. (z.d.). Durbin-Watson Significance Tables. In *University Of Notre Dame*. Geraadpleegd op 29 januari 2025, van https://www.google.com/url?client=internal-element-cse&cx=017319754557445836 37:tf-uiraghmq&q=https://www3.nd.edu/~wevans1/econ30331/durbin_watson_tables. pdf&sa=U&ved=2ahUKEwivrtm905mLAxUqhv0HHXFGFdMQFnoECAcQAQ&usg=A OvVaw3i3fbSv150lFbpl_uyiwJA

Van Der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes. *Review Of Educational Research*, *85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Wiliam, D. (2018). Feedback. In *Cambridge University Press eBooks* (pp. 3–28). https://doi.org/10.1017/9781316832134.003

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.03087

Wong, Z. Y., & Liem, G. A. D. (2021). Student Engagement: Current State of the Construct, Conceptual Refinement, and Future Research Directions. *Educational Psychology Review*, *34*(1), 107–138. https://doi.org/10.1007/s10648-021-09628-3

Zimmerman, B. J., & Schunk, D. H. (2011). *Self-Regulated Learning and Performance : An Introduction and an Overview* (pp. 15–26). Taylor & Francis Group. https://doi.org/10.4324/9780203839010-4

Appendix 1: A high-level explanation of how the application for automated feedback works

Google Apps Script

Most Google Apps (like Spreadsheets, Slides, Documents, Drive and Classroom) have access to an extension called ‘Google Apps Script’ which allows users to extend their files with custom code. Google provides an API, enabling users to programmatically interact with their files. The API includes a set of getter-methods (for example to get the current fill colour of a certain cell) and setter-methods (for example to set borders around certain cells).

The architecture

The application in this research embedded a small script in all students’ spreadsheets which sends the assignment name and spreadsheet-id to the server when the ‘check homework’ button is pressed. The server then evaluates the assignment in the spreadsheet and updates it by adding a feedback table.

Checking students’ work

In broad terms, the application checks the assignments as follows. First, the application checks a few preconditions that are required to properly check the assignment. For example, it checks the title of each sheet to identify it. If the precondition is not met, the student is prompted to fix the issue through the feedback.

Then the application starts checking the assignments themselves. It uses getter-methods to check each requirement set by the assignment. Formulas and text are checked using regex expressions. For each requirement that is not met, a piece of feedback is added to the feedback table. The application checks for specific mistakes in order to add specific tips when appropriate. Sometimes this results in a long list of feedback but usually the application only shows the most important feedback in order to manage students’ cognitive workloads.

Finally, the feedback table is added on a separate sheet using the setter-methods. When the student clicks the button again, the old feedback table gets removed and the new one gets added.

Appendix 2: The standardized instructions for the pre-test (translated to English)

Introduction

Welcome back everyone! We will skip the small talk about the holiday and save it for next lesson, because today we need all the time we have. We are going to do a Spreadsheets exam!

In front of me I've got an actual exam that was used during the exam week last year. We will be doing it today while treating it like an actual exam during the exam week. This means that we will do it in silence and independently. I will be checking all your work and you will receive the grades next week. Cheating is not allowed, but also doesn't make any sense, since the grade will not be officially registered. Its function is to demonstrate to you and me how well you already understand Google Spreadsheets, which will be beneficial for the lessons to come. It is normal to not understand most parts of the exam. After all, we haven't had any classes yet. However, don't give up. When learning digital tools it is often possible to figure it out on the go. Note the following two things:

- When hovering over the buttons, Google will show you the name of the button, which you may be able to link to the exercises in the exam.
- When you mess something up, you can use the 'undo button' [*show undo button*] to go back and try again.

Read the exercises carefully and good luck everyone!

The whiteboard:

[Classname], welcome back!

Today: Spreadsheets exam

- We adhere to exam week rules (silence / no cheating / etc.)
- Try to get as high as a grade as possible
- Don't give up! Figure things out

good luck :)

Appendix 3: the questionnaire about students' experiences

Note: this questionnaire has been translated to English for this paper.

Introduction
Thank you for making the effort to fill in this questionnaire! I would like to ask you a few questions about the 'check homework' button and the feedback.

Your responses will be fully anonymous to the outside world. However, they will not be anonymous to me. It's important to know that you may say absolutely anything. Preferably in the same way that you would talk about it behind my back. There is no need to be polite. Your honest answers help move the scientists forward. There will be no consequences for you personally and my opinion of you will not change.

Questions:

- What is your full name?

Students' name

- What class are you in?

Students' class

- How often do you use the 'check homework' button?

Never O - O - O - O - O All the time

- Could you explain why you have used it as much or little as stated above?

Students' explanation

- Could you explain how you use it when doing the assignments?

Students' explanation

- Do you feel like the button/feedback influences how much you learn about Google Spreadsheets? Please elaborate.

Students' explanation

- Do you feel like the button/feedback influences you to do more or less of the exercises? Please elaborate.

Students' explanation

- Do you feel like the practical exercises impacted how much you have learned about Google Spreadsheets? Please elaborate.

Students' explanation

- What are your general opinions about the button/feedback? For example, do you like or dislike it? You can use this question to share anything that you have not been able to share yet.

Students' explanation

Wrapping up

Thank you for participating! If you have any more questions, or comments you forgot to share, you may always email me or share them with me later. I'll see you next week!

Appendix 4: Checking assumptions for linear regression

Since the mediation analysis relies on linear regression, the data must meet specific assumptions to ensure valid inferences. This appendix demonstrates how the data have been tested. Note that the covariates ‘first quarter score’ and ‘level of schooling’ have also been taken into account to justify the regression analysis in appendix 5 where these are included. Hayes (2022) describes the following assumptions:

- Linearity
- Normal distribution of residuals
- Homoscedasticity
- Independence (i.e. no autocorrelation)

Additionally, this appendix checks for multicollinearity.

Linearity

This section investigates the collected data to assess the extent to which linear regression is justified compared to non-linear regression.

Formative assessment in relation to other variables

For the simple linear regressions involving the dichotomous independent variables of formative assessment and type of schooling and a single outcome variable, a linear relation is the only option (see Figure 12 for example).

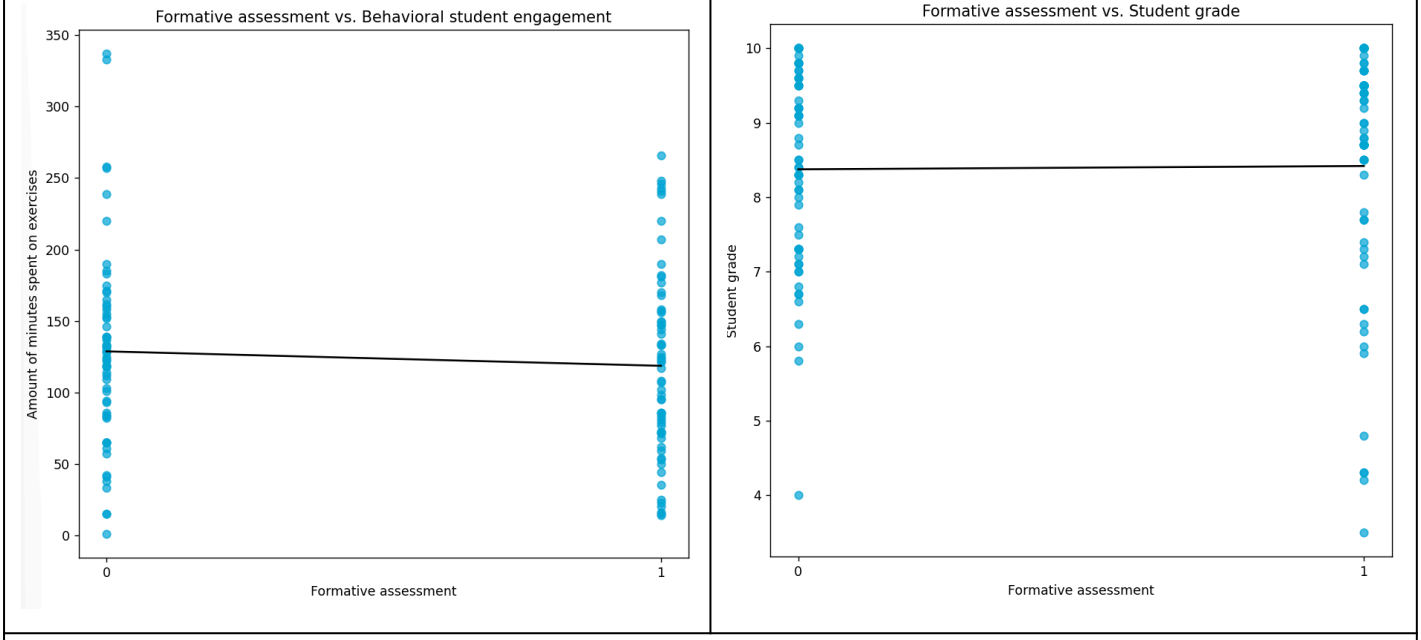

Figure 12: Scatterplots of amount of minutes spent on exercises and student grades in relation to the research group (control group = 0, experimental group = 1)

Behavioral student engagement in relation to student grades

In contrast, the relations involving the behavioral student engagement, the students first quarter, pre-test and final test scores are less straightforward.

Figure 13 shows the scatterplot of the multiple regression analysis. At first sight there seems to be a clear non-linear relation present in the data. Further inspection reveals that this is mostly due to a few extreme cases where the amount of time spent on exercises is particularly low.

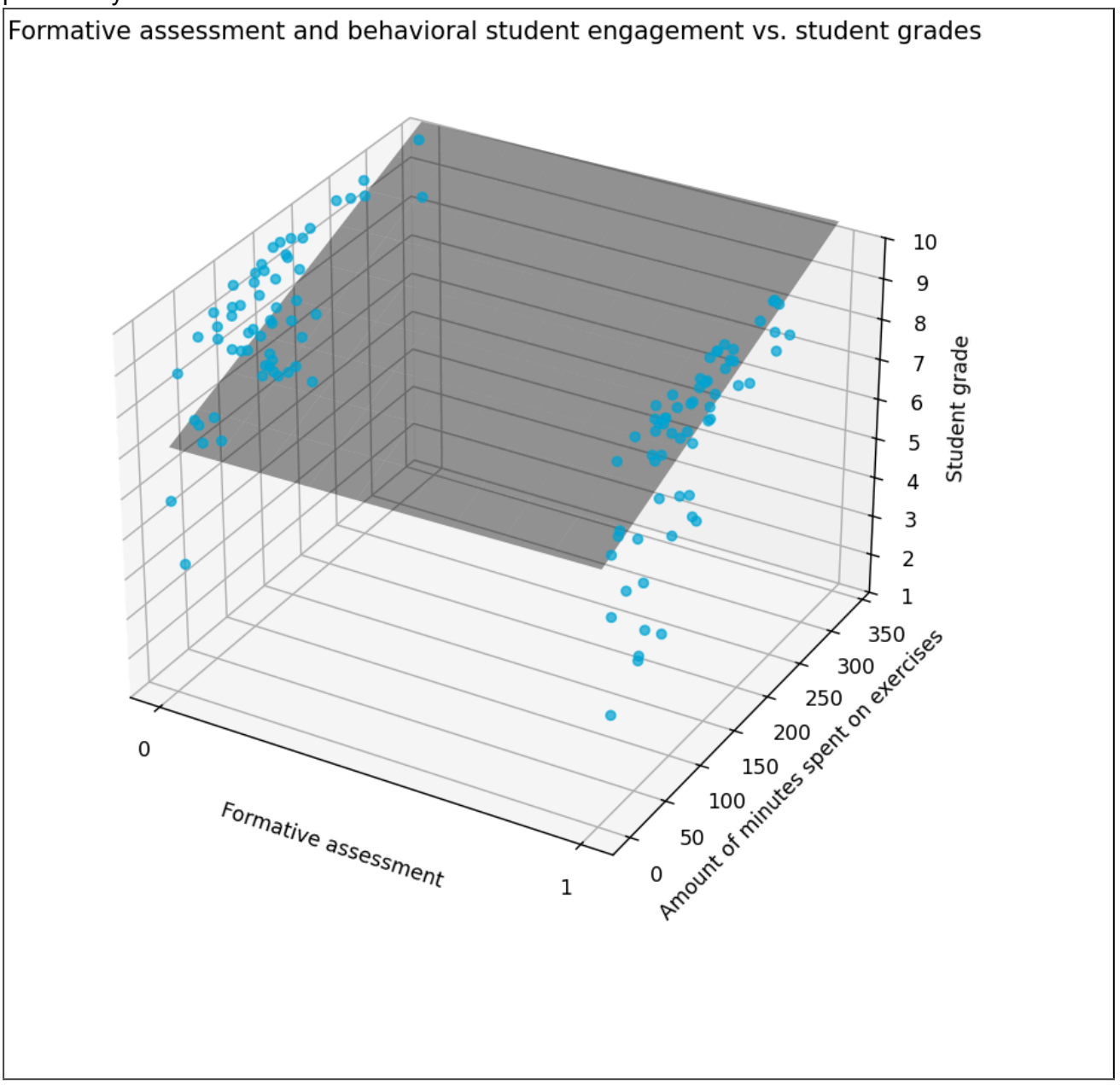

Figure 13: A scatterplot of student grades in relation to the amount of minutes spend on the exercises and the research group (control group = 0, experimental group = 1)

To get a better idea of the type of relationship, eleven different functions were fitted to the behavioral student engagement data for both the control group (see Figure 14) and the experimental group (see Figure 15).

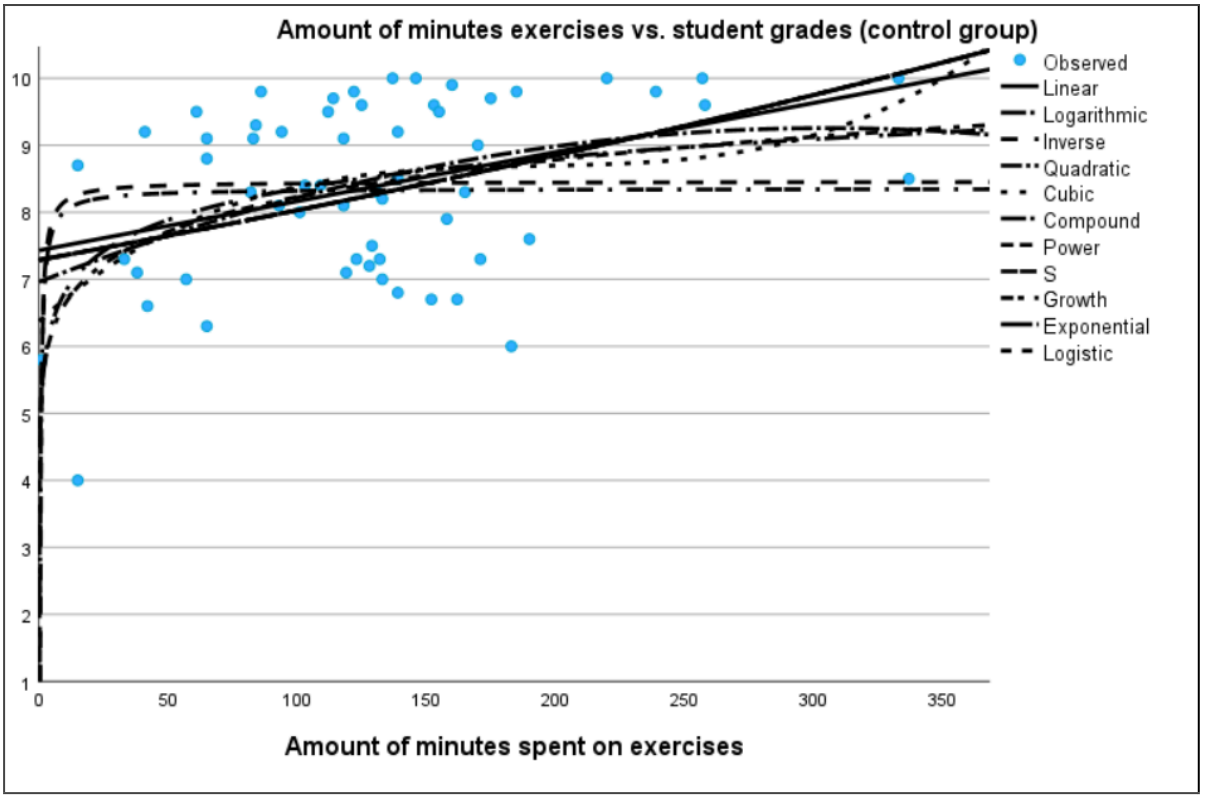

Figure 14: A scatterplot illustrating student grades in relation to the amount of time spent on exercises, with eleven different functions fitted to the data of the control group.

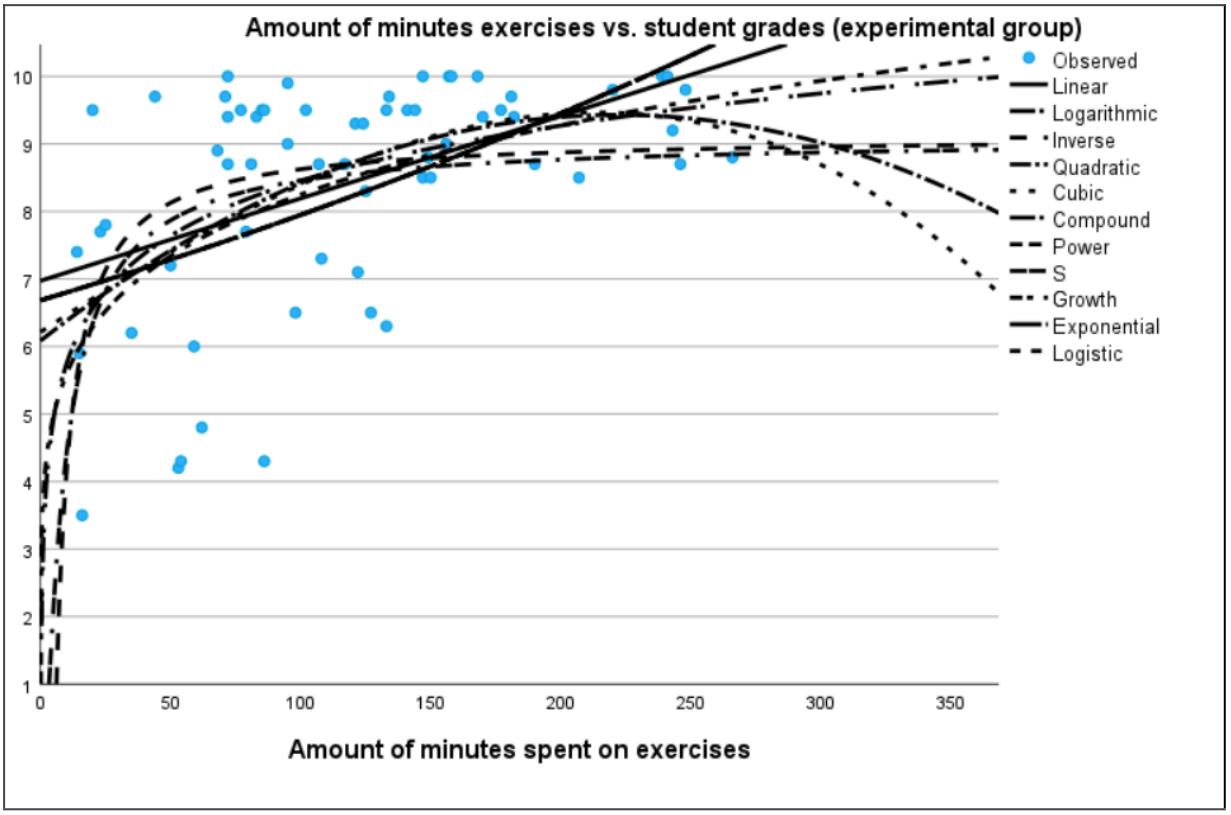

Figure 15: A scatterplot illustrating student grades in relation to the amount of time spent on exercises, with eleven different functions fitted to the data of the experimental group.

Table 3 quantifies how well these functions fit, using the coefficient of determination (R^2) and the F statistic. Roughly speaking, R^2 explains the percentage of variation that is explained by the fitted curve. The F statistic indicates whether the tested model is a better model than one without any explanatory variables.

Amount of minutes exercises (control group)					Amount of minutes exercises (experimental group)		
Equation	***R^2***	***F(df1, df2)***	**p value**		***R^2***	***F(df1, df2)***	**p value**
Power	.216	15.668(1, 57)	< .001		.247	20.046(1, 61)	< .001
Logarithmic	.195	13.815(1, 57)	< .001		.255	20.910(1, 61)	< .001
Cubic	.183	4.120(3, 55)	.010		.266	7.124(3, 59)	< .001
Quadratic	.162	5.406(2, 56)	.007		.265	10.843(2, 60)	< .001
Linear	.142	9.438(1, 57)	.003		.227	17.910(1, 61)	< .001
Compound	.141	9,367(1, 57)	.003		.215	16.731(1, 61)	< .001
Growth	.141	9,367(1, 57)	.003		.215	16.731(1, 61)	< .001
Exponential	.141	9,367(1, 57)	.003		.215	16.731(1, 61)	< .001
Logistic	.141	9,367(1, 57)	.003		.215	16.731(1, 61)	< .001
S	.091	5,734(1, 57)	.020		.194	14.727(1, 61)	< .001
Inverse	.083	5,190(1, 57)	.026		.197	14.952(1, 61)	< .001

Table 3: The coefficient of determination (R^2) values, F-statistics and corresponding p values, for various functions applied to the behavioral student engagement data

Table 3 shows that, according to the F-statistics, most functions provide a useful model. Furthermore, the table shows that a power function provides the best fit for the data, explaining 6.4% more variance than a linear function in the control group and 2% more in the experimental group. However, these are relatively small differences and since linear regression is more straightforward to interpret and more widely understood, this research deems it reasonable to use linear regression for the mediation analysis.

A similar analysis and conclusion is made for the relations involving the covariates. See figures 16, 17, 18 and 19 and tables 4 and 5.

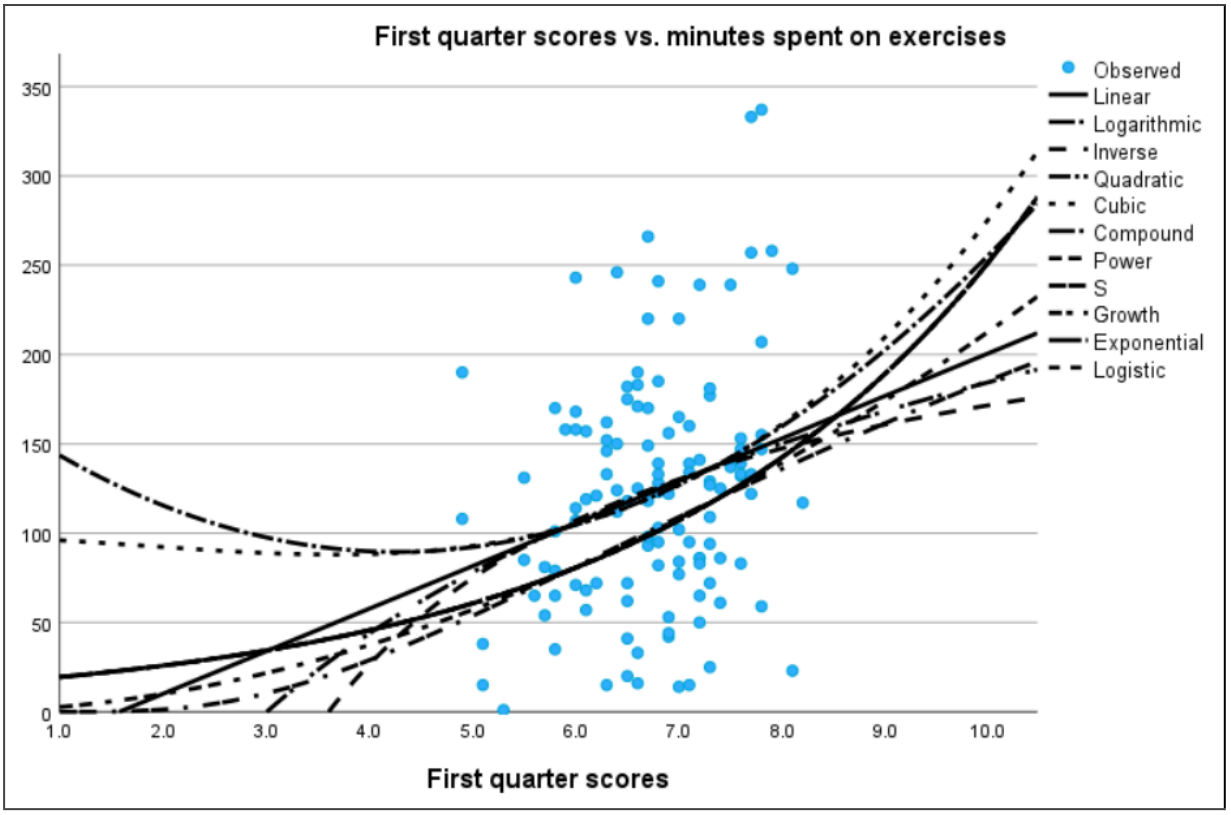

Figure 16: A scatterplot illustrating students' average first quarter scores in relation to the amount of time spent on exercises, with eleven different functions fitted to the data.

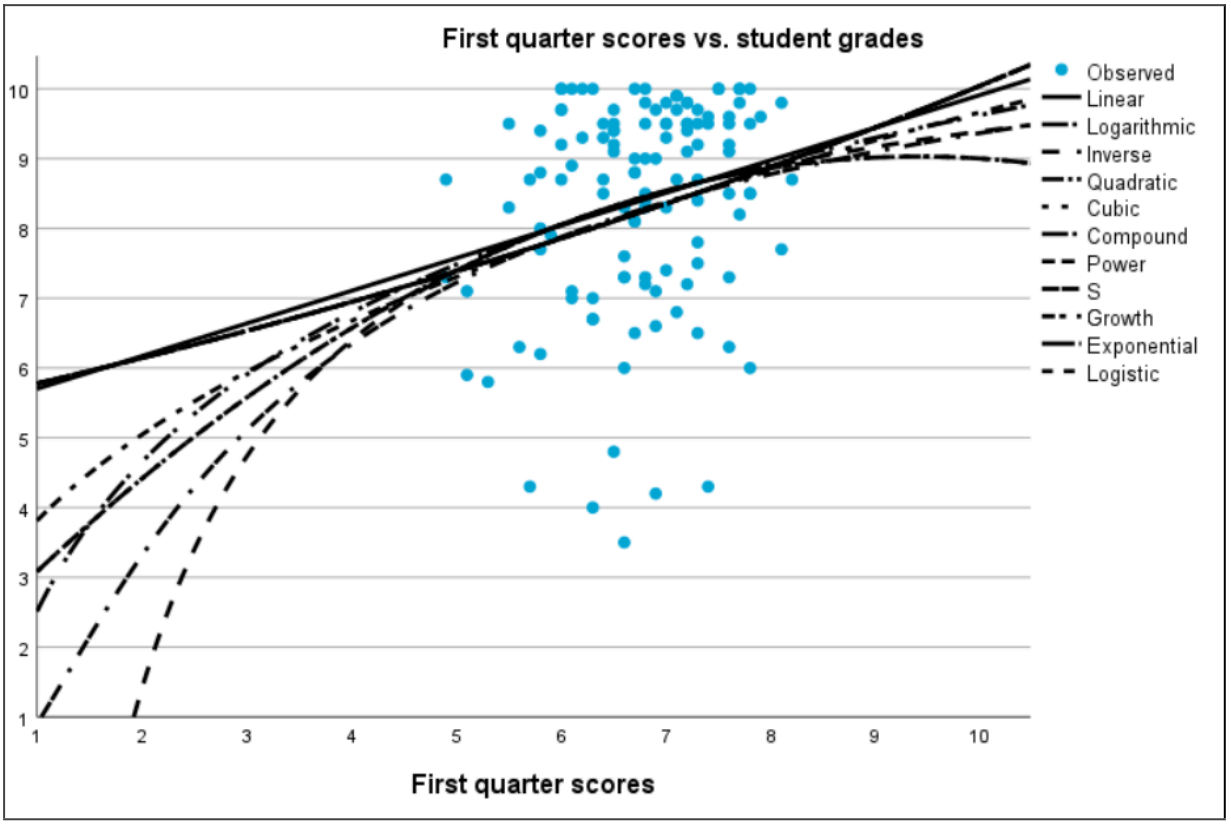

Figure 17: A scatterplot illustrating students' first quarter scores in relation to their final grades, with eleven different functions fitted to the data.

First quarter scores and minutes spent on exercises					First quarter scores and final test scores		
Equation	R^2	***F(df1, df2)***	**p value**		R^2	***F(df1, df2)***	**p value**
Linear	.066	8.489(1, 120)	.004		.050	6.282(1, 120)	.014
Logarithmic	.064	8.222(1, 120)	.005		.051	6.410(1, 120)	.013
Inverse	.062	7.882(1, 120)	.006		.051	6.479(1, 120)	.012
Quadratic	.069	4.376(2, 119)	.015		.051	3.202(2, 119)	.044
Cubic	.069	4.402(2, 119)	.014		.051	3.202(2, 119)	.044
Compound	.067	8.554(1, 120)	.004		.043	5.407(1, 120)	.022
Power	.069	8.929(1, 120)	.003		.044	5.464(1, 120)	.021
S	.071	9.229(1, 120)	.003		.044	5.465(1, 120)	.021
Growth	.067	8.554(1, 120)	.004		.043	5.407(1, 120)	.022
Exponential	.067	8.554(1, 120)	.004		.043	5.407(1, 120)	.022
Logistic	.067	8.554(1, 120)	.004		.043	5.407(1, 120)	.022

Table 4: The coefficient of determination (R^2) values, F-statistics and corresponding p values, for various functions applied to the data of students' first quarter scores

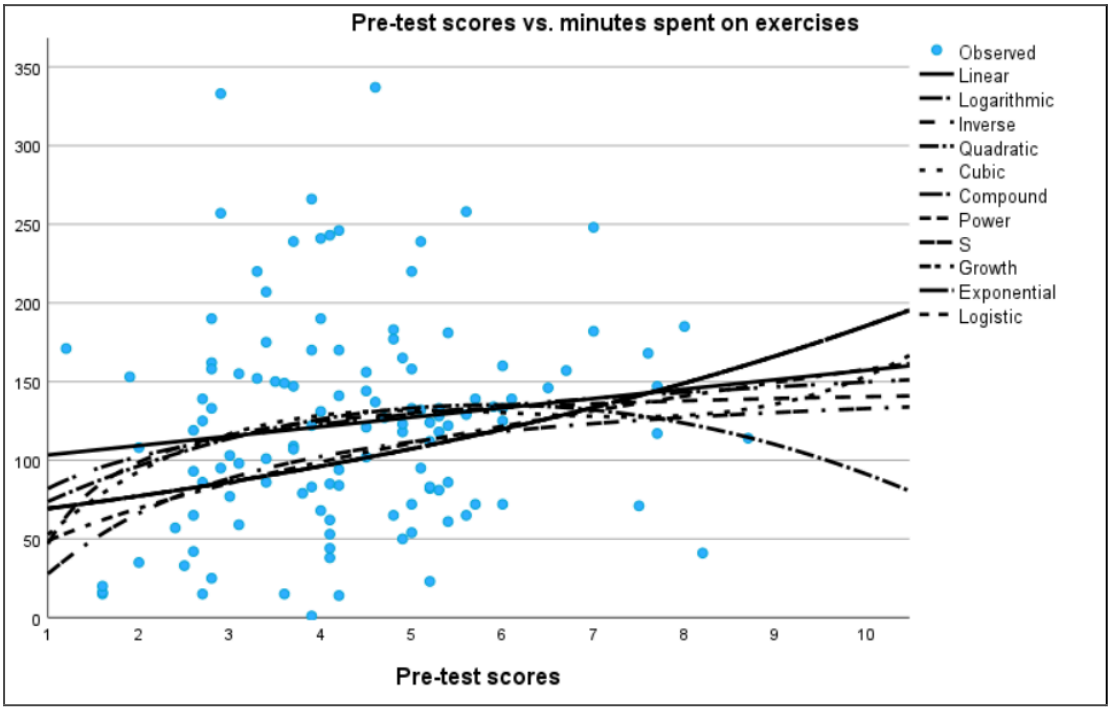

Figure 18: A scatterplot illustrating students' pre-test scores in relation to the amount of time spent on exercises, with eleven different functions fitted to the data.

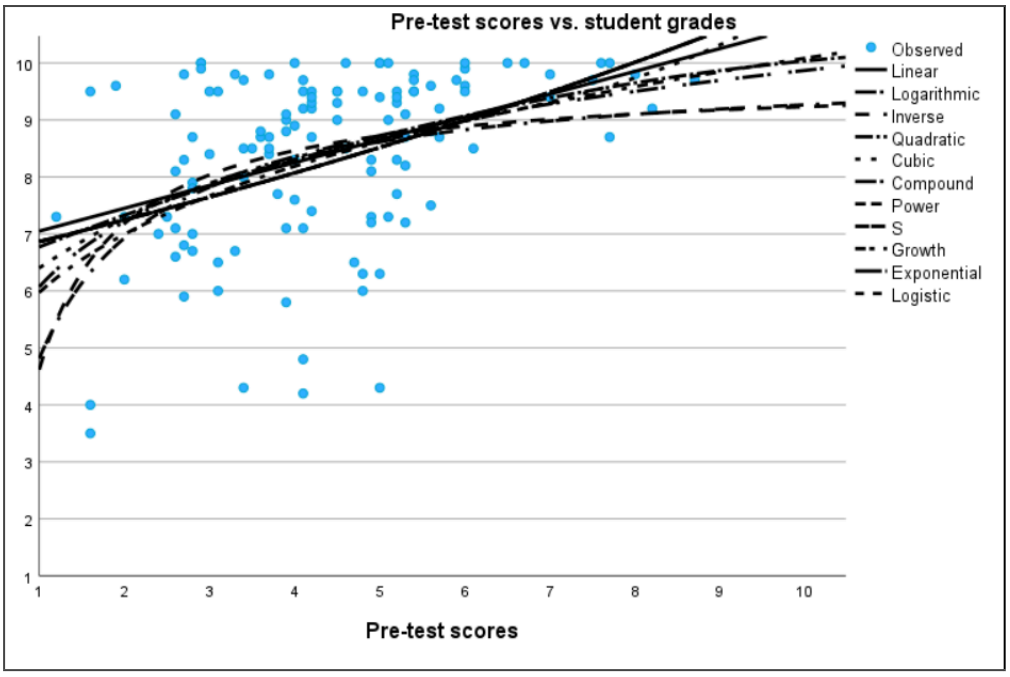

Figure 19: A scatterplot illustrating students' pre-test scores in relation to their final grades, with eleven different functions fitted to the data.

Pre-test scores and Minutes spent on exercises					Pre-test scores and final test scores		
Equation	**R^2**	***F(df1, df2)***	**p value**		**R^2**	***F(df1, df2)***	**p value**
Linear	.018	2.232(1, 120)	.138		.160	22.853(1, 120)	< .001
Logarithmic	.027	3.301(1, 120)	.072		.164	26.580(1, 120)	< .001
Inverse	.031	3.894(1, 120)	.051		.149	21.046(1, 120)	< .001
Quadratic	.033	2.037(2, 119)	.135		.163	11.547(2, 119)	< .001
Cubic	.036	1.487(3, 118)	.222		.165	7.746(3, 118)	< .001
Compound	.043	5.416(1, 120)	.022		.145	20.404(1, 120)	< .001
Power	.057	7.205(1, 120)	.008		.157	22.289(1, 120)	< .001
S	.062	7.944(1, 120)	.006		.151	21.296(1, 120)	< .001
Growth	.043	5.416(1, 120)	.022		.145	20.404(1, 120)	< .001
Exponential	.043	5.416(1, 120)	.022		.145	20.404(1, 120)	< .001
Logistic	.043	5.416(1, 120)	.022		.145	20.404(1, 120)	< .001

Table 5: The coefficient of determination (R^2) values for various functions applied to the data of students' pre-test scores

Normal distribution of residuals and Homoscedasticity

Table 6 and 7 show that the White tests do not provide evidence for heteroscedasticity, implying that the assumption for homoscedasticity is met. The Shapiro-Wilk tests imply that the residuals are not normally distributed. Because of the latter, a bootstrapping technique will be employed which does not rely on normality nor on homoscedasticity.

Regression of:	Shapiro-wilk statistic	p value	White test statistic	p value
IV: formative Assessment **Cov:** pre-test score **DV:** minutes spent on exercises	*df*: 122 *W*: 0.81	.002	*df*: 4 X^2: 1.741	.783
IV: formative Assessment **IV:** minutes spent on exercises **Cov:** pre-test score **DP:** Student grade	*df*: 122 *W*: .508	<.001	*df*: 119 X^2: 119.913	.459

Table 6: The results of the shapiro-wilk test and the White test on the regressions of the mediation analysis

Regression of:	Shapiro-wilk statistic	p value	White test statistic	p value
IV: formative Assessment **Cov:** pre-test score **Cov:** first quarter score **Cov:** type of schooling **DV:** minutes spent on exercises	*df*: 122 *W*: .071	.008	*df*: 12 X^2: 16.111	.186
IV: formative Assessment **IV:** minutes spent on exercises **Cov:** pre-test score **Cov:** first quarter score **Cov:** type of schooling **DP:** Student grade	*df*: 122 *W*: .547	< .001	*df*: 121 X^2: 122.000	.457

Table 7: The results of the shapiro-wilk test and the White test on the regressions of the mediation analysis, including the covariates 'first quarter score' and 'type of schooling'

Independence

Independence (also known as autocorrelation) indicates that data points do not depend on each other. Table 8 and 9 show the results of the Durban-Watson tests with the corresponding critical values according to the University of Notre Dame (z.d.). Both statistics are less than two but higher than the upper critical value, indicating independence of the data points.

Regression of:	Durbin-Watson	Lower critical value	Upper critical value
IV: formative Assessment **Cov:** pre-test score **DV:** minutes spent on exercises	1.838	1.502	1.582
IV: formative Assessment **IV:** minutes spent on exercises **Cov:** pre-test score **DP:** Student grade	1.767	1.482	1.604

Table 8: Results of the Durban-Watson test on the regressions of the mediation analysis and the critical values for a significance level of 0.01

Regression of:	Durbin-Watson	Lower critical value	Upper critical value
IV: formative Assessment **Cov:** pre-test score **Cov:** first quarter score **Cov:** type of schooling **DV:** minutes spent on exercises	1.857	1.461	1.625
IV: formative Assessment **IV:** minutes spent on exercises **Cov:** pre-test score **Cov:** first quarter score **Cov:** type of schooling **DP:** Student grade	1.759	1.441	1.647

Table 9: Results of the Durban-Watson test on the regressions of the mediation analysis and the critical values for a significance level of 0.01, including the covariates 'first quarter score' and 'type of schooling'

Multicollinearity

Multicollinearity occurs when the independent variables of a multiple linear regression correlate with each other. High correlation between independent variables poses difficulty when predicting one variable while controlling for the other. Table 10 demonstrates some significant correlations, but none of them are high (rule of thumb: lower than -0.7 or higher than 0.7). This research considered the multicollinearity sufficiently low to conduct the linear regressions.

Correlations	PRT	FQS	TOS	MSOE	FA
Pre-test scores (PRT)	-	$r_s(120) = .154$, $p = .089$	$r_s(120) = .331$, $p < .001$	$r_s(120) = .150$, $p = .098$	$r_s(120) = .070$, $p = .441$
First quarter scores (FQS)	-	-	$r_s(120) = .446$, $p < .001$	$r_s(120) = .203$, $p = .025$	$r_s(120) = -.053$, $p = .563$

Type of schooling (TOS)	-	-	-	$r_s(120) = .203$, $p = 0.25$	$r_s(120) = .076$, $p = .408$
Minutes spent on exercises (MSOE)	-	-	-	-	$r_s(120) = .-074$, $p = .416$
Formative Assessment (FA)	-	-	-	-	-

Table 10: Spearman's Rho correlations between all variables that act as independent variables together at least once in the mediation analysis.

Appendix 5: Supplementary results

This appendix presents results from the data collection and analysis that are insightful but not directly relevant to answering the research questions.

Descriptive statistics

The table and figures in this section present descriptive statistics of the quantitative data collected in this research. These data are not required to answer the research questions but may provide intuition about the data.

Assignment group	Schooling type	Amount of minutes spent		Final grades	
		Mean	Std. deviation	Mean	Std. deviation
Control group	Havo	89.6	60.6	7.4	1.3
	Vwo	144.5	65.5	8.8	1.1
	Total	**128.7**	**68.4**	**8.4**	**1.3**
Experimental group	Havo	116.9	77.5	8.0	1.7
	Vwo	119	62.2	8.5	1.7
	Total	**118.6**	**65.3**	**8.4**	**1.7**
Total	Havo	102.0	68.9	7.7	1.5
	Vwo	130.8	64.7	8.6	1.4
	Total	**123.5**	**66.7**	**8.4**	**1.5**

Table 11: Means and standard deviations of the quantitative research data by treatment groups and type of schooling

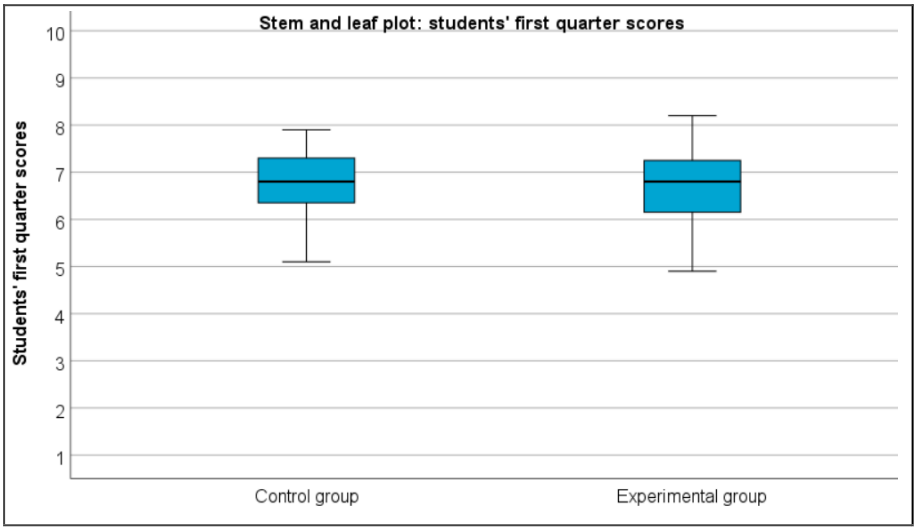

Figure 20: A stem and leaf plot of students' first quarter scores, comparing the control and experimental group

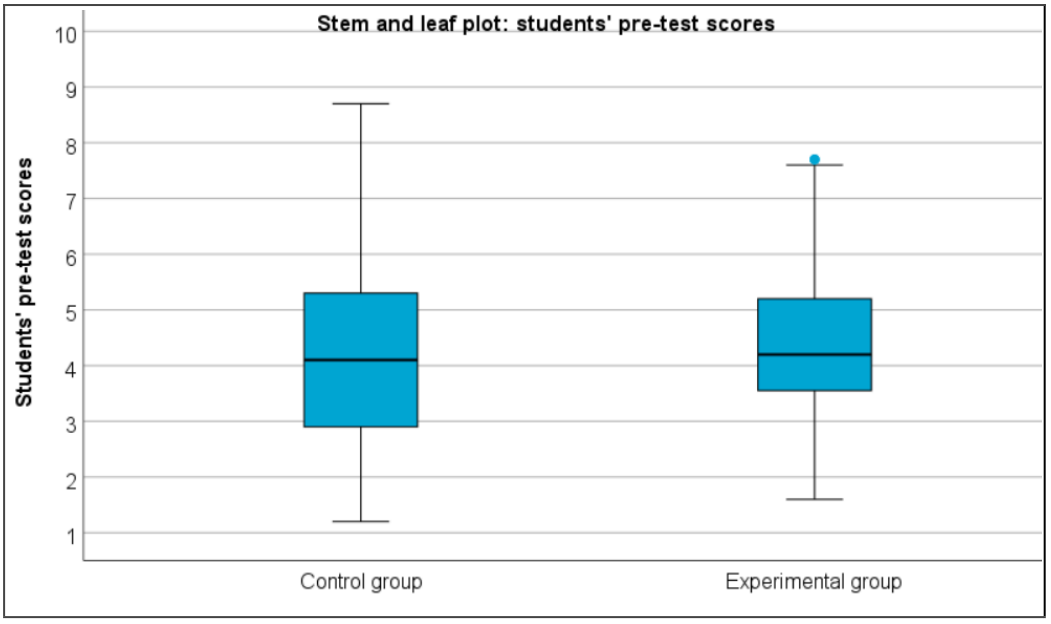

Figure 21: A stem and leaf plot of students' pre-test scores, comparing the control and experimental group

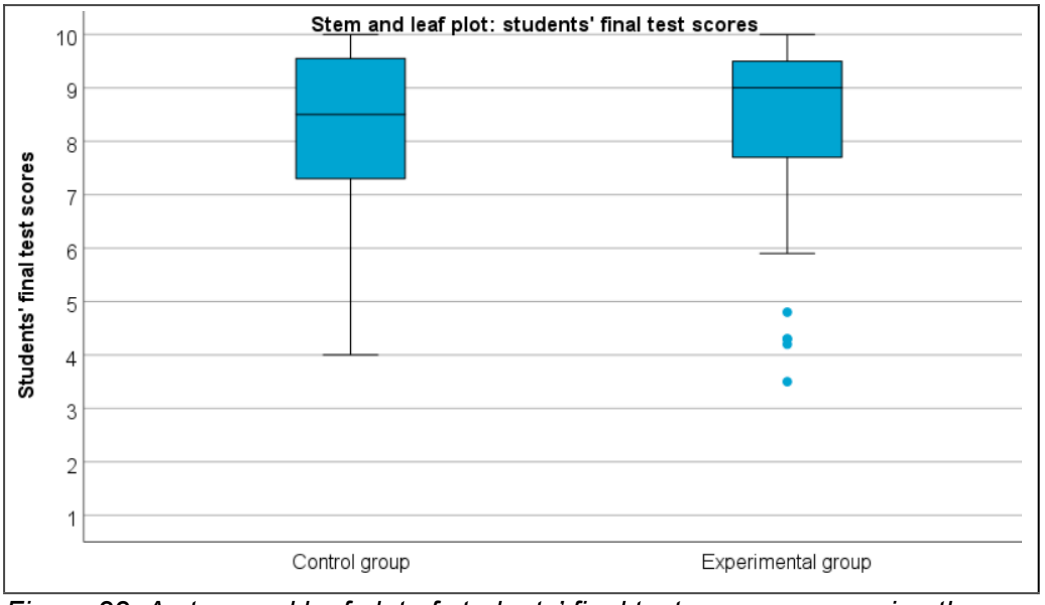

Figure 22: A stem and leaf plot of students' final test scores, comparing the control and experimental group

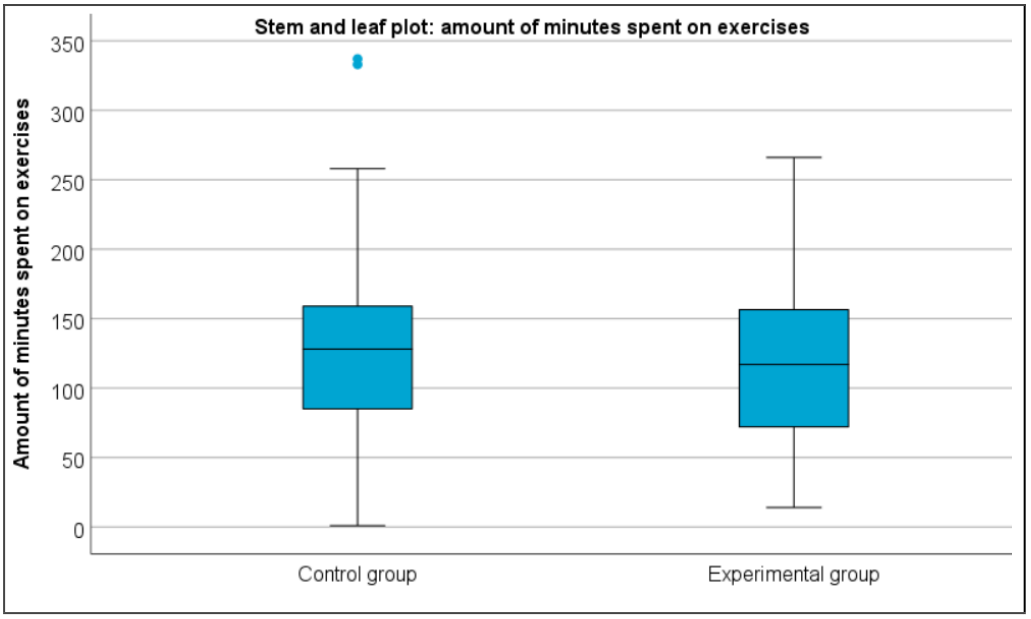

Figure 23: A stem and leaf plot of the amount of minutes spent on exercises, comparing the control and experimental group

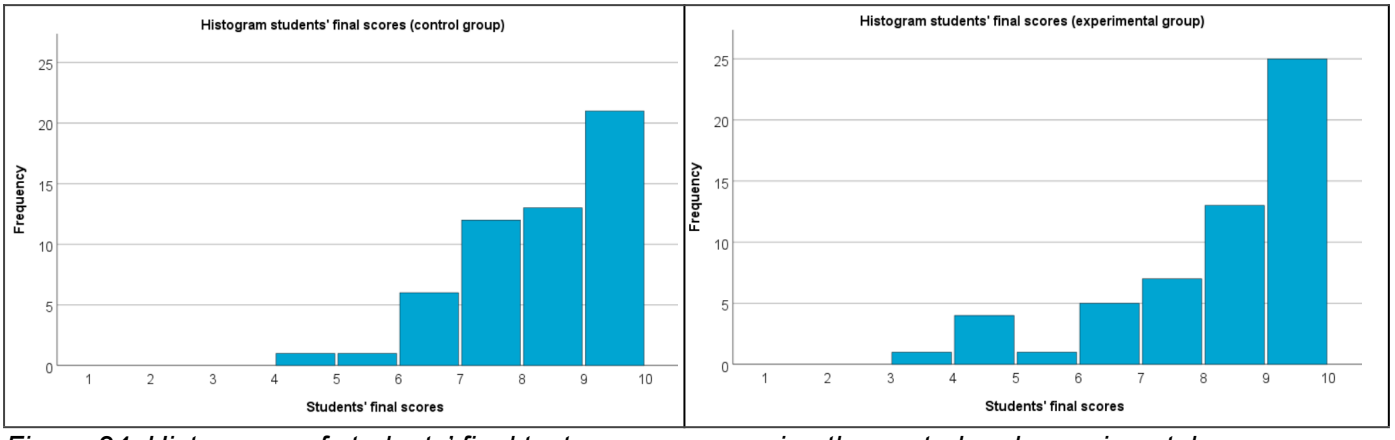

Figure 24: Histograms of students' final test scores, comparing the control and experimental group

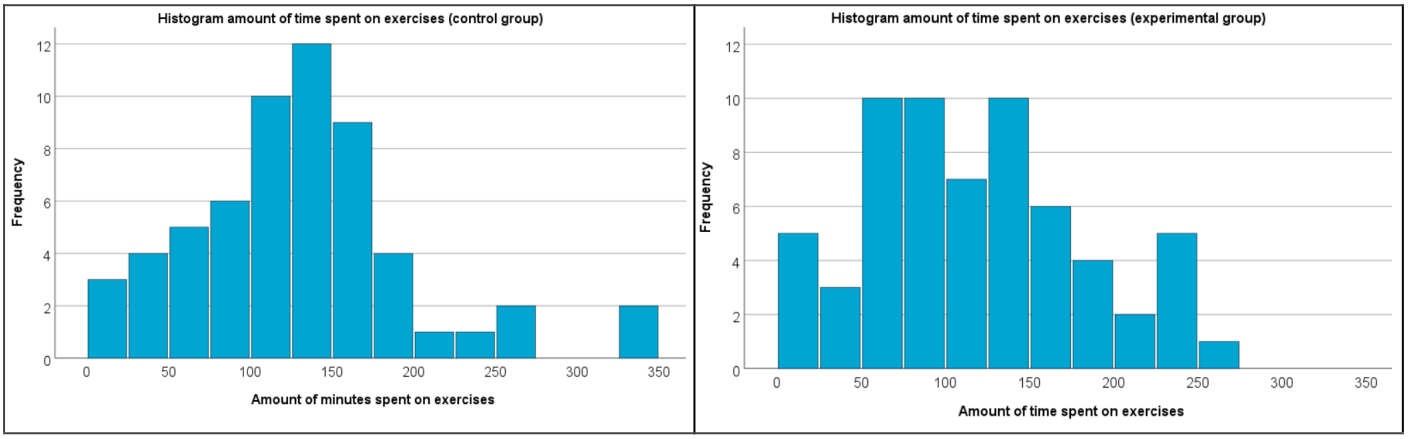

Figure 25: Histogram of the amount of minutes spent on exercise, comparing the control and experimental group

Supplementary results from the mediation analysis

The following tables go into more detail about the results of the mediation analysis:

- Table 12 provides more details on the regression analysis of the total effect.
- Table 13 provides details on the regression analysis with behavioral student engagement as the outcome variable.
- Table 14 provides details on the regression analysis with learning outcomes as the outcome variable.

Table 15 to 18 show similar results, but for the mediates analysis that included the covariates 'first quarter scores' and 'type of schooling'. This mediation analysis has been conducted as well since these data had already been collected for the assignment to treatment groups, allowing them to provide an extra layer of control this way. However, no additional inferences arose from the analysis.

The standardized coefficients resulting from linear regressions with a dichotomous independent variable are expressed as partially standardized coefficients and are denoted with the subscript: $_{ps}$.

The mediation analysis that excludes the covariates 'first quarter scores' and 'level of schooling'

Table 12 indicates the 'pre-test score' to be a significant predictor of the final test score, indicating that a point increase in the pre-test predicts a 0.401 point increase in the final test score. Considering that the first quartile of the pre-test data is 3.3 and the third quartile is 5.2, the coefficient should be used with caution for data outside of these values.

Total effect with outcome variable: learning outcomes				
Variable	**Coefficient (*B*)**	**Standardized coefficient (β)**	**LLCI**	**ULCI**
Formative assessment (c)	-0.033	-0.022_{ps}	-0.534	0.469
Pre-test scores (prt_c)	0.401	0.401	0.234	0.568

Table 12: The results of the analysis of the total effect including the covariates

Table 13 indicates no significant predictors of the amount of minutes spent on exercises.

Outcome variable: behavioral student engagement				
Relation	**Coefficient (*B*)**	**Standardized coefficient (β)**	**LLCI**	**ULCI**
Intercept	-102.147	-	64.166	140.128
Formative assessment (a)	-11.322	-0.170_{ps}	-35.154	12.173
Pre-test scores (prt_a)	6.223	0.141	-1.388	13.742

Table 13: Results of the regression analysis of computer assisted formative assessment on the amount of minutes spent on exercises with pre-test scores, first quarter scores and type of schooling (havo/vwo) as covariates.

Table 14 indicates the amount of minutes spent on exercises to be a significant predictor of the final test score, indicating that a minute increase in the amount of minutes spent predicts a 0.009 point increase in the final test score. Furthermore it shows that the pre-test score remains a significant predictor of the final test score when additionally controlling for the amount of minutes spent on exercises.

Outcome variable: learning outcomes				
Variable	**Coefficient (*B*)**	**Standardized coefficient (β)**	**LLCI**	**ULCI**
Intercept	5.781	-	4.967	6.594
Formative assessment (c')	0.065	0.043_{ps}	-0.376	0.503
Minutes spent on exercises (b)	0.009	0.381	0.005	0.012
Pre-test scores (prt_bc')	0.348	0.347	0.211	0.482

Table 14: Results of the regression analysis of computer assisted formative assessment and minutes spent on exercises on student final test scores with pre-test scores, first quarter scores and type of schooling as covariates.

The mediation analysis that includes the covariates 'first quarter scores' and 'level of schooling'

This section presents the results of a mediation analysis where 'first quarter scores' and 'level of schooling' are additionally taken into account as covariates. The three regression analyses were all statistically significant:

- For the total effect: R^2 = .443, $F(4, 117) = 7.146$, $p < .001$.
- For behavioral student engagement as the outcome variable: R^2 = .083, $F(4, 117) = 2.659$, $p = .036$.
- For learning outcomes as the outcome variable: R^2 = .313, $F(5, 116) = 10.591$, $p < .001$

Table 15 presents the results regarding the direct, indirect and total effect and relation *a*. The coefficients should be interpreted as the change in the dependent variable when the computer assisted formative assessment is applied compared to when it is not.

Total, direct and indirect effects and relation *a*				
Type of effect	**Coefficient (*B*)**	**Partially standardized coefficient (β)**	**LLCI**	**ULCI**
Relation *a*	-9.927	-0.149	-33.479	12.673
Direct effect (c')	0.052	0.035	-0.393	0.490
Indirect effect (ab)	-0.080	-0.053	-0.287	0.102
Total effect (c)	-0.028	-0.019	-0.527	0.471

Table 15: The direct, indirect and total effect resulting from the mediation analysis

Table 15 shows that neither the total effect, direct effect, nor indirect effect is significant on a 95% bootstrap confidence interval, providing insufficient evidence to reject the null hypotheses H_{10} and H_{30}. The same goes for the relation between computer assisted formative assessment and behavioral student engagement (relation *a*), providing insufficient evidence to reject the null hypothesis H_{20}.

Total effect with outcome variable: learning outcomes				
Variable	**Coefficient (*B*)**	**Standardized coefficient (β)**	**LLCI**	**ULCI**
Formative assessment (c)	-0.028	-0.019_{ps}	-0.527	0.471
Type of schooling (tos_c) (Havo=0, Vwo=1)	0.394	0.144	-0.288	1.076
Pre-test scores (prt_c)	0.347	0.346	0.172	0.521
First quarter scores (fqs_c)	0.243	0.116	-0.153	0.639

Table 16: The results of the analysis of the total effect including the covariates

Table 16 indicates the 'pre-test score' to be a significant predictor of the final test score, indicating that a point increase in the pre-test predicts a 0.347 point increase in the final test score. Considering that the first quartile of the pre-test data is 3.3 and the third quartile is 5.2, the coefficient should be used with caution for data outside of these values.

Outcome variable: behavioral student engagement				
Relation	**Coefficient (*B*)**	**Standardized coefficient (β)**	**LLCI**	**ULCI**
Intercept	-26.268	-	-155.842	101.734
Formative assessment (a)	-9.927	-0.149_{ps}	-33.479	12.673
type of schooling (tos_a) (Havo=0, Vwo=1)	9.889	0.065_{ps}	-24.545	42.893
Pre-test scores (prt_a)	3.881	0.088	-3.918	11.527
First quarter scores (fqs_a)	19.314	0.208	-0.871	39.937

Table 17: Results of the regression analysis of computer assisted formative assessment on the amount of minutes spent on exercises with pre-test scores, first quarter scores and type of schooling (havo/vwo) as covariates.

Outcome variable: learning outcomes				
Variable	**Coefficient (*B*)**	**Standardized coefficient (β)**	**LLCI**	**ULCI**
Intercept	5.173	-	2.852	7.379
Formative assessment (c')	0.052	0.035_{ps}	-0.393	0.490

Minutes spent on exercises (b)	0.008	0.357	0.005	0.012
Type of schooling (tos_bc') (Havo=0, Vwo=1)	0.314	0.091_{ps}	-0.346	1.020
Pre-test scores (prt_bc')	0.315	0.315[3]	0.161	0.473
First quarter scores (fqs_bc')	0.087	0.042	-0.268	0.462

Table 18: Results of the regression analysis of computer assisted formative assessment and minutes spent on exercises on student final test scores with pre-test scores, first quarter scores and type of schooling as covariates.

Table 18 indicates the amount of minutes spent on exercises to be a significant predictor of the final test score, indicating that a minute increase in the amount of minutes spent predicts a 0.008 point increase in the final test score. Furthermore it shows that the pre-test score remains a significant predictor of the final test score when additionally controlling for the amount of minutes spent on exercises.

[3] The standardized coefficient being equal to the unstandardized coefficient is not an error but occurs because the standard deviations of the pre-test scores and final scores are nearly equal.