Assessment for growth: fostering student learning through assessment innovations in medical education

Wijk, E.V. van

Part III: Feedback

Chapter 8

Does ‘summative’ count? The influence of the awarding of study credits on feedback use and test-taking motivation in medical progress testing

Elise V. van Wijk
Floris M. van Blankenstein
Jeroen Donkers
Roemer J. Janse
Jacqueline Bustraan
Liesbeth G.M. Adelmeijer
Eline A. Dubois
Friedo W. Dekker
Alexandra M.J. Langers

Advances in Health Science Education. 2024;29(5):1665-1688.
DOI: 10.1007/s10559-10324-4.

Abstract

Despite the increasing implementation of formative assessment in medical education, its' effect on learning behaviour remains questionable. This effect may depend on how students value formative, and summative assessments differently. Informed by Expectancy Value Theory, we compared test preparation, feedback use, and test-taking motivation of medical students who either took a purely formative progress test (*formative* PT-group) or a progress test that yielded study credits (*summative* PT-group). In a mixed-methods study design, we triangulated quantitative questionnaire data (n = 264), logging data of an online PT feedback system (n = 618), and qualitative interview data (n = 21) to compare feedback use, and test-taking motivation between the *formative* PT-group (n = 316), and the *summative* PT-group (n = 302). Self-reported, and actual feedback consultation was higher in the *summative* PT-group. Test preparation, and active feedback use were relatively low and similar in both groups. Both quantitative, and qualitative results showed that the motivation to prepare and consult feedback relates to how students value the assessment. In the interview data, a link could be made with goal orientation theory, as performance-oriented students perceived the *formative* PT as not important due to the lack of study credits. This led to low test-taking effort, and feedback consultation after the *formative* PT. In contrast, learning-oriented students valued the formative PT, and used it for self-study or self-assessment to gain feedback. Our results indicate that most students are less motivated to put effort in the test, and use feedback when there are no direct consequences. A supportive assessment environment that emphasizes recognition of the value of formative testing is required to motivate students to use feedback for learning.

Introduction

The notion that *'assessment drives learning'* is widely acknowledged [1, 2]. The way learning is driven may therefore differ with the focus of the assessment. Within medical education, the focus is mainly on assessment of learning [3]. These summative assessments evaluate students' learning by measuring their performance, often reported as a summative grade. In contrast to assessment *of* learning, assessment *for* learning uses formative assessments to enhance in-depth learning, and self-regulated learning (SRL) by providing ongoing feedback [4-13]. More specifically, formative feedback provides opportunities for reflection, identifying learning gaps, and adjusting learning, which are important aspects of SRL [5, 14, 15] In this way, feedback can also stimulate the use of learning strategies that enhance future learning performance [16]. With the growing consensus that assessment should promote learning, and in light of these positive learning effects, there is a shift in assessment *of* learning towards assessments *for* learning in medical education [3, 11]. However, to facilitate this shift, further elucidation of the complex relationship between assessment, learning, and the driving factors behind students' learning is needed.

One of the factors found to drive students' motivation to learn is increasing the weight of summative assessments [17]. Motivation to learn for an assessment also affects test-taking motivation: students' readiness to invest effort in a test [18-21]. Considering the lack of direct consequences of formative test results, students might be less motivated to put their best effort in these tests. This can be explained by the Expectancy-Value Theory (EVT), a conceptual framework frequently used in the context of test-taking motivation. This theory assumes that motivation for a task depends on expectancies of success, and perceived value given to the task [22]. Specifically, motivation for a task increases when people expect to be successful and when they find the task valuable for themselves. Test-taking effort is the main element of test-taking motivation, which, according to the EVT, is thus the direct outcome of expectancy and value. Most studies that investigated EVT in the context of test-taking motivation focused on the value component of EVT. Overall, these studies report positive relationships between value and test performance, and also between test effort and test performance [19, 23].

Another way to look at *'assessment drives learning'* is through the lens of the goal orientation theory. This theory states that the individual goal orientation affects motivation, which in turn guides behavioural responses [24]. Goal orientation can either rely on learning (mastery- or learning-oriented goals) or performance (performance-oriented goals). Learning-oriented students might take a different approach in making a test, and using its feedback than performance-oriented students, but so far the influence of goal orientation in different assessment conditions has not been investigated.

One way to investigate the differences between different assessment conditions is by using the medical progress test (PT), which is a frequently used assessment method in medical education. The PT is a longitudinal, comprehensive, and curriculum-independent test administered repeatedly to assess students' knowledge progress and provide feedback [25, 26]. The PT combines longitudinal testing with feedback, serving an important formative function, but in many educational contexts the results of PTs are also used for a summative pass/fail decision followed by the rewarding of study credits. As the PT covers the entire medical curriculum, it discourages test-directed studying, and encourages self-directed learning by using the feedback of the previous PT [26].

Implementing frequent PTs with a summative component, and the integrating purely formative PTs (no study credits involved) in a curriculum with other formative assessments has shown a positive impact on students' test-effort, perceived learning value, and feedback use [27-31]. However, some studies have not found the expected beneficial impact of feedback in purely formative PTs on learning [26, 32-35]. Different educational conditions affect the test-taking effort, and the perceived value of purely formative PTs [30]. These PTs have no direct consequences (i.e. no *'stakes'*) for study progress, which may lead to a lower perceived value, which in turn may result in less test-taking motivation and effort put in these tests [30, 36]. Besides an impact on test performance, this might also affect their use of feedback.

In summary, while assessment should promote learning (i.e. assessment *for* learning), the actual effect of formative assessments on learning is unclear. More specifically, it remains unclear how formative versus summative assessment affects students' feedback use, and test-taking motivation. The PT provides a unique opportunity to study this distinction, especially when we can compare a purely formative PT with a PT that also has a summative component. Understanding how students adapt their learning behaviour to formative versus summative assessment may help teachers optimize both functions of assessment, as it enables them to react to the student's behaviour in order to promote their learning process, and foster lifelong learning. Therefore, we aimed to investigate the effect of a PT with a summative component (*summative* PT), and a purely formative PT (*formative* PT) on medical students' (1) test preparation, (2) factors that influence test taking motivation, and the use of feedback, and (3) self-reported, and actual feedback use after the test.

Methods

Study design

We used a convergent mixed-methods approach with a subtle realism paradigm, involving a questionnaire, online Progress test Feedback system (ProF) logging data, and semi-structured interviews. The subtle realism paradigm combines a realist ontology (an objective reality independent of our perceptions) with a constructivist epistemology (our understanding of reality depends on our perspectives) [37, 38]. This paradigm aims at representing reality rather than attaining *"the truth"*, by triangulating different data sources, perspectives, and theories. We chose this approach as this best aligns with our research design, which attempts to represent, and deepen our understanding of reality ('feedback use in the context of different assessment conditions') by the triangulating different data sources, and theories. This paradigm allows us to integrate different perspectives while remaining flexible in interpreting our qualitative data. All data types were analysed separately and converged in a final interpretation phase, where we compared the results of the quantitative and qualitative data, and assessed whether the data confirmed or disconfirmed each other. Our qualitative results, using the existing theoretical frameworks of EVT and goal-orientation theory, helped us understand, and explain the observed and self-reported quantitative feedback behaviour.

Setting

The study was conducted at Leiden University Medical Center (LUMC) in the Netherlands. The medical curriculum in the Netherlands includes a three year preclinical Bachelor program and three year clinical Master program. The Bachelor program comprises several theoretical courses, assessed by written

summative assessments at the end of each course, and rewarded with study credits. Most courses also offer a formative assessment for practice, which is not mandatory for students to take. In the Master program, students undergo clinical rotations, assessed by a pass or fail decision based on supervisor feedback. Throughout their six years of medical school, all medical students take four PTs per year, resulting in a total of 24 test moments [25]. The PTs are taken in September (PT 1), December (PT 2), February (PT 3), and May (PT 4). The PT is a comprehensive written test of 200 multiple choice questions (MCQs), covering all relevant medical disciplines, and stratified in categories [26]. The MCQs include a question mark option that yields no points, and points are deducted for incorrect answers [39]. All participating students take an identical PT in an exam hall with proctoring. The final score on the PT is expressed as a percentage of the maximum attainable score, which is translated into *"Good", "Pass", or "Fail"*, based on the mean, and standard deviation of the students that participated in the same test moment as a relative standard. The scores of the four PTs in every academic year are combined, and translated into a summative decision, followed by the awarding of two study credits (of the in total 60).

After each PT, students can check their answers with an online answer key. For each answer a source is provided for further information, and for some answers a short explanation is given. Additionally, students receive their score and feedback via e-mail (*Appendix 1 – Supplemental Table 1*), and they can access feedback in ProF in the form of a table displaying their individual score, stratified by category and discipline, compared to the overall score of their peers. In ProF, their individual longitudinal test results are visualized in graphs as well [25]. There is no option to download the feedback displayed in ProF. Students receive information about the PT, and the use of ProF through a lecture in each of the bachelor years. Reflection on the feedback with their supervisor is optional.

Due to the COVID-19 restrictions, some of the PTs in the LUMC during the academic years 2020–2021 and 2021–2022 were taken from home by students, via a digital assessment platform. As the COVID-19 restrictions intensified during the pandemic (e.g. total lockdown), exam conditions varied as well. Some of the online PTs used online proctoring software, and were summative (e.g., PT1 and PT2 in 2021–2022). However, in February 2021 (PT3), we could not access the online proctoring system due to logistic reasons. Part of the students could take the PT in the exam hall, but its capacity was largely reduced due to COVID-19 regulations. As a result, the exam hall could only harbour one cohort, i.e. the third-year students. Second-year students took the PT from home, online and non-proctored. As a result, the PT was *summative* for third-year students, and *formative* for second-year students. *Figure 1* shows which PTs were formative, and summative for these two cohorts. We show the situation for these two cohorts only, because these cohorts are our main focus. Students were instructed to take the *formative* PT as a usual (proctored) PT, without using study materials, but without proctoring, we could not verify if students followed these instructions. Participation in these PTs was mandatory, but the test results were not taken into account for the rewarding of study credits. Therefore, these non-proctored PTs turned into purely formative assessments. Hereafter we will call this PT the *formative* PT, whereas the proctored PT that counts towards study credits will be called the *summative* PT.

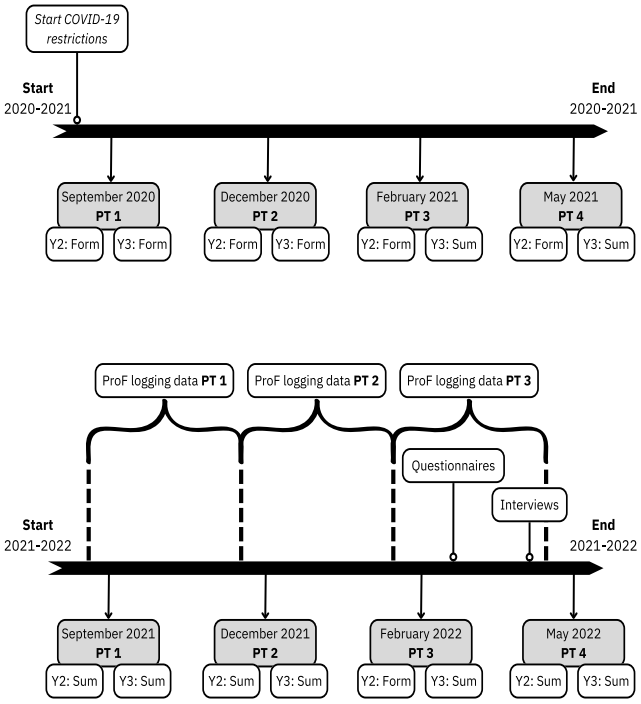


Figure 1. Timeline of progress tests and associated data collection during the academic year 2020-2021 (top) and 2021-2022. PT = progress tests; Y2 = second-year students; Y3 = third-year students; Form = formative; Sum = summative. In 2021-2022, PT3 was purely formative for Y2 students because there was no access to online proctoring, and summative for Y3 students.

Participants

All second-, and third-year bachelor medical students at the LUMC who participated in the PT session on February 2, 2022 (PT 3 of 2021–2022), were eligible for participation in the questionnaire part of the study, and all second- through six-year medical students at the LUMC were eligible for the interviews. The PT session on the February, 2, 2022 was purely formative (*formative* PT) for second-year students, while the result of the PT was taken into account for study credits in third-year students (*summative* PT). Inclusion criteria for semi-structured interviews were (1) participation in at least four of the six PTs between September 2020, and December 2021, and (2) participation in both a *formative* and *summative* PT. In total, 1286 students met our inclusion criteria. Students were sampled using maximum variation sampling based on ProF logging sessions, study-year, and PT results to ensure the representation of multiple perspectives [40]. Sampling of these students was informed by quantitative data, such as ProF logging sessions and study-year. The groups for the number of ProF logging sessions were based on the distribution among all students who participated in the PTs. The PT results were divided in two groups: *"fail"* or *"pass/good"*. If a student had failed on at least one PT, the student was assigned to the *"fail"* group (*n*=410). The other students were assigned to the *"pass/good"* group (*n*=876). We initially approached 140 students that met our sampling strategy, of whom 18 were interested. The distribution

of these students was a good representation of our sampling groups, so we invited all 18 students for an interview. After this initial sampling, second-year students were still underrepresented compared to third-year students, so we decided to sample and approach additional second-year students. Three students replied, who were all included in our study. In total, this resulted in 21 interviews, and a more equal distribution among second- (*n*=6) and third-year students (*n*=8).

Data collection

Questionnaire and ProF logging data

A questionnaire was completed either digitally or on paper (*Appendix 2 – Questionnaire*). It measured perceived assessment condition, test preparation, feedback consultation, and active use of feedback. *Perceived assessment condition* was measured with two MCQs ('formative, summative or don't know', and 'high, intermediate or low stakes'). These items were added to compare true assessment conditions with perceived assessment conditions. *Preparation* and *feedback consultation* after the PT were measured with two yes/no questions. Students were also asked to select explanatory reasons for their answers. *Active feedback use* was measured with the Active Use of Feedback (AUF) scale. This scale consists of seven 6-point, positively packed Likert-items, and is part of the validated, revised version of the Students Conceptions of Feedback (SCoF) Questionnaire [6]. Six of the seven original AUF scale items, and one item of the *'Enjoyment'* (ENJ) subscale were used. The items were adapted to the context of the PT (e.g., 'tutor' was replaced by 'progress test'). Two items were excluded because they did not apply to the specific context or were very similar to another item. The items were translated to Dutch using a forward-backward translation method. The content, and structure of the questionnaire were assessed by three master students using a thinking aloud method. Two weeks after the PT scores and feedback were made available, students received the digital questionnaire by e-mail. We also visited lectures, and working groups to hand out paper questionnaires. The students received up to two digital reminders. Age, PT grades, and ProF logging data of all students (both responders and non-responders of the questionnaire) were derived from the university's student administration system.

Interviews

We developed an interview guide to explore which factors affect feedback use in progress testing (*Appendix 3 – Interview guide*). The interview data were part of a more comprehensive study on factors influencing feedback use in progress testing [41]. In this study, we only selected interview data about students' perceptions of feedback use in the context of a formative and summative PT. Besides their own perceptions, we asked students to reflect on the ProF logging data from all bachelor, and master students in relation to formative and summative PTs during the COVID-19 pandemic (*Appendix 4 – Supplemental Figure 1*).

The principal investigator (EvW) conducted two pilot interviews with fourth-year medical students, which resulted in minor revisions in the interview guide to improve clarity and structure. The pilot interviews were not included in the study. EvW conducted the interviews with 21 students (*Appendix 5 – Supplemental Table 2*) via online meetings in Microsoft Teams in April and May 2022. Participants were invited by e-mail, and received an electronic gift card in return for participation. The interviews took 30–60 min, and were audiotaped. The audiotapes were transcribed verbatim, and anonymized before analysis. The timeline of the data collection from the different sources are depicted in *Figure 1*.

Data analysis

Questionnaire

Descriptive statistics were calculated for the demographics, and perceived assessment con-ditions. Standardized mean differences (SMD) were calculated to quantify baseline group differences between the *formative*, and *summative* PT-groups, and to explore potential response bias (non-responders versus responders) [42]. Logistic regression analyses were used to study the effect of assessment condition on test preparation, and feedback consultation. Cronbach's α was calculated to assess internal consistency of the AUF scale items. Differences between the formative, and summative PT-group were assessed by an unpaired t-test (total mean score on the AUF scale items), and chi-squared tests (multiple-choice questions on preparation, and feedback consultation). Subgroup analyses were performed on students from whom the perceived assessment condition (formative or sum-mative) matched the actual assessment condition. We used the actual assessment conditions for our main analyses, because students were well aware of the physical difference in tests condition (i.e., from home without any webcam observation versus in an exam hall with continuous supervision), and therefore we assumed that this would be a more important discriminative factor than the perceived formative or summative test condition.

ProF logging data

All ProF sessions were included for analysis, independent of the number of pageviews or duration of their session. The average number of ProF sessions per student was calculated for the PTs in September 2021 (PT 1), December 2021 (PT 2), and February 2022 (PT 3). We chose a time range of one week before the PT until one week before the subsequent PT to assess both feedback consultation before (for preparation) and after the PT. Linear regres-sion was used to estimate the effect of assessment condition on average number of ProF ses-sions for the PT in February 2022, adjusted for ProF-sessions on previous PTs (December 2022, and September 2021). Adjustment for ProF-sessions in December 2022, and September 2021 was done by adding the number of ProF-sessions around these PTs as two separate covariates in our linear regression formula. To cross-check the self-reported ProF consultation after the PT on the questionnaire, we analysed the ProF logging data of the responders in the week of the PT in February 2022 until the end of the questionnaire administration (6 weeks later). For both the questionnaire, and ProF logging data analysis, statistical significance was determined by a 95% confidence interval (CI) and $p<0.05$. Data were analysed using R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

Interviews

Data analysis started after four interviews, which led to small adjustments in the interview guide to specify the questions more. The remainder of the data analysis took place after all interviews were completed. Because the extensive literature on feedback use can be integrated by *a priori* themes that guide the deductive analysis, we used template analysis in which hierarchical coding and development of successive coding templates is used [43]. Our *a priori* themes were based on EVT in the context of TTM [18, 19, 22]. Two independent coders (EvW and FvB) coded interviews 1–6 in Atlas.ti. This was discussed afterwards together and with a third researcher (AL) to reach consensus on the initial template, which was then used to guide the coding of the next interviews. Analysis of interviews 7–14 was used to further revise the initial template (EvW and FvB) which in turn was used to code interviews 15–21, and develop the final template. Only minor revisions were made to the revised initial template, and no new themes related to

the research question raised in the development of the final template, indicating theoretical sufficiency after interview 14 [44, 45]. The final template was discussed with the research team (EvW, FvB, AL, JB). During the iterative process, elements of the EVT and goal-orientation theory were incorporated in the template. Eventually, EvW reread, and recoded all interviews with the final template to ensure all relevant information to answer the research question was included in the template. With this final template, a thematic-map was constructed to identify connections between the themes and codes. Member checking was done using the Synthesized Member Checking (SMC) method [46], and yielded no adjustments.

Reflexivity

We considered and discussed (inter)personal reflexivity throughout our data collection, and analysis process using a reflective diary and critical dialogues regarding our interpretations of the data [47]. The reflective diaries created awareness of personal expectations, assumptions, and reactions to the participants and data, and were used to guide the dialogues between the investigators. In interviewing the students, EvW experienced that she could easily relate to the participants, because of her own medical background and experience with the PT. This created an open atmosphere, in which the students felt comfortable to talk openly about their experiences and perceptions. Influenced by her scientific background in (bio)medicine, EvW attempted to attain as much objectivity and produce rigorous qualitative research by using maximum variation sampling, member checking, and reflexivity throughout the data collection and analysis. The other researchers were an educational consultant and researcher in medical education (FvB) and a medical doctor with experience in clinical teaching and educational research (AL). FvB has been trained to conduct research in an empirical way during his studies in cognitive psychology. As such, he supported using theoretical concepts from feedback literature to formulate *a priori* themes. This theory-driven approach may have influenced the results. AL is a member of the national PT working group and a PT examiner, which might have led to assumptions on study behaviour based on her experience with the PT and conversations with students in the past. Her involvement with the PT was very valuable in reflecting on the interview data, and placing it in the right context.

Results

Demographics and perceived assessment condition

Of 316 students who took the purely formative PT (*formative* PT-group), 113 students participated in the questionnaire (response rate: 35.8%). In the *summative* PT-group, 154 students participated in the questionnaire (response rate: 50.0%) from which 3 students were excluded due to incomplete reply to the questionnaire (*Appendix 6 – Supplemental Figure 2*). Responders (*n*=264) and non-responders (*n*=354) differed in fail/pass/good grade and average ProF logging sessions (mean (SD); 1.29 (1.60) versus 0.71 (1.87), for responders versus non-responders) (*Appendix 7 – Supplemental Table 3*). In both the formative and summative PT-group, 70% of the responders were female, and the distribution of the grades was similar among the groups (*Table 1*). Regarding the perceived stakes of the PT in February 2022, 50% versus 13% of students perceived the PT as low stakes, 42% versus 62% as intermediate stakes, and 7% versus 25% as high stakes for *formative* and *summative* PT-group respectively (*Appendix 8 – Supplemental table 4*). The perceived assessment conditions formative and summative can also be found in *Appendix 8 – Supplemental Table 4*.

Table 1. Baseline characteristics of the responders of the questionnaire in the *formative* and *summative* progress test group.

	Overall (*n*=264)	Formative Test (*n*=113)	Summative Test (*n*=151)	SMD[a]
Age, median (IQR)	21 (20, 21)	20 (20, 21)	21 (21, 22)	0.841
Female, n (%)	185 (70)	80 (71)	105 (70)	0.022
Grade, n (%)				
Fail	24 (9)	11 (10)	13 (9)	0.034
Pass	106 (40)	46 (41)	610 (40)	0.020
Good	134 (51)	56 (50)	78 (51)	0.040
Proportion passed earlier PTs, %				
Sep '21[b]	223 (87)	99 (88)	124 (87)	0.030
Dec '21	226 (87)	97 (86)	129 (88)	0.059

SMD = standardized mean difference; IQR = interquartile range; PT = progress test.

[a]A standardized mean difference >0.1 may point towards meaningful imbalance between groups.

[b]PTs of Sep '21 and Dec '21 were summative tests

In the following paragraphs we present the results for each research question: the effect of a *summative* PT and a *formative* PT on medical students' (1) test preparation (questionnaires, and interviews), (2) factors that influence test-taking motivation, and the use of feedback (interviews), and (3) self-reported and actual feedback use after the test (questionnaires and ProF logging data, and interviews).

Test preparation

Logistic regression showed no significant association between assessment condition and preparation for the PT (adjusted OR [aOR] 1.26, 95% CI 0.57–2.76) (*Table 2*). A similar result was found in the subgroup analysis (aOR 1.83, 95% CI 0.72–4.64).

Table 2. Test preparation, feedback consultation and active use of feedback of students in the *formative* and *summative* progress test-group.

	Formative Test	Summative Test	Crude OR (95% CI)	Adjusted OR (95% CI)[a]	p-value
True formative and summative					
Number of individuals	113	151			
Preparation, n (%)	14 (12)	28 (19)	1.61 (0.80-3.22)	1.26 (0.57-2.76)	0.568
Feedback consultation, n (%)					
Answer key	22 (19)	56 (37)	2.44 (1.38-4.32)	1.92 (1.04-3.55)	0.038
Feedback e-mail	89 (79)	126 (83)	1.36 (0.73-2.53)	1.00 (0.49-2.05)	0.996
Feedback ProF	41 (36)	86 (57)	2.32 (1.41-3.83)	1.92 (1.10-3.34)	0.021
None	20 (18)	13 (9)	2.28 (1.08-4.81)	1.86 (0.80-4.32)	0.149
ProF logging data[c], n (%)	26 (23)	58 (38)	2.09 (1.21-3.61)	1.89 (1.03-3.44)	0.039
Number of individuals[d]	90	135	**t-value**	**95% CI**	
Active use of feedback, mean (SD)	3.2 (0.9)	3.1 (0.9)	1.09	-0.10-0.36	0.275[e]
Perceived formative and summative[b]					
Number of individuals	79	128			
Preparation, n (%)	8 (10)	25 (20)	2.15 (0.92-5.05)	1.83 (0.72-4.64)	0.205
Feedback consultation, n (%)					
Answer key	18 (23)	48 (37)	2.03 (1.08-3.84)	1.47 (0.73-2.94)	0.280
Feedback e-mail	60 (76)	106 (82)	1.53 (0.76-3.04)	1.07 (0.48-2.39)	0.876
Feedback ProF	26 (33)	74 (57)	2.79 (1.55-5.02)	2.25 (1.18-4.31)	0.014
None	15 (19)	11 (9)	2.49 (1.08-5.75)	1.94 (0.74-5.05)	0.175
ProF logging data[c], n (%)	17 (22)	48 (38)	2.19 (1.15-4.17)	1.80 (0.88-3.66)	0.106
Number of individuals[d]	62	114	**t-value**	**95% CI**	
Active use of feedback, mean (SD)	3.2 (0.8)	3.1 (0.8)	0.94	-0.14-0.38	0.351[e]

[a]Adjusted for age and result progress test December 2021 (fail, pass, good).

[b]Subgroup analysis; Perceived formative/summative = students in the purely formative/summative test group who knew it was formative/summative.

[c]Real-time ProF logging data in week 05 (PT administration) until week 11 (end of questionnaire administration).

[d]Students who consulted feedback in e-mail or progress test feedback system.

[e]Unpaired t-test.

Regarding the reasons why students did not prepare for the PT, 27% of the students in the *formative* PT-group stated on the questionnaire that the PT was not important compared to 1% of the students in the *summative* PT-group ($p<0.001$, *Appendix 9 – Supplemental Table 5*). In the subgroup analysis this difference became more prominent (26 (37%) versus 0 (0%), $p<0.001$ for *perceived formative* versus *perceived summative*). Other reasons for not preparing were a lack of consequences and not knowing how to prepare.

In the interviews, many students mentioned that the lack of consequences and the possibility to look up answers in the formative PT affected their test preparation:

"My preparation for a formative PT is worse. I still look up some things in advance which I just want to know, but there is less pressure, so if it does not work out or if I don't really feel like doing it, then I think, well, if a question comes up I don't know, I can just look it up." (Interview #3)

Factors that influence test-taking motivation and feedback use

The value given to the *formative* and *summative* PT influenced students test-taking motivation, and determined how students behaved during the formative PT (i.e., test-taking behaviour). The majority of students valued the *summative* PT as more important compared to the *formative* PT, because of its' consequences for study progress, and the more formal test-setting compared to the *formative* PT (on location vs. at home). We call these students 'performance-oriented' (*Figure 2*, upper path):

"Ultimately, you take each test for the study credits. You follow the lessons to learn something, but I do not make a test to learn from it. I make a test to see if my learning was successful. And whether or not I can receive the credits so I can continue." (Interview #4)

On the other hand, 'learning-oriented' students valued the test and its' feedback as a moment of self-assessment and reflection, regardless of the assessment condition. Their main focus in both the *formative* and *summative* PT was to assess their current knowledge level, gain insights in their own strengths and weaknesses and learn from what they did wrong (*Figure 2*, lower path).

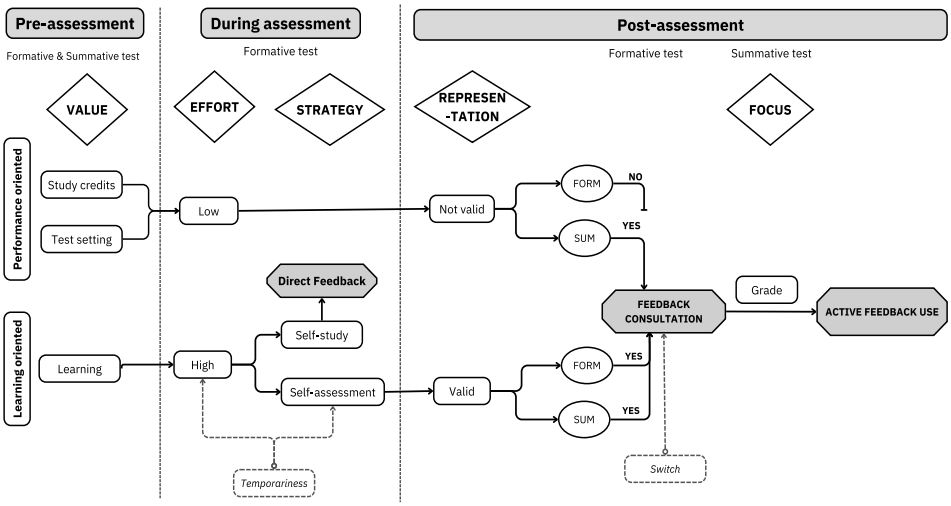


Figure 2. Thematic map showing the connections between the themes (triangular shapes on the top) and codes (boxes below the themes) pre-, during, and post-assessment for performance-oriented (upper path) and learning-oriented students (lower path). Form = formative progress test; Sum = summative progress test.

Test-taking behaviour: effort and strategy

We distinguished two subthemes within test-taking behaviour: effort and strategy. These themes only relate to the *formative* PT, because the low-stakes and lack of supervision were perceived as an opportunity to adapt their test-taking behaviour according to their values and goals in relation to the PT.

Learning-oriented students tended to put significant effort in the *formative* PT, as they wanted to be able to reflect effectively on their performance. In contrast, performance-oriented students put less effort in taking the *formative* PT, reflected by a higher proportion of guessing, looking up answers on the internet or being less focused during the test:

"I think that I guessed more of the answers in the online (formative) test when I recognized an answer vaguely from a previous course. I did not know the answer completely for sure, but I was doubting between three options and then I just guessed because it did not matter so much." (Interview #7)

Students employed different test-taking strategies in the *formative* PT, which could be divided in self-study and self-assessment. The self-study strategy was characterized by using study materials to look up answers during the test, mainly with the idea to learn directly from it. By looking up answers for questions, they generated instant feedback for themselves and hence used the *formative* test as a guide for self-study:

"Well, I thought if I look it up right away I will learn something from it, because then I know the answer. And if I will not look at it anymore afterwards, then I actually do not learn so much either, because I don't know if my answers were correct or incorrect." (Interview #9)

In the self-assessment strategy, students approached the *formative* PT as if it were a *summative* PT and refrained from looking up answers. They used the test as a realistic self-assessment of their current knowledge:

"When you get the result, that you have some sort of measurement of how good you actually are at it. Because otherwise (when using study material) I have the idea that it does not make sense at all to take that test." (Interview #10)

Contextual factors: Temporariness

Many students took into account that the *formative* PTs were only temporary and that in the near future, they would become summative again. This temporariness encouraged them to make the *formative* test just as seriously as the *summative* test, with an indirect focus on study credits relating to the performance-oriented mindset:

"Of course I could have looked it all up, but then I think you will fall at a certain moment. I think you cannot sustain that when the test is proctored again. And then it's only annoying that you're going to drop in your score again." (#Interview 14)

Self-reported and actual feedback use after the test

Feedback consultation

Students who took the *summative* PT reported consulting ProF (aOR 1.92, 95% CI 1.10–3.34) and the answer key (aOR 1.92, 95% CI 1.04–3.55) more often than students who made the *formative* PT. In *perceived formative* versus *summative*, the effect on feedback consultation in ProF became more evident (aOR 2.25, 95% CI 1.18–4.31) and the adjusted effect on the answer key consultation was not observed

(aOR 1.47, 95% CI 0.73–2.94) (*Table 2*). The sensitivity analysis with real-time ProF logging data of the responders showed the same trend as the self-reported data, but yielded lower overall numbers (26 (23%) versus 58 (38%), aOR 1.89, 95% CI 1.03–3.44; for *formative* versus *summative*; 17 (22%) versus 48 (38%), aOR 1.80 (0.88–3.66); for *perceived formative* versus *perceived summative*) (*Table 2*). Besides the sensitivity analysis using only ProF logging data of the responders, we also analysed the ProF logging data of all participating students including the non-responders. This analysis showed that there were more ProF logging sessions around the *summative* PT in February 2022 than around the *formative* PT (β:0.444, *p*<0.001). After adjustment for logging behaviour in earlier *summative* PTs in September 2021 and December 2022, this effect remained significant (β:0.251, *p*:0.003) (*Figure 3, Appendix 10 – Supplemental Table 6*).

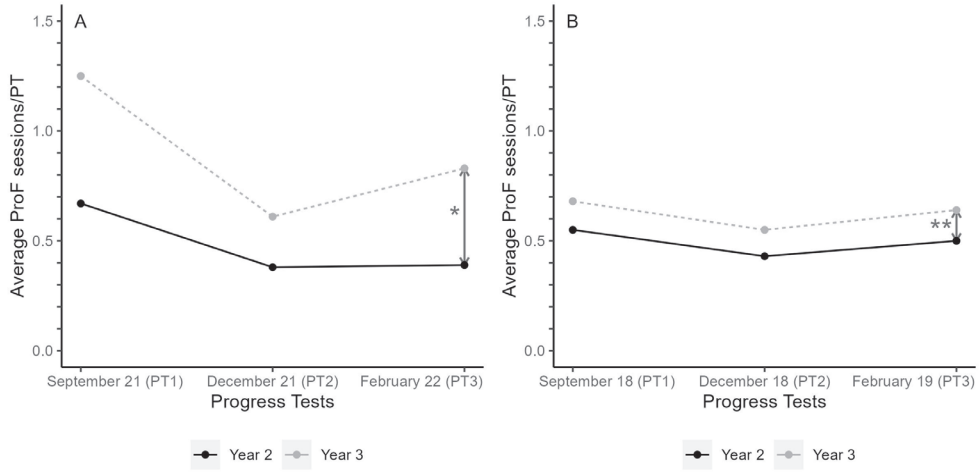


Figure 3. Average ProF sessions in year 2 (black line) and year 3 (dotted grey line) for the progress tests in September 2021, December 2021, and February 2022. Each point on the curve represents the average number of ProF sessions per student; *: crude beta: 0.444, p<0.001; adjusted beta: 0.251, p: 0.003.

Students who reported not to consult the feedback after the *formative* PT on the questionnaire more often considered the test as not important (15 (22%) versus 1 (2%), *p*<0.001 for *formative* versus *summative*) (*Table 3*). This resonates with the perceptions of the performance-oriented students:

"I think that you are more motivated when the test counts for study credits, so then afterwards you will be more interested in how you performed because it counts." (Interview #10)

However, qualitative data also revealed the learning-oriented students who valued the feedback of both assessment conditions for their learning (*Figure 2,* lower path):

"I look at the test result to know what questions I did wrong and to learn from it. And it doesn't matter to me whether it is formative or summative, because that remains the same. I still want to know which questions I got right and wrong. And I still want to know, I still want to learn from the things I did wrong. So, then it doesn't matter if the test was formative or summative." (Interview #3)

In the questionnaires, the *summative* PT-group more often found the feedback not useful (1 (1%) versus 8 (13%), *p*:0.015 for *formative* versus *summative*). Similar results were found for perceived assessment conditions (*Table 3*). Other reasons for not consulting feedback included no awareness of or not understanding ProF, not knowing how to use the feedback, and a lack of interest.

Table 3. Reasons for not using the progress test feedback system in the *formative* and *summative* progress test group.

	Formative Test	Summative Test	p-value[a]
True formative and summative			
Number of individuals	67	63	
No ProF use, n (%)			
Findability[c]	25 (37)	14 (22)	0.061
Time	24 (36)	16 (25)	0.178
Effort	7 (10)	1 (2)	0.062[d]
Importance	15 (22)	1 (2)	0.000
Grade	33 (49)	22 (35)	0.083
Utility	1 (1)	8 (13)	0.015[d]
Answer key	3 (4)	7 (11)	0.200[d]
Other	4 (6)	12 (19)	0.025
Perceived formative and summative[b]			
Number of individuals	48	52	
No ProF use, n (%)			
Findability	20 (42)	12 (23)	0.055
Time	16 (33)	15 (29)	0.628
Effort	5 (10)	1 (2)	0.102
Importance	13 (27)	1 (2)	0.000
Grade	21 (44)	20 (38)	0.591
Utility	0 (0)	6 (12)	0.027[d]
Answer key	1 (2)	5 (10)	0.207[d]
Other	3 (6)	10 (19)	0.054

ProF = progress test feedback system.
[a]Chi-squared test.
[b]Subgroup analysis; Perceived formative/summative = students in the formative/summative test group who knew it was formative/summative.
[c]Findability: ***"I do not know where I can find the feedback"***; Time: ***"I did not have time to look at the feedback"***; Effort: ***"I did not put effort in this progress test"***; Importance: ***"I thought this progress test was not important"***; Grade: ***"I got a pass/good for this progress test"***; Utility: ***"I find the feedback not useful"***; Answer key: ***"I already checked my answers with the answer key"***.
[d]Fisher's exact test

Qualitative data revealed that representation of the *formative* test also played a role in feedback consultation (*Figure 2*). Performance-oriented students indicated less interest in the feedback of the *formative* PT, because their low test-taking effort in taking the formative PT did not provide a valid representation of their own knowledge level. Therefore, the feedback was less meaningful to them:

"I think I took a quick look at ProF. That I just looked at that line, but that I thought yes, it is probably now higher than it should be. So I did not attach much value to it." (Interview #4)

This was also the case for students who used study material during the *formative* PT. Besides, these students found it more useful to receive direct feedback during the PT. In contrast, for learning-oriented students who used the *formative* PT as self-assessment the test result was a valid representation,

and they were interested to consult the feedback to assess their strengths and weaknesses.

ProF consultation was relatively high after the first *formative* PT (September 2020, *Appendix 4 – Supplemental Figure 1*). The interviewed students mentioned that this could be explained by curiosity right after switching to formative testing (*Figure 2, 'Switch'*):

"The first time it is always exciting, oh new and what would be my result now that it's online for the first time. And is there a difference maybe with the paper version that I always had. So then it is a bit more interesting and if you've done a few then you just think oh it's going fine, whatever." (Interview #18)

Active feedback use

The internal consistency (Cronbach's α) of the 6-point Likert scale items was 0.85 (>0.80: acceptable) [48, 49]. After deletion of item 7 of the subfactor ENJ the Cronbach's α remained 0.85. We found no difference in the mean total score on the items of AUF and ENJ (3.2 (0.9) versus 3.1 (0.9), t(223): 1.09, 95% CI -0.10-0.36). Comparing perceived assessment conditions yielded the same result (*Table 2*). On item level, item 7 (enjoy) had the highest score (5 (4–6), whereas item 3 (setting goals) and 6 (changing learning) had the lowest score (2 (1–3) in both groups (*Appendix 11 – Supplemental Table 7*).

Students who were interested in the test result and feedback often only consulted the feedback without actively using it. They seemed to use the feedback as a 'thermometer' to assess if they were still at the right 'temperature'. If they were still on the right track, they did not feel the urgency to change anything and engage with the feedback: *"If it ain't broke, don't fix it"(# Interview 4).* An insufficient grade on the other hand was or will be an incentive to act on the feedback and use it to prepare for the next PT:

"Suppose, if I had failed I would think oh dear, then I will really look at what I did wrong, which subject and really do that because you still want to get those study credits." (Interview #18)

Although the formative PT was also graded, grade focus only occurred in the *summative* PT, mainly because an insufficient grade on the *formative* PT had no consequences. Thus, no need was felt to act on the feedback (*Figure 2*).

Discussion

In this mixed-methods study, we compared the effect of a purely formative PT (*formative* PT) with a PT with a summative component (*summative* PT) on medical students' feedback use and test-taking motivation. We triangulated quantitative and qualitative interview data to explain these in the context of a *formative* versus a *summative* PT. Our thematic map (*Figure 2*), based on our qualitative data, in which EVT and goal-orientation frameworks were integrated helped explain our quantitative results, and provided a nuanced picture of the different ways students approached the feedback in the *formative* and *summative* PT. Test preparation was relatively low for both PT assessment conditions and did not differ between groups. Qualitative data showed that test-taking motivation and feedback use relate to how students value the assessment. Performance-oriented students valued the *summative* PT as more important because of its' consequences for study progress, and learning-oriented students valued the PT feedback for their own learning, regardless of the assessment condition.

These orientations influenced their test-taking behaviour (effort and strategy), and feedback consultation (representation of *formative* PT results). Self-reported questionnaire data showed more ProF consultation and use of the answer key after the summative PT compared to the *formative* PT. Actual feedback use, measured by ProF logging data, showed the same results. Students in the *formative* PT-group who did not consult PT feedback more often reported the *formative* PT as unimportant, reflecting the perceptions of performance-oriented students. However, self-reported active feedback use after the PT was relatively low in general and did not differ between groups, which was mainly determined by grade focus.

Test preparation, feedback consultation and test-taking motivation

We measured test preparation and feedback consultation with different data sources. The ProF logging data demonstrate that, in general, students made limited use of ProF to consult feedback, which is important to take into account with the interpretation of our data. Despite low use of ProF, our data can contribute to a better understanding of feedback behaviour. Both questionnaire and interview results suggest that the motivation to prepare and consult feedback relates to how students value the assessment. The interview results revealed that experienced utility value (i.e. usefulness) and attainment value (i.e. importance) of the PT affected test-taking effort, the important component of test-taking motivation, which influenced feedback consultation [22]. This positive relation between value and effort has also been found in test-taking motivation with test performance as outcome [19, 23, 50].

Moreover, our interview data showed that students valued the different PT assessment conditions based on whether they were orientated towards performance or learning. This aligns with the goal-orientation theory, which states that performance-orientated students focus on achievement based on normative standards (i.e. study credits), whereas learning-orientated students focus on achievement based on learning [24]. It seems that students' goal orientation guided their test effort and engagement with the feedback. Although students did not explicitly state goals for the PT in our study, they did show a more general focus on either learning or performance. Below, we elaborate on students' performance and learning orientation in this study.

Performance-oriented students

Performance-orientated feedback behaviour was revealed by our qualitative interview data, and confirmed by our quantitative results. The interviews showed that students found the summative PT more important and valuable, which led to less test-taking effort and feedback consultation after the *formative* PT. Quantitative data confirmed this performance orientation as self-reported, and actual feedback consultation was higher after the *summative* PT. Also, the perception that the PT was not important, and thus ProF consultation or preparation was not needed, was more profound in students who participated in the *formative* PT. These results are visualized in the upper path of our thematic map (*Figure 2*). The performance-oriented students mainly focus on the direct personal consequences (i.e., study-credits) of the test in the pre-assessment phase, see no value in investing effort in the *formative* tests (i.e., low effort during assessment), leading to an invalid representation of their test result, and a decreased motivation to consult the feedback of the *formative* test post-assessment. The study credits in the summative test, on the other hand, motivated these students to consult the feedback.

Learning-oriented students

Our qualitative analysis suggested that students were not also focused on learning. Thus, the interview data further deepened and nuanced our understanding of students' feedback use. Learning-orientated students valued the PT and its feedback as part of their learning process, regardless of the assessment condition (*Figure 2,* lower path). These students took the *formative* test seriously, invested high effort, used it for self-assessment or self-study, and were motivated to consult the feedback of the *formative* test. This aligns with previous research in surgical residents showing that formative assessments promoted a learning-oriented motivation [51]. The strategy of self-study was interesting in that the test itself was used as tool to pay attention to knowledge gaps and generate direct feedback. It is more likely that these students benefit from formative assessments and engage in more self-regulated learning compared to students adopting the performance orientation [52-54].

Some learning-oriented students indirectly also focused on the study credits considering that the *formative* PTs would switch back to *summative*. *Figure 2* shows that this contextual factor (*'Temporariness'*) influenced test-taking effort, and strategy in the *formative* test of these students (with the dotted arrows). They decided to put high effort in the *formative* test, and use it as self-assessment, to make sure they were at the right level to pass the upcoming summative test. Although these students predominantly focus on learning in the *formative* PTs, they do not completely let go their performance-orientation for the study credits of the future *summative* PTs.

Active use of feedback

Besides students' feedback consultation in the e-mail, ProF or by using the answer key, which can be considered a more passive use of feedback, we also measured active use of feedback after the PT by the AUF scale items in our questionnaire [6]. Although students enjoyed receiving feedback, active use of feedback after the PT was relatively low and no difference was found between the groups. The interview data also showed that most students did not actively use the feedback, and that they tended to act only on the feedback when they failed on the *summative* PT. This is illustrated in *Figure 2* in the post-assessment phase, where all students, regardless of their orientation, were driven by the grade in their decision to act on the feedback after consultation. This suggests that failure drove using feedback, regardless of students' learning orientation. However, we could only find qualitative evidence for this, as too few students failed the PT to provide quantitative evidence. As described in previous literature, grade focus strongly limits the likelihood to engage with feedback after a sufficient summative grade [55, 56]. However, students stated even less engagement with the feedback after the *formative* PT, because they lacked a feeling of urgency to change something as this test had no direct consequences for their study progress. These findings are in line with earlier studies on progress testing, where the effect of the feedback on learning was questionable [26, 32, 33, 35, 57]. Although students used the feedback to monitor their progress, and identify strengths and weaknesses in these studies, there was no direct influence on future learning [32, 33].

Implications for practice

Our results suggest that the desired positive effect of formative testing on the learning process is limited in progress testing, with students mainly focusing on performance. Introducing more formative assessments in medical education requires a change in shift in focus towards the learning process (learning-oriented)

rather than the outcome (performance-oriented), and enhancing students' feedback literacy: their ability to effectively engage with and utilize feedback [58]. This involves creating a supportive environment in which students are encouraged to develop feedback literate skills [59]. An example of such a system emphasizing the value of assessment for learning is the programmatic assessment approach. In this approach assessments are no longer divided in formative and summative, but rather represent a continuum of stakes (from low to high). Heeneman *et al.* demonstrated positive results on feedback use of embedding a formative PT in a programmatic assessment system, in which the reflection on the PT, and guidance in the feedback process by mentors in the curriculum is embedded [28]. A supportive assessment environment that emphasizes the understanding of the concept, and purpose of formative testing is key in motivating students and support learning [60, 61].

Strengths and limitations

In the present study we had the unique opportunity to make a direct comparison between two conditions of the same test in one medical curriculum. Except for the assessment conditions, the educational setting was exactly the same for all students and feedback was provided to all students, which facilitated the assessment of the (additional) effect of the summative component over the formative component of assessment on feedback use. Moreover, we analysed both assigned and perceived test conditions, which showed the same trend. Additionally, triangulation of quantitative, and qualitative data was used to increase validity and create a more in-depth understanding of student's values. The triangulation of three data sources also adds to the credibility of our conclusion that the formative PT was associated with less feedback use than the summative PT.

This study also has some limitations. Firstly, this study was conducted at only one medical school, which could limit the transferability to other settings. Secondly, we cannot completely rule out that the difference in study-years between the groups affected feedback behaviour. As third-year students are more experienced with the PT, possibly having a more serious attitude towards their study, this might have resulted in a higher baseline level of feedback use. Nevertheless, test preparation was similar between groups, and the effect found in the ProF logging data remained significant after adjusting for previous ProF use in both years. Moreover, the interview data clarified that the formative and summative component of the PT played a significant role in their feedback behaviour, regardless of their study progress. Thirdly, the responders to our questionnaires were overall students with more ProF logging sessions, and the response rate of the students in the formative PT-group was relatively low. However, the ProF logging data of all students, both responders and non-responders, point towards the same conclusion that feedback consultation was higher after the summative PT. Fourthly, the assessment of a more longitudinal pattern of ProF logging behaviour under summative conditions was hindered by changes in PT conditions before September 2021 (COVID-19) and after February 2022 (new adaptive format). Finally, it must be noted that this study focused on the PT, which is a longitudinal, repetitive and comprehensive assessment. We cannot be sure to what extent these results can be adapted to other contexts, such as a context with a different assessment structure or to end-of-course examinations. The perception of feedback and the feedback behaviour in these other contexts is an interesting question for future research. Moreover, additional research is needed to understand the interaction between the different goal orientations and feedback use.

Conclusion

In conclusion, this study found that students make little use of PT feedback. When they do use PT feedback, a *summative* PT is associated with more feedback consultation compared to a *formative* PT, which can be explained by lower overall test-taking motivation in the *formative* PT and a performance-orientation. Nonetheless, qualitative data also showed learning-oriented students who found the *formative* PT useful and important for their learning, emphasizing that the perceived value of assessment is key to the learning effect of formative testing. Active use of feedback after the PT was low in both assessment conditions and seemed to be affected mostly by high-stakes consequences (i.e., not obtaining enough study credits due to failing the *summative* PT). This might be partly because reflection, and guidance in the feedback process were not embedded in the curriculum. Therefore, it is important to consider the introduction of formative assessments in the medical curriculum very carefully, and make sure students understand its value and are supported in the feedback process.

References

1. Al-Kadri HM, Al-moamary MS, Roberts C, Van der vleuten CPM. Exploring assessment factors contributing to students' study strategies: Literature review. Medical Teacher. 2012;34(sup1):S42-S50.
2. Newble DI, Entwistle NJ. Learning styles and approaches: implications for medical education. Medical Education. 1986;20(3):162-75.
3. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. Medical Teacher. 2011;33(6):478-85.
4. Berkhout JJ, Helmich E, Teunissen PW, van der Vleuten CPM, Jaarsma ADC. Context matters when striving to promote active and lifelong learning in medical education. Medical Education. 2018;52(1):34-44.
5. Black P, Wiliam D. Assessment and Classroom Learning. Assessment in Education: Principles, Policy & Practice. 1998;5(1):7-74.
6. Brown GTL, Peterson ER, Yao ES. Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. British Journal of Educational Psychology. 2016;86(4):606-29.
7. Castro MABE, de Almeida RLM, Lucchetti ALG, Tibiriçá SHC, da Silva Ezequiel O, Lucchetti G. The Use of Feedback in Improving the Knowledge, Attitudes and Skills of Medical Students: a Systematic Review and Meta-analysis of Randomized Controlled Trials. Medical Science Educator. 2021;31(6):2093-104.
8. Koh LC. Refocusing formative feedback to enhance learning in pre-registration nurse education. Nurse Education in Practice. 2008;8(4):223-30.
9. Kulasegaram K, Rangachari PK. Beyond "formative": assessments to enrich student learning. Advances in Physiology Education. 2018;42(1):5-14.
10. Schuwirth LWT, van der Vleuten CPM. The use of progress testing. Perspectives on Medical Education. 2012;1(1):24-30.
11. Scott IM. Beyond 'driving': The relationship between assessment, performance and learning. Medical Education. 2020;54(1):54-9.
12. Seligman L, Abdullahi A, Teherani A, Hauer KE. From Grading to Assessment for Learning: A Qualitative Study of Student Perceptions Surrounding Elimination of Core Clerkship Grades and Enhanced Formative Feedback. Teaching and Learning in Medicine. 2021;33(3):314-25.
13. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Medical Education. 2019;53(1):76-85.
14. Hattie J, Timperley H. The Power of Feedback. Review of Educational Research. 2007;77(1):81-112.
15. Zimmerman BJ. Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. American Educational Research Journal. 2008;45(1):166-83.
16. Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Studies in Higher Education. 2006;31(2):199-218.
17. Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: an unavoidable truth? Anatomical Sciences Education. 2009;2(5):199-204.
18. Baumert J, Demmrich A. Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. European Journal of Psychology of Education. 2001;16(3):441-62.
19. Cole JS, Bergin DA, Whittaker TA. Predicting student achievement for low stakes tests with effort and task value. Contemporary Educational Psychology. 2008;33(4):609-24.
20. Thelk AD, Sundre DL, Horst SJ, Finney SJ. Motivation Matters: Using the Student Opinion Scale to Make Valid Inferences About Student Performance. The Journal of General Education. 2009;58(3):129-51.
21. Wise SL, DeMars CE. Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. Educational Assessment. 2005;10(1):1-17.
22. Eccles JS, Wigfield A. From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. Contemporary Educational Psychology. 2020;61:101859.
23. Zilberberg A, Finney SJ, Marsh KR, Anderson RD. The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. International Journal of Testing. 2014;14:360-84.
24. Elliot AJ, Dweck CS. Competence and Motivation: Competence as the Core of Achievement Motivation. Handbook of competence and motivation. New York, NY, US: Guilford Publications; 2005. p. 3-12.
25. Tio RA, Schutte B, Meiboom AA, Greidanus J, Dubois EA, Bremers AJA, et al. The progress test of medicine: the Dutch experience. Perspectives on Medical Education. 2016;5(1):51-5.
26. Van Der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. Medical Teacher. 1996;18(2):103-9.
27. Dijksterhuis MGK, Schuwirth LWT, Braat DDM, Scheele F. An exploratory study into the impact and acceptability of formatively used progress testing in postgraduate obstetrics and gynaecology. Perspectives on Medical Education. 2013;2(3):126-41.
28. Heeneman S, Schut S, Donkers J, van der Vleuten C, Muijtjens A. Embedding of the progress test in

an assessment program designed according to the principles of programmatic assessment. Medical Teacher. 2017;39(1):44-52.

29. Norman G, Neville A, Blake JM, Mueller B. Assessment steers learning down the right road: Impact of progress testing on licensing examination performance. Medical Teacher. 2010;32(6):496-9.
30. Schüttpelz-Brauns K, Karay Y, Arias J, Gehlhar K, Zupanic M. Comparison of the evaluation of formative assessment at two medical faculties with different conditions of undergraduate training, assessment and feedback. GMS journal for medical education. 2020;37(4):Doc41.
31. Wade L, Harrison C, Hollands J, Mattick K, Ricketts C, Wass V. Student perceptions of the progress test in two settings and the implications for test deployment. Advances in Health Sciences Education. 2012;17(4):573-83.
32. Aarts R, Steidel k, Manuel BAF, Driessen EW. Progress testing in resource-poor countries: A case from Mozambique. Medical Teacher. 2010;32(6):461-3.
33. Given K, Hannigan A, McGrath D. Red, yellow and green: What does it mean? How the progress test informs and supports student progress. Medical Teacher. 2016;38(10):1025-32.
34. Wrigley W, Van Der Vleuten CP, Freeman A, Muijtjens A. A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. Medical Teacher. 2012;34(9):683-97.
35. Yielder J, Wearn A, Chen Y, Henning MA, Weller J, Lillis S, et al. A qualitative exploration of student perceptions of the impact of progress tests on learning and emotional wellbeing. BMC Medical Education. 2017;17(1):148.
36. Barry CL, Horst SJ, Finney SJ, Brown AR, Kopp JP. Do Examinees Have Similar Test-Taking Effort? A High-Stakes Question for Low-Stakes Testing. International Journal of Testing. 2010;10(4):342-63.
37. Fetters MD, Curry LA, Creswell JW. Achieving Integration in Mixed Methods Designs—Principles and Practices. Health Services Research. 2013;48(6 Pt 2):2134-56.
38. Maxwell JA, Mittapalli K. Realism as a Stance for Mixed Methods Research. SAGE Handbook of Mixed Methods in Social & Behavioral Research: SAGE Publications, Inc.; 2010. p. 145-68.
39. Lord FM. Formula scoring and number-right scoring. Journal of Educational Measurement. 1975;12(1):7-11.
40. Onwuegbuzie A, Collins K. A Typology of Mixed Methods Sampling Designs in Social Science Research. The Qualitative Report. 2015.
41. van Wijk EV. Understanding students' feedback use in medical progress testing: A qualitative interview study (Manuscript submitted for publication). 2023.
42. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011;46(3):399-424.
43. Brooks J, McCluskey S, Turley E, King N. The Utility of Template Analysis in Qualitative Psychology Research. Qualitative Research in Psychology. 2015;12(2):202-22.
44. Dey I. Grounding grounded theory: guidelines for qualitative inquiry. San Diego: Academic Press; 1999;282 p.
45. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & Quantity. 2018;52(4):1893-907.
46. Birt L, Scott S, Cavers D, Campbell C, Walter F. Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? Qualitative Health Research. 2016;26(13):1802-11.
47. Olmos-Vega FM, Stalmeijer RE, Varpio L, Kahlke R. A practical guide to reflexivity in qualitative research: AMEE Guide No. 149. Medical Teacher. 2022;0(0):1-11.
48. Lance CE, Butts MM, Michels LC. The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? Organizational Research Methods. 2006;9:202-20.
49. Nunnaly JC, Bernstein, I.H. Psychometric theory. 3rd ed. New York, NY, US: McGraw-Hill;1994.
50. Penk C, Schipolowski S. Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. Learning and Individual Differences. 2015;42:27-35.
51. Lund S, D'Angelo JD, Gardner AK, Stulak J, Rivera M. General surgery resident motivation: the effect of formative compared to summative simulated skills assessments. Global Surgical Education - Journal of the Association for Surgical Education. 2022;1(1):55
52. Ames C. Classrooms: Goals, structures, and student motivation. Journal of Educational Psychology. 1992;84:261-71.
53. Pintrich PR. The Role of Goal Orientation in Self-Regulated Learning. Handbook of Self-Regulation: Elsevier; 2000. p. 451-502.
54. Schunk DH, Pintrich PR, Meece JL. Motivation in education: theory, research, and applications. 3rd ed. Upper Saddle River, N.J.: Pearson/Merrill Prentice Hall. 2008;433.
55. Harrison CJ, Könings KD, Schuwirth L, Wass V, van der Vleuten C. Barriers to the uptake and use of feedback in the context of summative assessment. Advances in Health Sciences Education. 2015;20(1):229-45.
56. Winstone NE, Nash RA, Rowntree J, Parker M. 'It'd be useful, but I wouldn't use it': barriers to university students' feedback seeking and recipience. Studies in Higher Education. 2017;42(11):2026-41.
57. Schüttpelz-Brauns K, Kadmon M, Kiessling C,

Karay Y, Gestmann M, Kämmer JE. Identifying low test-taking effort during low-stakes tests with the new Test-taking Effort Short Scale (TESS) – development and psychometrics. BMC Medical Education. 2018;18(1):101.

58. Molloy E, Boud D, Henderson M. Developing a learning-centred framework for feedback literacy. Assessment & Evaluation in Higher Education. 2020;45(4):527-40.
59. Carless D, Winstone N. Teacher feedback literacy and its interplay with student feedback literacy. Teaching in Higher Education. 2023;28(1):150-63.
60. Heeneman S, Oudkerk Pool A, Schuwirth LWT, van der Vleuten CPM, Driessen EW. The impact of programmatic assessment on student learning: theory versus practice. Medical Education. 2015;49(5):487-98.
61. Nouns ZM, Georg W. Progress testing in German speaking countries. Medical Teacher. 2010;32(6):467-70.

Appendix

Appendix 1 – Supplemental Table 1

Supplemental Table 1A. Feedback of the progress test with the results per category provided by e-mail.

		Individual				Test moment group (*n*=57)							
Description categories	**Number of questions**	**Correct**	**Incorrect**	**?**	**Score**	**Correct**	**Std**	**Incorrect**	**Std**	**?**	**Std**	**Score**	**Std**
01 Respiratory system	13	69	31	0	56	68	13	28	12	4	7	57	18
02 Musculoskeletal system	17	59	41++	0--	38	58	11	31	9	11	10	44	14
03 Mental Health Care	16	69	31+	0	58-	75	14	20	12	5	7	68	18
04 Reproductive system	11	45-	55++	0	27--	58	15	29	13	13	13	48	18
05 Blood, lymph, heart and circulation	24	58	25	17+	48	60	13	29	11	11	9	48	17
06 Hormones and metabolism	13	46-	46++	8	-29	57	13	31	14	12	10	46	17
07 Skin and connective tissue	12	83	17	0	78	80	10	17	10	3	6	74	13
08 Personal, social and prevention aspects	17	29--	71++	0--	4--	52	14	38	14	11	10	35	19
09 Digestive system	17	71	29	0--	61	66	12	26	11	8	7	57	15
10 Kidneys and urinary tract	16	69	25	6	59	71	13	21	11	7	8	63	16
11 Nervous system and senses	17	47--	47++	-6	28--	62	13	26	12	12	11	53	16
12 Knowledge about skills	23	48	39	13	33	49	11	40	11	11	9	32	14
Total	196	57-	38++	5-	42-	62	8	29	6	9	6	51	9

-/--/++/+ low respectively high in comparison with the total group. Results are presented in percentages. Std = standard deviation. ? = question mark option use.

Supplemental Table 1B. Feedback of the progress test with the results per discipline provided by e-mail.

		Individual				Test moment group (n = 57)							
Description disciplines	**Number of questions**	**Correct**	**Incorrect**	**?**	**Score**	**Correct**	**Std**	**Incorrect**	**Std**	**?**	**Std**	**Score**	**Std**
Anatomy	12	58	33	8	46	60	15	34	14	6	9	48	20
Biochemistry, molecular and cellular biology and genetics	18	50	44++	6--	34	46	14	31	12	24	14	34	17
Pharmacology	8	62	25	12	54	65	15	27	14	8	9	54	20
Physiology	11	73	27+	0-	62	73	17	18	12	9	12	65	21
Patho-, immuno- en microbiology	10	50	40	10	33-	57	15	34	15	10	10	44	19
Basic-, supportive subjects	**59**	**58**	**36+**	**7-**	**44**	**58**	**9**	**29**	**7**	**13**	**8**	**47**	**10**
Epidemiology/statistics	7	71+	29	0-	57+	55	23	32	15	12	21	41	26
Metamedica	5	20--	80++	0-	-23--	51	23	38	23	11	14	32	33
Psychiatry/psychology	12	67	33++	0-	54-	73	14	20	12	7	10	65	17
Social medicine	3	33	67+	0-	0	42	26	51	28	8	15	17	37
Behavioural scientific/other subjects	**27**	**56**	**44++**	**0-**	**35-**	**61**	**13**	**30**	**9**	**9**	**10**	**47**	**15**
Surgery	16	69	31	0-	56	67	13	27	12	6	8	56	17
Dermatology/ENT/ opthalmology	14	57	36	7	44	63	14	29	14	8	10	53	18
Geriatrics	8	62	38+	0	44	68	17	29	16	3	6	55	23
Obstetrics/Gynaecology	7	43--	57++	0-	21--	60	14	28	17	13	14	49	19
Family medicine	20	40--	55++	5	21--	61	12	34	12	4	5	49	16
Internal medicine	26	73	19	8+	67	73	11	22	9	5	5	64	14
Paediatrics	12	50-	42++	8	32-	60	15	28	13	12	12	48	19
Neurology	7	43	43+	14	19-	50	17	32	17	18	19	37	21
Clinical subjects	110	57-	37++	5	43--	65	8	28	7	7	6	54	10

-/--/++/+ low respectively high in comparison with the total group. Results are presented in percentages. Std = standard deviation. ? = question mark option use.

Appendix 2 – Questionnaire

In the context of the study on the effect of different assessment conditions of the progress test on learning behaviour we would like to conduct a short questionnaire about the last progress test on the 2nd of February 2022.

This questionnaire **only** concerns the **progress test on the 2nd of February 2022.**
ProF refers to the online feedback system of the progress test.

For part of the students the result of this progress test did not count towards the awarding of credits, while for another part of the students it did. Indicate what applies to you.

1. Did the result of this progress test count towards the awarding of credits?
    - o Yes
    - o No
    - o Don't know

2. How important was this progress test for you (e.g. for obtaining credits, for your study progress, personal reasons). Choose one answer option.
    - o Low
    - o Intermediate
    - o High

The following questions relate to the preparation prior to the progress test on the 2nd of February 2022.

3. Did you prepare for this progress test? Choose one answer option.
    - o Yes
    - o No

*In case you answered **question 3** with "**yes**", you can continue with question 5 and skip question 4. In case you answered **question 3** with "**no**", continue with question 4.*

4. Why did you not prepare for this VGT? Multiple answers possible.
    - o I had no time to prepare
    - o I did not feel like preparing
    - o I always pass my progress test without preparation
    - o I got a pass/good for my previous progress test
    - o I thought this progress test was not important
    - o Other: .....

The following questions relate to the consultation of the feedback after the progress test on the 2nd of February 2022.

5. Did you check the answers of this progress test with the answer key?
    - o Yes
    - o No

6. Did you look at the feedback of this progress test in the email?
   - o Yes
   - o No

7. Did you consult ProF to look at the feedback of this progress test?
   - o Yes
   - o No

*In case your answer to **question 9** was "**no**", you can continue with question 11 and skip question 10. In case you answered "**yes**" to **question 9**, you can continue with question 10 and skip question 11.*

8. What is the reason that you did not look at the feedback in ProF? Multiple answers possible.
   - o I do not know where I can find the feedback
   - o I did not have time to look at the feedback
   - o I did not put effort in this progress test
   - o I thought this progress test was not important
   - o I had a pass/good for this progress test
   - o I find the feedback not useful
   - o I already checked my answers with the answer key
   - o Other: .....

9. Which section of ProF did you look at? Multiple answers possible.
   - o Progress total score (longitudinal)
   - o Total score of this progress test (moment)
   - o Progress on discipline score (longitudinal)
   - o Discipline score of this progress test (moment)
   - o Progress on category score (longitudinal)
   - o Category score of this progress test (moment)
   - o Progress on cluster score (longitudinal)
   - o Cluster score of this progress test (moment)
   - o I do not know

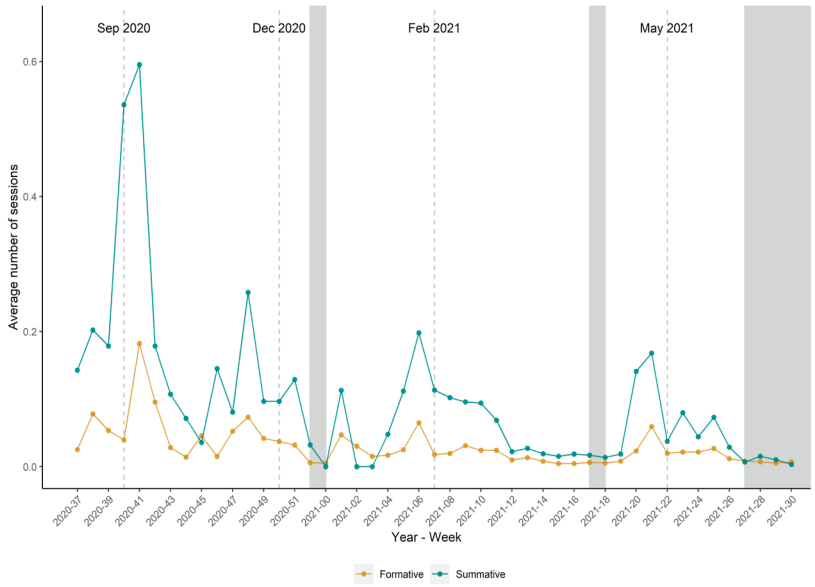
*The following seven items relate to the **consultation of feedback in the email or ProF after** the progress test on the 2nd of February 2022. For the statements below, indicate the extent to which you agree or disagree with each statement (1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = either agree or disagree, 5 = somewhat agree, 6 = agree, 7 = strongly agree).*

1. I actively use the feedback to help me improve.
2. I pay attention to the feedback.
3. I use the feedback to set goals for the next progress test.
4. I look at the feedback to see what I did wrong.
5. The feedback makes me try harder.
6. The feedback changes the way I learn and study.
7. I enjoy getting the feedback.

Appendix 3 – Interview guide

Part 1. Own feedback experiences
1. Do you prepare for the progress test?
How do you prepare?
What determines whether you prepare for the progress test?
2. Do you consult the result of the progress test?
Which methods do you use to consult the test result?
What determines whether you look at the test result? And what is the role of the test condition in this?
3. Do you use the result of the progress test?
What do you do with this information?
What determines whether you use the feedback? And what is the role of the test condition in this?
4. Are you aware of the online feedback system (ProF)? Only asked if students did not mention ProF yet.
Why are you not using this feedback system?
What do you think is the reason that you are not aware of ProF?
Part 2. Reflection and interpretation of ProF logging data (graph)
1. Can you describe what you see?
2. What do you think when you see these data?
3. How would you explain and/or interpret these data?
Part 3. Perception of progress test and feedback
1. What is your perception of the progress test? And which place does it have in your study program?
2. What is your perception of the way(s) the test result is presented to you?
Do you have any suggestions to improve this?

Appendix 4 – Supplemental Figure 1



Supplemental Figure 1. The average number of ProF sessions for the progress tests of September 2020, December 2020, February 2021, and May 2021 divided in students who participated in the formative (yellow line) or summative condition (blue line) which is shown to the interviewees for reflection and possible explanations for the trend.

Appendix 5 – Supplemental Table 2

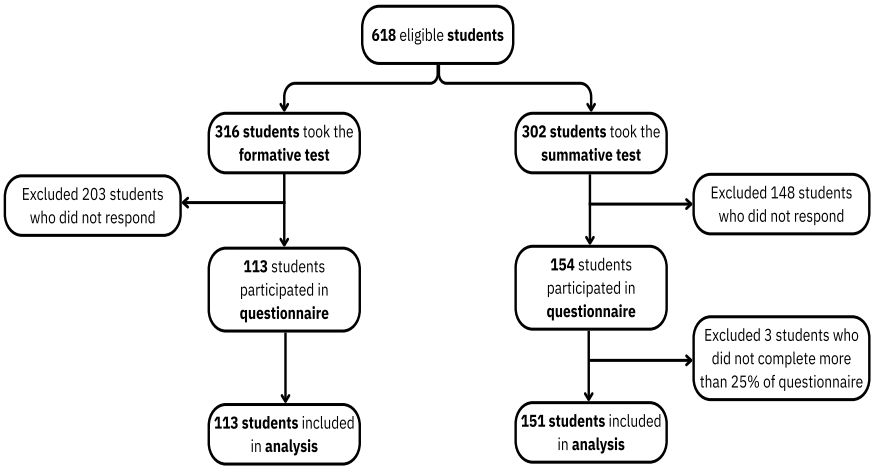
Supplemental Table 2. Descriptive characteristics of interviewees.

Number of logging sessions[a]	0		1		2 to 4		>5		
Grade	Fail	Pass/ Good	Fail	Pass/ Good	Fail	Pass/ Good	Fail	Pass/ Good	Total year (M/F)
Year 2	11[c]	3[c], 10[c]		21[c]		20[b],19[c]			6 (1/5)
Year 3	7[c]	18[c]	4[b]	9[c]		5[b], 8[b]	2[c]	15[b]	8 (4/4)
Year 5	6[c]		14[c]		16[c]	17[b]			4 (1/3)
Year 6	13[b]			1[c]		12[b]			3 (2/1)
Total Fail and Pass/Good (M/F)	4 (1/3)	3 (0/3)	2 (1/1)	3 (0/3)	1 (0/1)	6 (5/1)	1 (0/1)	1 (1/0)	

[a]ProF logging sessions from September 2020 to January 2021. M = male, F = female.

[b]Male. [c]Female.

Appendix 6 – Supplemental Figure 2



Supplemental Figure 2. Flowchart of the participants of the questionnaire.

Appendix 7 – Supplemental Table 3

Supplemental Table 3. Baseline characteristics of responders and non-responders of the questionnaire.

	Responders (*n*=264)	Non-responders (*n*=354)	SMD[a]
Age, median (IQR)	21 (20, 21)	21 (20, 22)	0.107
Female, n (%)	185 (70)	244 (69)	0.022
Grade, n (%)			
Fail	24 (9)	64 (18)	0.266
Pass	106 (40)	196 (55)	0.304
Good	134 (51)	94 (27)	0.508
Proportion passed earlier PTs, %			
September 2021	223 (87)	286 (85)	0.058
December 2021	226 (87)	249 (73)	0.355
ProF sessions, mean (SD)[b]	1.29 (1.60)	0.71 (1.87)	0.334

SMD, standardized mean difference. IQR, interquartile range. PT: progress test.

[a]A standardized mean difference >0.1 may point towards meaningful imbalance between groups.

[b]Period of consultation for each PT starts the week before the PT and ends the week before the next PT.

Appendix 8 – Supplemental Table 4

Supplemental Table 4. Students' idea of assessment condition and perceived stakes of the progress test.

	Formative Test (*n*=113)	Summative Test (*n*=151)
Students' idea of condition (Q1), n (%)		
Formative	79 (70)	8 (5)
Summative	11 (10)	128 (85)
Don't know	23 (20)	14 (9)
Perceived stakes (Q2), n (%)		
Low	57 (50)	19 (13)
Intermediate	48 (42)	94 (62)
High	8 (7)	37 (25)

Q1: *Did the result of this progress test count towards the awarding of credits?*

Q2: *How important was this progress test for you (e.g. for receiving credits, for your study progress, personal reasons)?*

Appendix 9 – Supplemental Table 5

Supplemental Table 5. Reasons for not preparing for the *formative* or *summative* progress test.

	Formative Test	Summative Test	p-value[a]
True formative and summative			
Number of individuals	99	123	
No preparation, n (%)			
Time[c]	35 (35)	42 (34)	0.765
Motivation	18 (18)	11 (9)	0.036
Need	69 (70)	92 (75)	0.543
Grade	39 (39)	46 (37)	0.671
Importance	27 (27)	1 (1)	<0.001
Other	5 (5)	6 (5)	1.000[d]
Perceived formative and summative[b]			
Number of individuals	71	103	
No preparation, n (%)			
Time	27 (38)	35 (34)	0.491
Motivation	14 (20)	10 (10)	0.050
Need	45 (63)	78 (76)	0.134
Grade	29 (41)	39 (38)	0.584
Importance	26 (37)	0 (0)	<0.001
Other	4 (6)	4 (4)	0.715[d]

[a]Chi-squared test.

[b]Subgroup analysis; Perceived formative = students in the formative test group who knew it was formative; Perceived summative: students in the summative test group who knew it was summative.

[c]Time: *"I had no time to prepare"*; Motivation: *"I did not feel like preparing"*; Grade: *"I got a pass/good for my previous progress test"*; Importance: *"I thought this progress test was not important"*.

[d]Fisher's exact test.

Appendix 10 – Supplemental Table 6

Supplemental Table 6. Average ProF logging sessions for each progress test

	Number of individuals	Average ProF sessions, mean (95% CI)	p-value[b]
PT September 2021[c]			
Year 2	316	0.67 (0.52-0.82)	
Year 3	305	1.25 (1.03-1.46)	
PT December 2021			
Year 2	316	0.38 (0.19-0.57)	
Year 3	305	0.61 (0.44-0.79)	
PT February 2022[d]			
Year 2	316	0.39 (0.27-0.51)	
Year 3	305	0.83 (0.66-1.01)	<0.001

ProF = progress test feedback system; PT = progress test.

[a]Population based on participants of the PT in February.

[b]Unpaired t-test.

[c]Period of consultation for each PT starts the week before the PT and ends the week before the next PT.

[d]PT February 22 summative for year 3, formative for year 2.

Appendix 11 – Supplemental Table 7

Supplemental Table 7. Median scores (IQR) of the 6-point Likert scale items of *Active use of feedback* and its subfactor *Enjoyment* in the *formative* and *summative* progress test-group.

	Formative Test	Summative Test
True formative and summative		
Number of individuals[a]	91	135
Feedback use, median (IQR)		
Item 1[c]	3 (2-3)	3 (2-3)
Item 2	3 (3-4)	3 (2-4)
Item 3	2 (1-3)	2 (1-3)
Item 4	4 (3-5)	4 (3-5)
Item 5	3 (2-4)	3 (2-4)
Item 6	2 (1-3)	2 (1-3)
Item 7	5 (4-6)	5 (4-6)
Perceived formative and summative[b]		
Number of individuals	62	114
Feedback use, median (IQR)		
Item 1	3 (2-3)	3 (2-3)
Item 2	3 (3-4)	3 (2-4)
Item 3	2 (1-3)	2 (3-4)
Item 4	4 (3-5)	4 (3-4)
Item 5	3 (2-4)	3 (2-4)
Item 6	2 (1-3)	2 (1-3)
Item 7	5 (4-6)	5 (4-6)

IQR = interquartile range.

[a]Students who consulted feedback in e-mail or progress test feedback system.

[b]Subgroup analysis; Perceived formative= students in the formative test group who knew it was formative; Perceived summative = students in the summative test group who knew it was summative.

[c]Item 1 = I actively use the feedback to help me improve; Item 2 = I pay attention to the feedback; Item 3 = I use the feedback to set goals for the next progress test; Item 4 = I look at the feedback to see what I did wrong; Item 5 = The feedback makes me try harder; Item 6 = The feedback changes the way I learn and study; Item 7 = I enjoy getting the feedback.