Received: 13 March 2023 | Accepted: 10 February 2024

DOI: 10.1111/bjet.13447

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Comparing reusable, atomic feedback with classic feedback on a linear equations task using text mining and qualitative techniques

Filip Moons[1,2] | Alexander Holvoet[3] | Katrin Klingbeil[4] | Ellen Vandervieren[2]

[1]Freudenthal Institute, Utrecht University, Utrecht, The Netherlands

[2]Antwerp School of Education, University of Antwerp, Antwerp, Belgium

[3]Faculty of Economics and Business, KU Leuven, Leuven, Belgium

[4]Didaktik der Mathematik (Fakultät für Mathematik), Universität Duisburg-Essen, Essen, Germany

Correspondence
Filip Moons, Freudenthal Institute, Utrecht University, Princetonplein 5, PO Box 85170, Utrecht 3508 AD, The Netherlands.
Email: f.moons@uu.nl

Funding information
Fonds Wetenschappelijk Onderzoek, Grant/Award Number: 1S95920N

Abstract: In this crossover experiment, we investigated the impact of a statement bank, enabling the reuse of previously written feedback (SA condition), on 45 math teachers' feedback for 60 completed linear equation tests, compared to traditional pen-and-paper feedback (PP condition). In the SA condition, teachers were encouraged to use atomic feedback, a set of formulation requirements that makes feedback items significantly more reusable. A previous study found that significantly more feedback was written in the SA condition but did not investigate the content of the feedback. To address this gap, we employed a novel approach of combining text mining with qualitative methods. Results indicate similar wording and sentiments in both conditions. However, SA feedback was more elaborate yet general, focusing on major and minor strengths and deficits, while PP feedback was shorter but more concrete, emphasising main issues. Despite low feedback quality in both conditions, the statement bank led to less effective diagnostic activities, implying that teachers' careless use of statement banks, although convenient, might lead to lower-quality feedback.

KEYWORDS

atomic, feedback, handwritten tasks, reusable feedback, semi-automated assessment, statement banks

Practitioner notes

What is already known about this topic

- High-quality feedback should strike a balance between the volume and focus on the main issues, as more feedback does not necessarily equate to better feedback. Feedback should analyse a student's solution whenever possible: interpreting mistakes and communicating that interpretation as feedback.
- Text mining identifies meaningful patterns and new insights in text using computer algorithms.
- When teachers can reuse already given feedback using a software tool (statement bank), they tend to write more feedback instead of saving time.

What this paper adds

- Feedback is compared when teachers could use a tool to reuse already given feedback (referred to as 'statement banks') versus a scenario without such a tool. Both approaches observed similar word frequencies, sentiments and amounts of erroneous, descriptive and corrective feedback. However, feedback with a statement bank tended to be more elaborate yet less specific to individual student solutions. In contrast, feedback without the tool was shorter but more concrete, focusing on main issues. Overall, the tool for reusing feedback directed teachers towards less effective diagnostic activities.
- The paper introduces a novel methodological approach by combining text mining with qualitative techniques in educational research. While text mining provides an overall understanding of differences and similarities in feedback approaches, qualitative methods are essential for in-depth analysis of content characteristics and feedback quality.

Implications for practice and/or policy

- Statement banks can support teachers by giving more feedback, but in order to improve feedback quality, further measures are necessary (eg, improving pedagogical content knowledge).
- Teachers may not confuse handiness with quality: statement banks can help, but when used carelessly, teachers tend to describe and correct students' work instead of analysing underlying (mis-)conceptions using it. Continued attention to feedback quality remains necessary when using such tools.

INTRODUCTION

Feedback has been recognised as a crucial component in learning processes (Hattie & Timperley, 2007; Shute, 2008). While many studies in educational technology have considered the effect of the modes of delivery of feedback (eg, Gleaves & Walker, 2013; Ryan et al., 2019) and the effect of immediate versus delayed feedback (eg, Candel et al., 2020; Lefevre & Cox, 2017), in this paper, we focus on the role of technology as it helps teachers to write feedback on a mathematics task. More specifically, we compare the written feedback reports composed by teachers under two conditions: the semi-automated condition (SA) in which they could use software to reuse previously written feedback items working like a statement bank (Denton & McIlroy, 2018; Moons et al., 2022) and the paper-and-pencil condition (PP) in which teachers could not reuse feedback, resembling regular feedback on a paper-and-pencil task (Chang et al., 2012), but instead of being handwritten, it is being typed. We will abbreviate the conditions as **SA** and **PP** in the rest of the paper and also refer to them as feedback approaches.

Ideally, written feedback reports should strike a balance between the volume and focus on the main issues as more feedback does not necessarily mean better feedback (Glover & Brown, 2006). Indeed, Evans (2013) indicates that feedback should not be so specific and detailed that students do not have to think for themselves anymore. Chiles (2021) calls this balance the 'goldilocks principle': feedback should be concise and accurate since 'too much feedback can be overwhelming for students and lead them to disengage with it.' It seems to be best to link feedback directly to overarching learning intentions and break it into small, achievable steps. As such, feedback should be more than solely corrective: it should indicate the what, how and why of problems in the students' work (Gibbs & Simpson, 2005), address misconceptions (Yang & Lu, 2021) and identify actions the student can take to improve (Sadler, 2010).

Providing feedback may be tedious and time-consuming: 49% of the teachers in the European Union and 53% of all British teachers complain about having too much assessment work (Education, Audiovisual and Culture Executive Agency, 2021; Gibson et al., 2015). One of the well-known coping mechanisms to overcome this workload is shortening feedback (Price et al., 2010) or using rubrics or marking sheets (Denton & Rowe, 2015).

Atomic feedback

In this research project, we take a slightly different approach to provide written feedback to handwritten mathematics tasks more efficiently. After all, handwritten tasks remain important to train higher-order thinking skills and genuine problem-solving in mathematics education as students can express themselves more freely (Bokhove & Drijvers, 2010; Hoogland & Tout, 2018). Therefore, we propose a semi-automated (SA) approach: handwritten solutions are scanned, then teachers write feedback items and the computer saves them so they can easily be reused when other students make similar mistakes (Moons et al., 2022).

How to write feedback that can easily be reused for other students? Long pieces of classic feedback are often too targeted to a specific student (Winstone et al., 2017). Hence, we suggest atomic feedback (see Figure 1): a collection of form requirements for written feedback that have been shown to make feedback significantly more reusable (Moons et al., 2022). To write an atomic feedback item, teachers must:

1. identify independent errors,
2. write small feedback items for each error separately or
3. if an error reflects a structural mistake/misconception (Gusukuma et al., 2018; Movshovitz-Hadar et al., 1987; Schnepper & McCoy, 2013), create two feedback items:
   a. one item containing feedback on the misconception in general and
   b. one or more sub-items addressing specific mistakes.

Atomic items ultimately form a point-by-point list covering only items relevant to a student's solution. The list can be hierarchical in order to *cluster* items that belong together. Clustering ensures that feedback items can be written as atomically as possible. It prevents teachers from writing overly specific items because it provides an orderly way to present related feedback to students (eg, through thematic clustering or a visual presentation of both general and specific feedback on the same error).

A comparison of classic (PP condition) and atomic feedback (SA condition) is presented in Figure 1. This comprehensive example demonstrates that classic feedback reports can

Student's solution

Manipulate the formula: $A = 2\pi rh + 2\pi r^2$ to h

$$\frac{A}{2\cdot\pi\cdot r} = h + 2\pi r^2$$

$$\frac{A - 2\pi r^2}{2\cdot\pi\cdot r} = h$$

Classic feedback

Mind the fact that the dominant operation on the right-hand side of the equation is an addition! The division of the left-hand side by $2\pi r$ is, therefore, not helpful. Moreover, $2\pi r$ is a common factor of the right-hand side, but the sum wasn't completely divided by it (second addend not divided). Although your final answer is correct, the way it is written makes it look like a coincidence. Going from the first to the second step, you would normally subtract $2\pi r^2$ from both sides, meaning that it shouldn't be placed directly in the numerator, as you should make the denominators the same.

Atomic feedback

- First step
  - Dominant operation on the right side is an addition!
    - ⋆ Division of left-hand side is not helpful
    - ⋆ $2\pi r$ is a common factor of the right side, but:
      - · sum wasn't completely divided by it
      - · the second addend was not divided
- Second-step
  - Your final answer is correct, but:
    - ⋆ It looks like a coincidence.
    - ⋆ You should subtract $2\pi r^2$ from both sides.
    - ⋆ Mistake with making the denominators the same!
      - · $2\pi r^2$ shouldn't be directly in the numerator.

FIGURE 1 A comparison between classic (PP) and atomic (SA) feedback.

be rephrased as atomic. This paper compares all feedback reports from the PP and SA conditions. In the SA condition, teachers were encouraged to write atomic feedback, but it is important to mention that all SA feedback reports will be considered; and not all of them adhere to the definition of atomic feedback (see Moons et al., 2022).

Research aims

In Moons et al. (2022), it was demonstrated that feedback items meeting the atomic feedback requirements were significantly more reused than non-atomic items ($p < 0.001$, odds ratio: 2.6). This finding suggests that writing feedback items atomically enhances their reusability. Additionally, no significant differences in time investment were observed between the PP and SA conditions. However, teachers participating in the SA condition wrote significantly more feedback characters compared to the PP condition ($p = 0.02$, Cohen's $d = 0.41$), approaching a medium effect size. Despite these findings, an important research question remains unanswered:

> [RQ] What similarities and differences do the SA and PP feedback approaches have regarding form, content and quality?

To address this question, we will employ text mining techniques (Ferreira-Mello et al., 2019) and conduct a qualitative analysis (MacLure, 2013) on the feedback from both conditions. The qualitative analysis will investigate content characteristics and quality by coding the feedback reports. Through text mining, we will analyse word frequencies, sentiment, bigrams and word correlations to compare the form and content characteristics of the two feedback approaches.

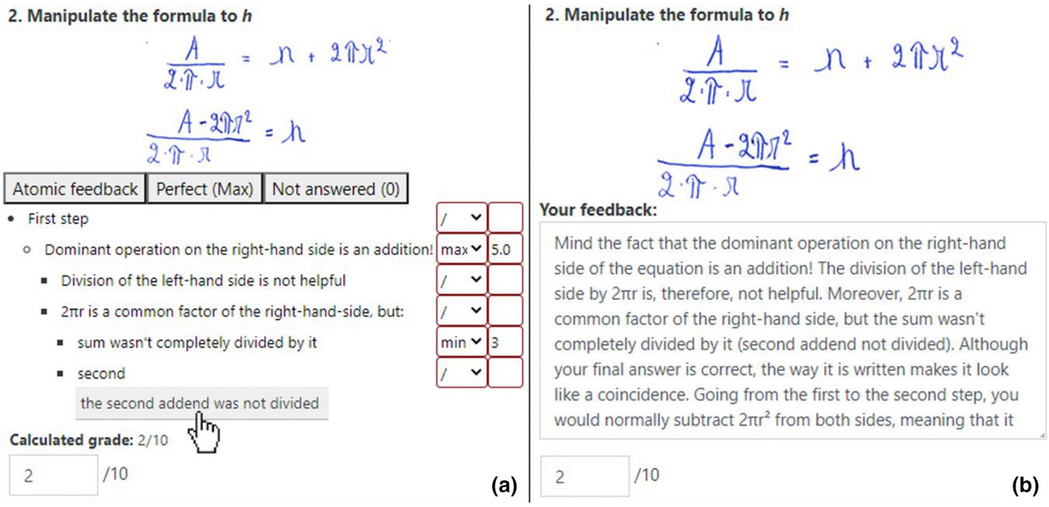
By addressing this research question, we aim to achieve two broader objectives. Firstly, we seek to gain a deeper understanding of how the utilisation of a statement bank, specifically reusing feedback, influences the characteristics of the resulting written feedback. This investigation will shed light on the impact of utilising pre-existing feedback statements on the form, content and quality of the feedback provided. Secondly, we aim to explore the methodological approach of combining text mining and qualitative analysis to compare feedback. While text mining has been extensively used in higher education to analyse student course feedback (eg, Grönberg et al., 2021), and qualitative approaches have been employed in combination (Hujala et al., 2020), the integration of these methodologies to compare feedback represents a relatively novel and promising application.

MATERIALS AND METHODS

Materials

Semi-automated assessment tool for SA and text box for PP

For the SA condition, a self-developed plugin in Moodle was used. While providing feedback on students' solutions, teachers always had three options in this condition: formulating atomic feedback, indicating that a solution was perfect or indicating that the question was not answered (Figure 2a). They were able to use keyboard shortcuts to create a hierarchical list of feedback items. When a teacher typed something, the system searched the feedback items that had already been entered to detect possible matches for auto-completion (Figure 2a). The system searched only within the feedback items that the teacher had already entered for that particular question. In the PP condition, the teachers received only a text box to type feedback (see Figure 2b), with no possibility of reusing feedback. In both conditions, teachers were also asked to give each solution a score out of 10.



2. Manipulate the formula to *h*

$$\frac{A}{2\cdot\pi\cdot r} = h + 2\pi r^2$$

$$\frac{A-2\pi r^2}{2\cdot\pi\cdot r} = h$$

- First step /
  - Dominant operation on the right-hand side is an addition! max 5.0
    - Division of the left-hand side is not helpful /
    - 2πr is a common factor of the right-hand-side, but: /
      - sum wasn't completely divided by it min 3
      - second /

the second addend was not divided

Calculated grade: 2/10

2 /10

2. Manipulate the formula to *h*

$$\frac{A}{2\cdot\pi\cdot r} = h + 2\pi r^2$$

$$\frac{A-2\pi r^2}{2\cdot\pi\cdot r} = h$$

Your feedback:

Mind the fact that the dominant operation on the right-hand side of the equation is an addition! The division of the left-hand side by 2πr is, therefore, not helpful. Moreover, 2πr is a common factor of the right-hand side, but the sum wasn't completely divided by it (second addend not divided). Although your final answer is correct, the way it is written makes it look like a coincidence. Going from the first to the second step, you would normally subtract $2\pi r^2$ from both sides, meaning that it

2 /10

FIGURE 2 Screens of the tool in the SA condition (a) and PP condition (b).

Test on linear equations

We developed a test on linear equations in cooperation with a ninth-grade math teacher for this study. The test consisted of three items: (1) solving an equation, (2) manipulating a formula (see Figures 1 and 2) and (3) a modelling question consisting of a word problem (see Figure 8). The three items were combined to form a traditional test on linear equations. Solutions of 60 ninth-grade students (14–15 years old) from one secondary school in Flanders (Belgium) were used in this study. The test and solution key can be found in Appendix A in Moons et al. (2022).

Methods

Teacher participants

A total of 45 secondary mathematics teachers from Flanders with at least 3 years of working experience volunteered to participate in the study (28 women, 17 men). They were sampled using announcements in math teaching magazines. The average age of the participating teachers was 40.2 years (*SD* = 10.3). The first time we organised the experiment with nine math teachers, we noticed several methodological imperfections in our study design. Therefore, this first attempt was used as a pilot study (*n* = 9) to refine the actual study (*n* = 36). A description of the pilot study and the adaptations to the actual study can be found in Appendix B in Moons et al. (2022).

Study design

The study was set up as a randomised crossover study (Bose & Dey, 2009) with two conditions: SA and PP. During a full working day, the teachers started in one condition in the morning and swapped to the other condition in the afternoon. The experiment was executed in the summer of 2020. Unfortunately, due to COVID-19 measures, only nine participants at a time were allowed in the computer laboratory. Therefore, the experiment was repeated seven times.

Each teacher gave feedback to all 60 solutions of the linear equation test, with a quasi-random selection of 30 solutions being assessed under the SA condition and the other 30 solutions under the PP condition. To mitigate order effects inherent in crossover experiments (Ratkowsky et al., 1993), half of the teachers started under the SA condition and the other half started under the PP condition. The day started with training for all participating teachers. The training focused on working with the SA tool and the PP text field in Moodle and on how to formulate atomic feedback. The linear equation test was never mentioned during the training, and a geometry task obtained from other students was used instead as a demonstration. At the end of the training, teachers were asked to treat the students' solutions in the experiment in the same way that they would treat their own students. No training was provided for providing content-rich feedback and they were not informed about the research questions. Teachers had to be themselves above all.

The quasi-random selection of 30 solutions in each condition for each teacher ensured: (1) Comparability of the feedback between the conditions. Each solution was included in the SA condition of 18 teachers and the PP condition of the other 18 teachers, ensuring that both conditions comprised feedback to the same solutions an equal number of times. (2) To balance the conditions for each teacher, we ranked all 60 students' tests based on the grades of the pilot study and made three groups: high, moderate and low. Each condition contained 10 tests from each group. (3) The order in which the solutions were presented in each condition was random to avoid any bias caused by task familiarity or fatigue.

Data analysis procedures

Text mining

First, the provided feedback was explored using text mining techniques (Kwartler, 2017; Silge & Robinson, 2017). Text mining transforms unstructured text into a structured format to identify meaningful patterns and new insights using computer algorithms. It can be seen as a qualitative research method 'using quantitative techniques' (Yu et al., 2011).

A difficulty in applying text mining techniques is that many possible analyses can be employed. As this paper aims to compare the given feedback to the same mathematics tests in two conditions, we carefully applied techniques allowing us to find differences and similarities between these two feedback approaches. More specifically, we compared word frequencies, did a sentiment analysis and compared the Markov chains of bigrams and the pairwise correlations of both conditions. These techniques were inspired by the book of Silge and Robinson (2017). More advanced approaches, such as LDA topic modelling, were executed but did not provide meaningful insights for our research question and are, therefore, not reported. We deliberately left out any significance tests in the text mining part as these tests are often overpowered when analysing on the level of words, making the sample sizes too large (Faber & Fonseca, 2014), or the test is executed on outcomes of an analysis that requires cautious interpretation (such as sentiment analysis), further supporting our decision. All the analyses were done using R.

Since the teachers participating in the study provided feedback in Dutch, all analyses were conducted in this language. In the pre-processing data phase, we removed all Dutch frequently used words (like 'a', 'the' and 'of' in English) using a predefined lexicon (Benoit et al., 2021), a conventional first step in text mining analyses. In the final data analysis step, the results were automatically translated to English using the DeepLr package (Zumbach & Bauer, 2021) to make the results interpretable for an international audience; hereby losing some specific language characteristics of Dutch (abbreviations and concatenations).

Qualitative analysis

For the qualitative exploration of feedback, Busch et al. (2015a, 2015b) developed a codebook to assess teachers' diagnostic competencies that take into account the quality features of feedback described in the introduction like the number of deficits/strengths (Chiles, 2021; Evans, 2013), focusing on misconceptions (Yang & Lu, 2021), diagnostic activity (Gibbs & Simpson, 2005) and giving hints for improvement (Sadler, 2010). The codebook of Busch et al. (2015a) was especially suitable as it was directed to mathematics tasks. Moreover, it does not include categories related to standard classroom settings, in which teachers can direct more personal messages to students; aligning with the study design with participating teachers not knowing the students. Only minor changes were made to the codebook for this study: the categories were defined more rigorously to achieve higher interrater reliability (see Table 2), and we made a distinction between categorisable and not further categorisable feedback.

Two authors of this paper acted as independent raters for the qualitative analysis who coded blindly, meaning the coders could not see each other's codes in the process. The level of analysis was the full feedback report of the teacher on a student's solution to the word problem (see Figure 8). Four iterations were necessary to arrive at high interrater reliability; the final Cohen's kappa coefficients can be found in Table 2. A random selection of about 100 feedback reports from the pilot study was used in each iteration. At the end of every iteration, the differences in coding were thoroughly discussed and some definitions of categories were refined (codebook described below). When coding the actual study data, feedback reports were coded student by student, and all feedback was checked for correctness by placing the student's solution next to it. Each researcher coded the teachers' feedback from 30 of the 60 students.

The codebook consists of *categorisable feedback* and *not further categorisable feedback*. The latter consists of *erroneous feedback*, *incomprehensible feedback* or *only addressing a solution was perfect, totally wrong or left blank*. A feedback report can only have one of these codes, meaning that erroneous or incomprehensible feedback reports were deliberately excluded from further classification.

The remaining categorisable feedback is to be coded into five sub-categories:

Concreteness judges how 'specific' the feedback is. For example, feedback containing only *'Order of operations!*' is *general*, while *'x = 14.4? This can not be the number of answers!*' points to *concrete* feedback. As a guideline, the two independent raters used the following question to decide between general and concrete: 'Can this feedback without any adjustment be applied to another student who did something else?' If yes, the feedback is classified as general; if not, the feedback is concrete. Concrete and general were mutually exclusive: as soon as something concrete was mentioned in the feedback, the whole report was characterised as concrete.

The **focus of the diagnosis** counts how many *deficits* and *strengths* the feedback addresses.

The **diagnostic activity** differentiates among *analysis*, *description* or *correction*: correction entails a teacher pointing to a mistake and giving the right solution (eg, '*amount of correct answers: 30-4-x*'). In contrast, description references a teacher addressing deficits without correction (eg, '*wrong equation'*). Finally, analysis signifies the teacher interpreting the student's mistakes and reporting that interpretation as feedback (eg, '*You swapped answer and number of correct points in the choice of the unknown, 5x is the number of points gained with the correct answers, x the number of correct answers'*). Merely noticing an error is seen as description, merely giving the right solution as correction. To ensure interrater reliability, a feedback report could only be coded into one diagnostic activity. When several diagnostic activities were identified in a feedback report, analysis was always preferred over correction, and correction always over description.

The **quality features of the diagnosis** contain four aspects, which were not mutually exclusive:

- *Explanation for deficits available*: the feedback contains a statement explaining why something is wrong in the solution (eg, *'Why subtraction? Points should be added!*'). Explanation as a quality feature should not be confused with the diagnostic activity analysis: it can also be a more general expression of a mistake without interpretation at the student level.
- *Gives hints for improvement*: the feedback contains statements indicating how the solution should be improved in a possible future review (eg, *'Keep points and number of questions well apart!*'). A hint cannot contain the correct solution since the need for a future overhaul is then eliminated.
- *Notes that parts are missing in the student's solution*: the feedback explicitly refers to something that should have been in the solution (eg, *'Write down the choice of the unknown*').
- *Points to misconceptions*: the feedback contains statements to known misconceptions in mathematics education (Movshovitz-Hadar et al., 1987) or misunderstandings in the student's reasoning (eg, *'You fail to see that your solution is impossible since there are more answers correct than questions'*).

RESULTS AND DISCUSSION

Text mining

Comparing word frequencies

A common first step in text mining is to compare word frequencies (Silge & Robinson, 2017). The frequency of a word is the proportion of the number of times a word occurs out of the total word count. Figure 3 gives a scatter plot of the used words in both feedback conditions. Words close to the identity line have similar relative frequencies in both conditions. It is apparent from this plot that most words scatter around this line, meaning that the majority of the words appear in both feedback approaches with a similar relative frequency. For example, 'attention' and 'both' appeared almost equally frequently in SA and PP. The observation that most words appeared in both feedback approaches with an almost equal relative frequency was confirmed by calculating Pearson's correlation coefficient of the word frequencies in both conditions. It returned a high, positive correlation of $r(928) = 0.89$ with 95% CI [0.87, 0.90].

Words far from the identity line are, proportionally speaking, found more in one condition than the other. For example, 'super' and 'beautiful' were found more in PP feedback, while 'perfect' was found more in SA feedback. A likely reason is the default presence of a 'Perfect' button that could be used for correct solutions in the SA condition (see Figure 2a). In the PP condition, teachers always had to write something themselves, and it seems they

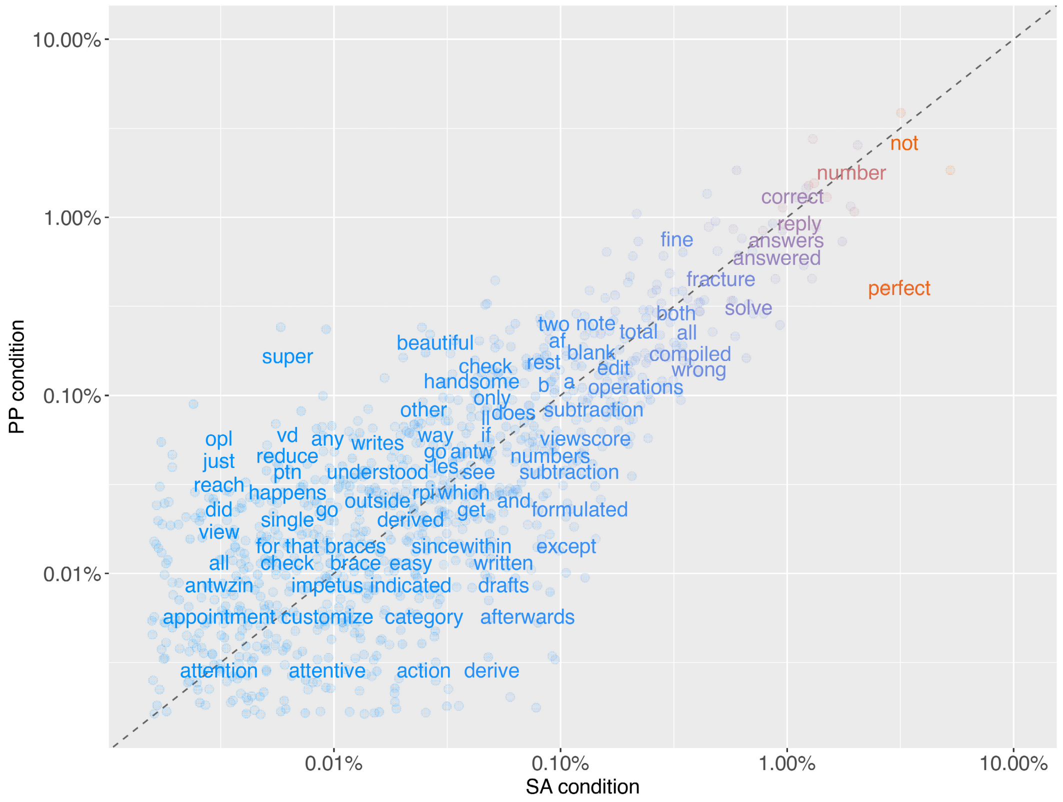


FIGURE 3 Comparing the word frequencies of SA and PP feedback.

naturally chose a more diverse range of encouraging words. Also notable is the increased presence of many abbreviations in the PP condition, which DeepL understandably failed to translate, like ‘opl’ (Dutch abbreviation for ‘solution’), ‘vd’ (= ‘of the’), ‘ptn’ (= ‘points’) or ‘antw’ (= ‘answer’). Teachers shortening feedback is one of the well-known coping mechanisms described in the literature (Price et al., 2010) to overcome the workload stemming from giving feedback. The semi-automated system seems to discourage teachers from using abbreviations all too often, as they can reuse feedback items.

Sentiment analysis based on given scores

As we will analyse the words' sentiment, an insightful step is to look at the distribution of words spent on perfect, good, moderate and bad students' solutions in both conditions, based on teachers' given scores out of 10. We used an arbitrary division in scores to categorise all the words. Scores less than 5 were classified as belonging to bad solutions, those corresponding to scores between 5 and 7 were classified as moderate, and those corresponding to scores greater or equal to 7 but lower than 10, were classified as good. Perfect solutions had 10 out of 10 points. The number of words in each solution type was counted and turned into percentages, leading to the distributions in Figure 4.

The distribution of both feedback approaches looks essentially the same: proportionally, an almost equal amount of words is spent on bad solutions. SA feedback features slightly more feedback on moderate answers than PP feedback, which has proportionally more words coupled with good and perfect answers.

Although the word distribution in Figure 4 is some kind of sentiment analysis, in text mining, analysing the sentiment of a text is often done by using a pre-existing lexicon that assigns a polarity score to individual words (like ‘beautiful’ = 1, ‘incorrect’ = −1); subsequently, the sentiment of the whole text can be determined by taking the mean (Silge & Robinson, 2017). For the Dutch language, the PATTERN lexicon (De Smedt & Daelemans, 2012) gives words a polarity score ranging from −1 (very negative), 0 (neutral) to 1 (very positive). For example, the following PP feedback has a mean polarity score of −0.65 (negative to very negative):

> Wrong choice of the unknown. A solution is found by guessing. However, guess cannot be right, you cannot give 95 wrong answers to 30 questions. No check.

In contrast, the SA feedback below received a mean polarity score of 0.72 (positive to very positive):

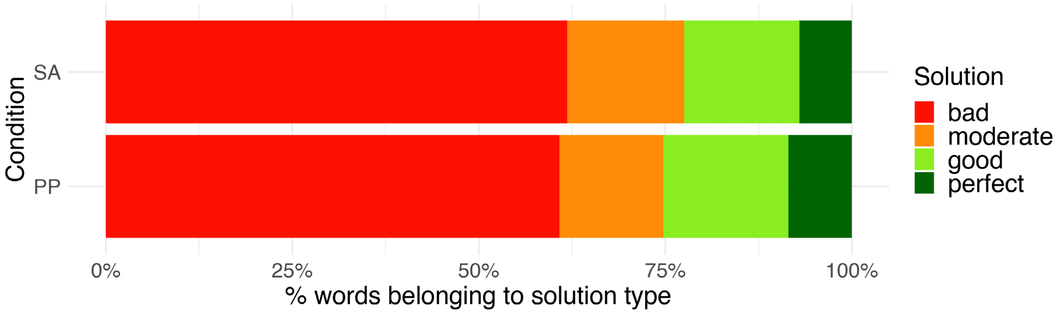


FIGURE 4 Comparing the distribution of words (above: SA/below: PP) spent on different solution types (red: bad/orange: moderate/light green: good/dark green: perfect).

TABLE 1 Mean polarity score and standard deviation overall and for each solution type.

		Overall without perfect	Solution type			
	Overall		Bad	Moderate	Good	Perfect
SA condition	0.384 ± 0.592	−0.056 ± 0.410	−0.097 ± 0.382	−0.076 ± 0.385	0.096 ± 0.438	0.938 ± 0.196
PP condition	0.239 ± 0.451	0.022 ± 0.400	−0.013 ± 0.366	−0.038 ± 0.354	0.160 ± 0.467	0.616 ± 0.253

- Good choice of the unknown
- Good representation of the second unknown
- The equation that you have set up is perfect.
- The solution of the equation is perfect.
- You did not formulate an answer.

All feedback reports were analysed with this lexicon by taking the mean polarity score of all the words in the report. Next, we looked at the overall mean, the overall mean without the perfect solution type and the mean for each solution type. The results can be found in Table 1.

The sentiment analysis suggests that overall, the feedback in the SA and PP conditions has a neutral tone when perfect solutions are not considered. Moreover, the feedback tones are relatively equal when comparing the solution types in both conditions.

Lastly, some caution is necessary when interpreting this sentiment analysis. For example, the reason why we considered 'overall without perfect' as a separate column in Table 1 is because including the perfect solutions induces a bias in favour of positive tones in the SA condition, as the button 'perfect' yielded feedback just saying 'perfect', with a polarity score of 1. The greater variety of appreciation words in the PP condition can sometimes include words or abbreviations not included in the sentiment lexicon; as such, the polarity score is sometimes estimated to be somewhat lower than 1, while the feedback reports are equally positive for these perfect solutions. Moreover, like in many sentiment analyses, the context was not taken into account; making statements like 'not good' having a polarity score of 0.6 as the 'not' is not seen as a word that reverses the polarity score; note, however, that the word usage of both conditions is almost equal (see previous paragraph), so the bias due to not including context is probably almost the same in both conditions.

Cluster analysis: Markov chains of bigrams and pairwise correlations

To increase the readability of the plots in this paragraph, we limit ourselves to the feedback given in question 2 of the linear equations test (see Figures 1 and 2) in both conditions. Figure 5 depicts the Markov chains of SA feedback (blue) and PP feedback (red). It visualises the pairs of consecutive words (= bigrams). As a cut-off, we have chosen a minimum of 10 co-occurrences. Although it represents a directed graph, we have omitted the arrows to increase readability.

Apparently, SA feedback features a denser linking structure between consecutive words. However, as reusing feedback is the main characteristic of this feedback, this was expected as some pairs will have been reused frequently, while PP feedback contains slight variations in word pairs. Nevertheless, some similar clusters arise in both feedback conditions. For example, noticing that double arrows should be used between the different intermediate steps was a cluster in both conditions. Interestingly, in the SA condition, the word 'notation' also appears in this cluster. Using titles as a way of clustering feedback is one of the characteristics of atomic feedback, of which 'notation' is a clear example. If we examine the other clusters, other structuring elements in SA are found: 'calculation rules', 'step 1', etc., which do not appear in PP. SA with atomic feedback seems to foster teachers to structure feedback using titles, a phenomenon that does not emerge in PP feedback.

Finally, pairwise correlations of the words in the same feedback reports were compared. Pairwise correlations differ from the bigrams in Figure 5 as they do not link words succeeding each other but connect words often appearing together in the same feedback report (not necessarily consecutive). As SA contains many reused words, a denser correlation network is again to be expected. To marginally mitigate this bias in favour of SA, Spearman's rank correlation coefficients were used, which compare ranks instead of frequencies. In Figures 6 and 7, the correlation networks of the feedback on question 2 can be found. The different

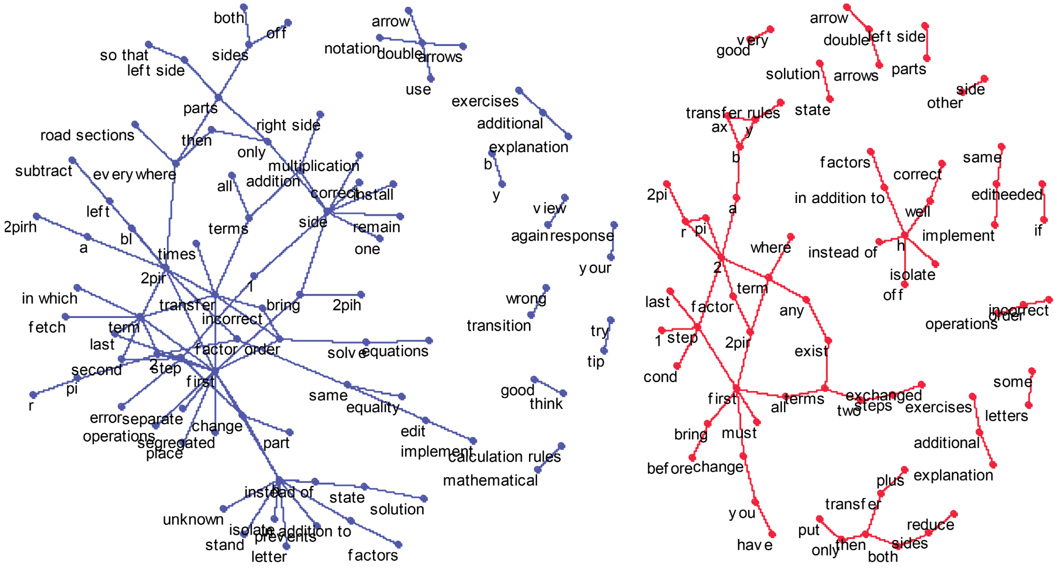


FIGURE 5 Markov chains of bigrams for SA (blue) and PP (red) on question 2.

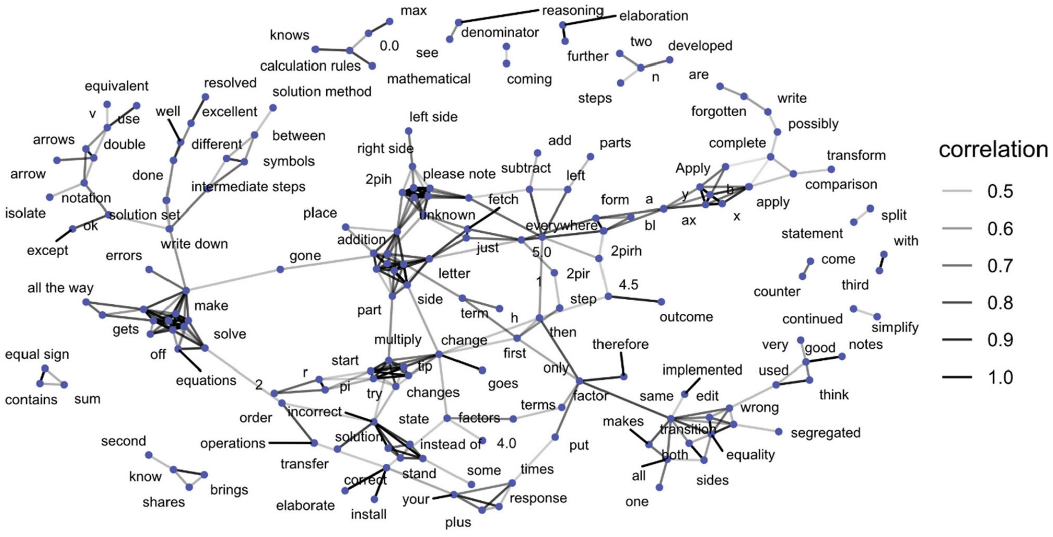


FIGURE 6 Correlation network of SA feedback given to question 2.

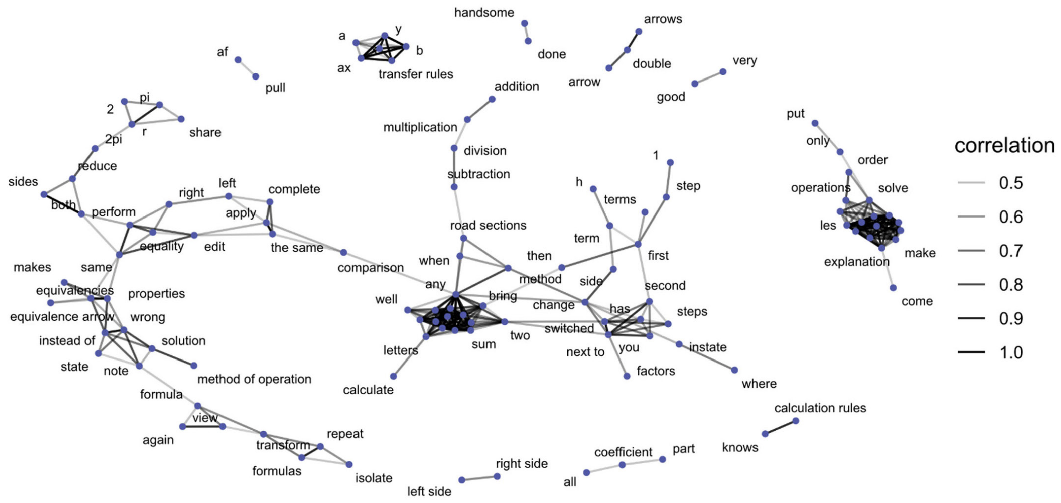


FIGURE 7 Correlation network of PP feedback given to question 2.

clusters refer to the same student's mistakes. Although the bias in favour of SA should be remembered, the difference in the largest cluster suggests that PP limits itself more often to short statements like ‘reduce both sides’ and ‘isolate *h*’. In contrast, SA feedback seems to provide more information.

Qualitative analysis

Table 2 shows the results of the qualitative analysis. All percentages represent the proportion of feedback reports out of all feedback reports in that condition. The number of deficits and strengths between SA and PP were compared using a Mann–Whitney *U* test. All other reported *p*-values stem from two-sample *z*-tests for proportions, comparing for every category if the proportion of feedback reports differs between SA and PP. The Pearson correlation coefficient ρ_{teacher} correlates the number of times a characteristic was chosen in both conditions for each teacher (or the number of deficits/strengths addressed). A strong ρ_{teacher} (>0.7) for a characteristic indicates that the prevalence of the characteristic was consistent for teachers' feedback reports across both conditions. ρ_{student} reports the correlation on the level of the student solution.

Observed differences between SA and PP feedback

Overall, the results indicate SA feedback is less tailored to the student's solution than PP feedback: the SA reports are almost equally likely to be labelled as general or concrete (39.65% and 38.99%), whereas PP condition yielded much more concrete feedback (47.52%). However, SA seems more detailed: significantly more deficits and strengths were addressed in this condition; in contrast, PP seems more centred on the main issues in the solution. A frequently observed use of SA, which is general and can address different deficits and strengths, is using it as a sort of checklist, as the feedback below illustrates:

TABLE 2 Results of the qualitative analysis.

	κ	SA ($n=913$)	PP ($n=947$)	p-value	ρ_{teacher}	ρ_{student}
Categorisable feedback		*78.64%*	*78.04%*	0.631		
Concreteness						
General***	0.82	39.65%	30.52%	<0.001	0.47	0.74
Concrete***	0.75	38.99%	47.52%	<0.001	0.51	0.83
Focus of the feedback						
Number of deficits**	0.89[a]	1.73 ± 1.28	1.57 ± 1.08	0.003	0.48	0.87
Number of strengths*	0.89[a]	0.79 ± 1.05	0.65 ± 0.84	0.038	0.60	0.51
Diagnostic activity						
Analysis*	0.88	5.15%	7.60%	0.038	0.22	0.51
Correction	0.87	16.21%	16.79%	0.726	0.68	0.64
Description	0.84	56.63%	52.80%	0.103	0.58	0.68
Quality features						
Explanation for deficits available*	0.58	9.42%	12.57%	0.030	0.17	0.85
Gives hints for improvement	1.00	19.72%	23.23%	0.072	0.68	0.68
Notes parts that are missing	0.49	15.55%	14.36%	0.478	0.65	0.72
Points to misconceptions	0.82	4.93%	5.07%	0.889	0.31	0.85
Not further categorisable feedback		*21.36%*	*21.96%*	0.538		
Erroneous feedback	0.79	4.60%	4.96%	0.711	−0.02	0.82
Incomprehensible feedback**	1.00	1.20%	0.11%	0.003	−0.04	0.25
Only addresses solution is entirely correct*	–[b]	11.17%	14.68%	0.024	0.19	0.68
Only addresses question is left blank*	–[b]	3.40%	1.90%	0.044	−0.20	0.90
Only addresses solution is entirely wrong	–[b]	0.99%	0.32%	0.072	0.52	0.30

[a] Intra-class correlation coefficient.

[b] Automatically coded.

*$p<0.05$; **$p<0.01$; ***$p<0.001$.

- Choosing the unknown
  - You are confusing the distinction between the number of questions and points received.
- Setting up and solving the equation
  - You did not include the unanswered questions
  - Your equation is simpler than the equation to solve the question, but the solution is right.

Concerning the diagnostic activity, we see that there are significantly more feedback reports analysing where it went wrong with a student's solution in PP, from which an example is given:

> Please try again with x being the number of correct answers. Indeed, you know that for 26 questions, he got points. So you express the number of unanswered questions in terms of x. When setting up the equation, you noted 120 instead of 102. You have to take into account the 5 points per correct question.

SA feedback reports tended to use more description and correction as diagnostic activity. Moreover, notice the low correlation (0.22) of teachers concerning analysis: teachers who analysed some solutions in one condition did not necessarily use that diagnostic activity as often in the other, suggesting that the SA system discourages teachers from providing feedback reports that analyse student's mistakes. One possible explanation is that teachers intuitively use SA too much as a checklist, preventing them from interpreting the interplay of intermediate steps the students took.

The significantly lower number of explanations given in the SA condition (the only significant difference in quality features) can be seen in the same vein. PP often addresses a particular mistake, on which the teacher sometimes gives an extra word of explanation. SA more often addresses all the mistakes in a solution, but treats these more superficially, without much extra information.

If we look at the differences in the '*not further categorisable feedback*', we see that SA feedback is more often incomprehensible. However, it concerns only 1.2% of all feedback reports. Almost all of these stem from the same teacher who consistently used the hierarchical list of atomic feedback items in a confusing way by using opposite appreciation words in the parent items and child items, for example:

- Good formula
- Bad formula

It is striking that the readily available buttons 'Perfect' and 'No answer' in the SA condition had opposite effects. Just noticing a solution was 'good' or 'cleverly done' happened significantly more often without a button (!) in PP. This is consistent with the earlier observation that PP contains fewer deficits/strengths, but seems somewhat contradictory to the text mining analysis where 'perfect' was a more prevalent word in the SA condition. In the SA condition, however, feedback reports for perfect solutions sometimes contained a complete list of all things that went well or noticed something that still could be added, such as a check if the obtained solution could be correct. In contrast, teachers did not hesitate to use the 'No answer' button in the SA condition when a solution was missing, while in the PP condition, they tended to give some hints on how to start solving the question, wrote some encouraging statements or asked the student what the underlying problem was (eg, time issue or not enough understanding):

> The question was left blank. Did you have enough time?

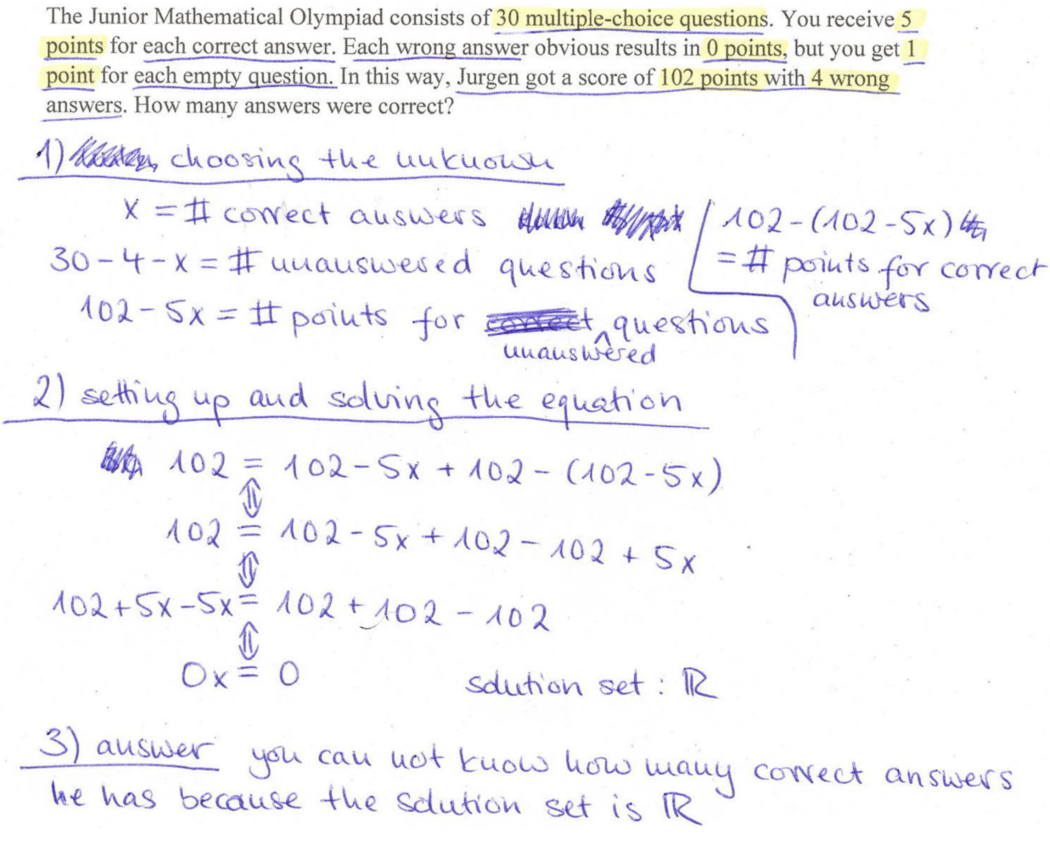
Observed feedback quality

Both the text mining as well as the qualitative analysis allow us to evaluate the overall quality of both feedback conditions.

From the text mining analysis, it follows that abbreviations are more common in the PP condition compared to the SA condition. Avoiding abbreviations is often presented in guidelines for providing feedback as it makes texts more easily interpretable (Wager & Wager, 1985). Moreover, the SA condition contained many more structuring elements, which partly follows from the definition of atomic feedback. However, it remains to be investigated if abbreviations and structuring elements really affect students' understanding of the given feedback.

From the qualitative analysis, a disappointing outcome in Table 2 is that almost 1 of 20 feedback reports is erroneous in both conditions. In other words, when feedback is handed out in an average classroom of compulsory education that, according to OECD (2012), consists of 21 students; one student will receive incorrect feedback. These errors might be due to routine like teachers noting a common mistake that did not occur (they probably interpreted the solution too quickly) or saying the solution is perfect, while the intermediate steps contain arithmetic errors. Nevertheless, more severe erroneous feedback was noticed, too: sometimes students choosing an alternative (but correct) solution path for the question and not arriving at the correct answer only received negative feedback in which their solution method was also (falsely) rejected. Luckily, erroneous feedback pops up coincidentally, as the within-teachers correlation of −0.02 shows it is not a consistent characteristic of teachers. In contrast, some solutions lead to erroneous feedback more often in both conditions ($\rho_{\text{student}} = 0.82$).

From the introduction, we know that feedback should be more than solely corrective (Gibbs & Simpson, 2005). However, *corrective* and *descriptive* feedback are the most popular diagnostic activities in both conditions. More than half of the reports are descriptive, only noticing mistakes without any action the student can take to improve. Indication about the what, how and why of problems in the students' solution (Gibbs & Simpson, 2005) aligns with *analysis* as a diagnostic activity, with a worryingly low proportion in Table 2. Part of the explanation is that some simple mistakes are not analysable, such as a small calculation error due to the absent-mindedness of the student: in such cases, a teacher can only notice the error. Consequently, the feedback would be coded as 'description' or 'correction'. Nevertheless, analysis is not only lacking in these cases but also when the student solution is well analysable like the one in Figure 8. In this solution, the student makes a well-documented circular argument (Reusser & Stebler, 1997). By using the same given information twice, the student is left with an equation leading to an infinite number of solutions. Only 5 of the 36 teachers (14%) responded to this fallacy with feedback that analysed



The Junior Mathematical Olympiad consists of 30 multiple-choice questions. You receive 5 points for each correct answer. Each wrong answer obvious results in 0 points, but you get 1 point for each empty question. In this way, Jurgen got a score of 102 points with 4 wrong answers. How many answers were correct?

1) choosing the unknown

$x = \#$ correct answers

$30 - 4 - x = \#$ unanswered questions

$102 - 5x = \#$ points for ~~correct~~ unanswered questions

$102 - (102 - 5x) = \#$ points for correct answers

2) setting up and solving the equation

$$102 = 102 - 5x + 102 - (102 - 5x)$$
$$\Updownarrow$$
$$102 = 102 - 5x + 102 - 102 + 5x$$
$$\Updownarrow$$
$$102 + 5x - 5x = 102 + 102 - 102$$
$$\Updownarrow$$
$$0x = 0$$

solution set : $\mathbb{R}$

3) answer you can not know how many correct answers he has because the solution set is $\mathbb{R}$

FIGURE 8 An 'analysable' student's solution to the word problem.

it; the other teachers gave descriptive feedback just noticing simple facts (eg, '*equation is wrong*') or corrective feedback. Of those five teachers, just one analysed this solution in the SA condition. SA feedback seems to engage teachers less in giving feedback analysing the student's solution, compromising overall feedback quality. However, this example shows that factors other than the condition, such as pedagogical content knowledge (Depaepe et al., 2013) or awareness of feedback quality criteria, seem to play an essential role in the feedback quality as well.

As mentioned in the introduction, shorter feedback is not necessarily worse for students (Chiles, 2021; Evans, 2013; Glover & Brown, 2006). However, the overall low number of deficits and strengths in Table 2 in both conditions gives pause for thought. While coding, we noticed a lot of 'incomplete' feedback reports, like the following PP feedback given to the solution in Figure 8:

> The first two lines are enough in your choice of the unknown. Equation is not set up correctly.

One may wonder what students can learn from this feedback: they probably already figured out that the equation was incorrect as infinitely many correct answers seem a highly unlikely outcome. And what about the other things they wrote? The phenomenon was seen many times while coding: feedback addressing the deficits at the start of the student's solution; next, it concludes: 'as a result, the rest of your solution is also wrong', not saying anything about deficits and strengths in the rest of the student's solution. Some feedback seemed just too short to be meaningful to a student. This phenomenon occurred more in PP feedback as the number of addressed deficits was significantly lower.

CONCLUSIONS

To wrap up this paper, we collected all our observations in Table 3. With this explorative study comparing SA and PP feedback using text mining and qualitative techniques, we identified some essential characteristics of both feedback approaches.

First, we discovered similarities in both approaches. From the text mining analysis, we distilled that the word usage and frequency are equal in both conditions (S1), equal amounts of feedback were spent on bad, moderate and good solutions (S2) and feedback reports featured predominantly the same sentiments (S3). From the qualitative analysis, we know corrective and descriptive feedback appeared equally often as diagnostic activity in both conditions (S4), as well as giving hints, pointing at misconceptions and parts that are missing (S5). Writing erroneous feedback was also independent of the condition: it appeared almost equally often in both conditions (S6).

Many differences can be attributed to the observation of Price et al. (2010) that teachers often shorten feedback to reduce the workload of giving it. The need for this coping mechanism was profoundly reduced in the SA condition where teachers could reuse their feedback items: it contains more feedback (D1; Moons et al., 2022), fewer abbreviations (D2), addresses more mistakes and strengths (D5) and is more elaborate in describing mistakes (D9). However, this apparent comprehensiveness of SA feedback does not greatly improve the content quality: SA is often used as a checklist of all things that could go well/wrong, leading to more general feedback (D4). In contrast, PP feedback seems to be more focused on the main issues (D5), is more concrete and tailored to the student's solution (D4) and gives more short explanations of the observed deficits (D7, D9). More importantly, PP included more reports that analyse the student's solution (D6). When solutions were perfect, PP feedback used various appreciation words without much more, while SA often had some

TABLE 3 Observed similarities and differences between SA and PP feedback.

	SA feedback	PP feedback
Similarities	*Stemming from text mining*	
	(S1) Similar in both word usage and (relative) frequency	
	(S2) Equal distributions of feedback belonging to bad, moderate and good solutions	
	(S3) Equal sentiments in both feedback approaches.	
	Stemming from the qualitative analysis	
	(S4) Equal amounts of descriptive and corrective feedback	
	(S5) Both give hints for improvement, note parts that are missing and point to misconceptions an almost equal amount of times	
	(S6) Almost 1 of 20 feedback reports is erroneous	
Differences	*Stemming from a previous study* (Moons et al., 2022)	
	(D1) More feedback	(D1) Fewer feedback
	Stemming from text mining	
	(D2) Limited use of abbreviations	(D2) Abbreviations common
	(D3) Many structuring elements such as section titles	(D3) No structuring elements like titles
	Stemming from the qualitative analysis	
	(D4) More general, often used as a kind of checklist of right/wrong intermediate steps	(D4) More concrete and specific for the student's solution
	(D5) Addresses more deficits and strengths, including minor issues	(D5) Focuses mainly on main issues, less on minor deficits or strengths
	(D6) Feedback analysing the student's solution less common	(D6) Feedback analysing the solution is more common
	(D7) Less explanations of mistakes	(D7) More explanations of mistakes
	Stemming from both text mining as the qualitative analysis	
	(D8) Empty questions get 'No answer' as feedback	(D8) Empty questions often receive an encouraging statement to get the student started
	(D9) More elaborate feedback on mistakes	(D9) More short statements on mistakes
	(D10) Perfect solutions are often labelled with 'perfect', although more often accompanied by side remarks	(D10) Perfect solution praised with a variety of appreciation words, with no extra remarks

extra comments included in this case (D10); this is surprising as the SA condition featured a ready-to-use 'Perfect' button, which was not present in the PP condition. In contrast, the ready-to-use 'No answer given' button had the opposite effect (D8): in PP feedback, some encouragements or questions were included in the teachers' feedback when a question was left blank, while teachers in the SA condition mostly used the button. Finally, structuring elements like titles (D4) is an essential characteristic of SA; however, it is surprising that teachers did not naturally structure their feedback in the PP condition.

While our research question is answered in the previous paragraphs, we also aimed: (1) to learn how the use of statement banks (reusing of feedback) in general changes written feedback, and (2) how we can combine text mining with a qualitative analysis to compare feedback.

For the first aim, we should acknowledge some study limitations: the self-developed semi-automated assessment tool and the requirement for teachers to write atomic feedback make some similarities and differences rather specific for this research setting (eg, the 'No answer' button), making not all observed feedback characteristics applicable to the general use of statement banks. However, we still can conclude that when statement banks are deployed carelessly, it naturally drags teachers into less effective ways of giving feedback compared to classic written feedback not using statement banks. The feedback becomes more general, and the structure becomes centred on both major and minor aspects of the work that apply to many students, most likely because they can easily be repeated. Without a statement bank, the teachers' feedback is shorter but more focused on the main flaws in the students' work. Therefore, our main advice is to make teachers aware of this danger and that even when using statement banks, the rules of effective feedback remain key. The research context supports this claim: teachers were not informed about what constitutes content-rich feedback and were asked to treat the students equally in both conditions, making the mediocre feedback quality in both conditions not surprising, but the worsening when using statement banks all the more problematic.

Second, from combining text mining with ‘classical’ qualitative techniques, we learned that text mining gave us an overall idea about the differences and similarities in both feedback types; mainly in terms of the form of the feedback like abbreviations being more common in the PP condition and structuring elements in the SA condition, while the sentiments of the feedback were largely comparable. However, some phenomena were not discoverable using text mining only, like D10: Figure 3 suggests ‘Perfect’ dominated the SA condition, which turned out to be a more subtle story when combined with the qualitative analysis. Moreover, to make statements about the content and quality of the feedback, qualitative analysis was indispensable. Text mining for education (Ferreira-Mello et al., 2019) is a promising research field, but in our view, it is not yet a self-sufficient methodology for comparing texts.

One critical follow-up question remains: how does a student interpret SA and PP feedback? We listed similarities and differences, but the litmus test is to see how students can act on the given feedback; a fruitful idea for further research.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of several people. We are particularly grateful to Katie De Jonge (SGI Lennik) and her students for providing us with 60 tasks on linear equations, as well as to the 45 mathematics teachers from all over Flanders (Belgium), who spent an entire summer day at our university assessing these tasks.

FUNDING INFORMATION

This research is funded by a doctoral fellowship (1S95920N) granted to Filip Moons by FWO, the Research Foundation of Flanders (Belgium).

CONFLICT OF INTEREST STATEMENT

No conflicts of interest.

DATA AVAILABILITY STATEMENT

Data will be made available upon reasonable request.

ETHICS STATEMENT

Ethical clearance for this study was obtained from the Ethics Committee of the University of Antwerp. The committee approved the study design and the procedures for data management, consent and protecting the privacy of the participants.

ORCID

Filip Moons https://orcid.org/0000-0002-5368-3429
Alexander Holvoet https://orcid.org/0000-0002-7549-3247
Katrin Klingbeil https://orcid.org/0009-0003-6552-0701

Ellen Vandervieren https://orcid.org/0000-0003-1569-5274

REFERENCES

Benoit, K., Muhr, D., & Watanabe, K. (2021). *stopwords: Multilingual Stopword Lists*. https://CRAN.R-project.org/package=stopwords

Bokhove, C., & Drijvers, P. (2010). Digital tools for algebra education: Criteria and evaluation. *International Journal of Computers for Mathematical Learning*, *15*(1), 45–62. https://doi.org/10.1007/s10758-010-9162-x

Bose, M., & Dey, A. (2009). *Optimal crossover designs*. World Scientific. https://doi.org/10.1142/6878

Busch, J., Barzel, B., & Leuders, T. (2015a). Die Entwicklung eines Instruments zur kategorialen Beurteilung der Entwicklung diagnostischer Kompetenzen von Lehrkräften im Bereich Funktionen. *Journal für Mathematik-Didaktik*, *36*(2), 315–338. https://doi.org/10.1007/s13138-015-0079-8

Busch, J., Barzel, B., & Leuders, T. (2015b). Promoting secondary teachers' diagnostic competence with respect to functions: Development of a scalable unit in continuous professional development. *ZDM*, *47*(1), 53–64. https://doi.org/10.1007/s11858-014-0647-2

Candel, C., Vidal-Abarca, E., Cerdán, R., Lippmann, M., & Narciss, S. (2020). Effects of timing of formative feedback in computer-assisted learning environments. *Journal of Computer Assisted Learning*, *36*(5), 718–728. https://doi.org/10.1111/jcal.12439

Chang, N., Watson, A. B., Bakerson, M. A., Williams, E. E., McGoron, F. X., & Spitzer, B. (2012). Electronic feedback or handwritten feedback: What do undergraduate students prefer and why? *Journal of Teaching and Learning with Technology*, *1*(1), 1–23.

Chiles, M. (2021). *The feedback pendulum*. John Catt Educational Ltd.

De Smedt, T., & Daelemans, W. (2012). “Vreselijk mooi!” (terribly beautiful): A subjectivity lexicon for Dutch adjectives. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey* (pp. 3568–3572). http://www.lrec-conf.org/proceedings/lrec2012/pdf/312_Paper.pdf

Denton, P., & McIlroy, D. (2018). Response of students to statement bank feedback: The impact of assessment literacy on performances in summative tasks. *Assessment & Evaluation in Higher Education*, *43*(2), 197–206. https://doi.org/10.1080/02602938.2017.1324017

Denton, P., & Rowe, P. (2015). Using statement banks to return online feedback: Limitations of the transmission approach in a credit-bearing assessment. *Assessment & Evaluation in Higher Education*, *40*(8), 1095–1103. https://doi.org/10.1080/02602938.2014.970124

Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education*, *34*, 12–25. https://doi.org/10.1016/j.tate.2013.03.001

Education, Audiovisual and Culture Executive Agency. (2021). *Teachers in Europe: Careers, development and well being*. Publications Office. https://doi.org/10.2797/915152

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*(1), 70–120. https://doi.org/10.3102/0034654312474350

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(4), 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.ebo

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *WIREs Data Mining and Knowledge Discovery*, *9*, e1332. https://doi.org/10.1002/widm.1332

Gibbs, G., & Simpson, C. (2005). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, *1*, Article 1.

Gibson, S., Oliver, L., & Dennison, M. (2015). *Workload challenge: Analysis of teacher consultation responses* (Vol. *355*). Department for Education London.

Gleaves, A., & Walker, C. (2013). Richness, redundancy or relational salience? A comparison of the effect of textual and aural feedback modes on knowledge elaboration in higher education students' work. *Computers & Education*, *62*, 249–261. https://doi.org/10.1016/j.compedu.2012.11.004

Glover, C., & Brown, E. (2006). Written feedback for students: Too much, too detailed or too incomprehensible to be effective? *Bioscience Education*, *7*(1), 1–16. https://doi.org/10.3108/beej.2006.07000004

Grönberg, N., Knutas, A., Hynninen, T., & Hujala, M. (2021). Palaute: An online text mining tool for analyzing written student course feedback. *IEEE Access*, *9*, 134518–134529. https://doi.org/10.1109/ACCESS.2021.3116425

Gusukuma, L., Bart, A. C., Kafura, D., & Ernst, J. (2018). Misconception-driven feedback: Results from an experimental study. In *Proceedings of the 2018 ACM Conference on International Computing Education Research, Espoo, Finland* (pp. 160–168). https://doi.org/10.1145/3230977.3231002

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

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: Pressures and tensions. *ZDM*, *50*(4), 675–686. https://doi.org/10.1007/s11858-018-0944-2

Hujala, M., Knutas, A., Hynninen, T., & Arminen, H. (2020). Improving the quality of teaching by utilising written student feedback: A streamlined process. *Computers & Education*, *157*, 103965. https://doi.org/10.1016/j.compedu.2020.103965

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.

Lefevre, D., & Cox, B. (2017). Delayed instructional feedback may be more effective, but is this contrary to learners' preferences?: Timing of TBI feedback. *British Journal of Educational Technology*, *48*(6), 1357–1367. https://doi.org/10.1111/bjet.12495

MacLure, M. (2013). Chapter 9 classification or wonder? Coding as an analytic practice in qualitative research. In R. Coleman & J. Ringrose (Eds.), *Deleuze and research methodologies* (pp. 164–183). Edinburgh University Press. https://doi.org/10.1515/9780748644124-011

Moons, F., Vandervieren, E., & Colpaert, J. (2022). Atomic, reusable feedback: A semi-automated solution for assessing handwritten tasks? A crossover experiment with mathematics teachers. *Computers and Education Open*, *3*, 100086. https://doi.org/10.1016/j.caeo.2022.100086

Movshovitz-Hadar, N., Zaslavsky, O., & Inbar, S. (1987). An empirical classification model for errors in high school mathematics. *Journal for Research in Mathematics Education*, *18*(1), 3–14. https://doi.org/10.2307/749532

OECD. (2012). How many students are in each classroom? In *Education at a Glance 2012*. OECD. https://doi.org/10.1787/eag_highlights-2012-25-en

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, *35*(3), 277–289. https://doi.org/10.1080/02602930903541007

Ratkowsky, D. A., Evans, M. A., & Alldredge, J. R. (1993). *Cross-over experiments: Design, analysis and application*. Dekker.

Reusser, K., & Stebler, R. (1997). Every word problem has a solution—The social rationality of mathematical modeling in schools. *Learning and Instruction*, *7*(4), 309–327. https://doi.org/10.1016/S0959-4752(97)00014-5

Ryan, T., Henderson, M., & Phillips, M. (2019). Feedback modes matter: Comparing student perceptions of digital and non-digital feedback modes in higher education. *British Journal of Educational Technology*, *50*(3), 1507–1523. https://doi.org/10.1111/bjet.12749

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*(5), 535–550. https://doi.org/10.1080/02602930903541015

Schnepper, L. C., & McCoy, L. P. (2013). Analysis of misconceptions in high school mathematics. *Networks: An Online Journal for Teacher Research*, *15*(1), 625. https://doi.org/10.4148/2470-6353.1066

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* (1st ed.). O'Reilly.

Wager, W., & Wager, S. (1985). Presenting questions, processing responses, and providing feedback in CAI. *Journal of Instructional Development*, *8*(4), 2–8. https://doi.org/10.1007/BF02906047

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, *52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

Yang, K.-H., & Lu, B.-C. (2021). Towards the successful game-based learning: Detection and feedback to misconceptions is the key. *Computers & Education*, *160*, 104033. https://doi.org/10.1016/j.compedu.2020.104033

Yu, C., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *The Qualitative Report*, *16*(3), 730–744. https://doi.org/10.46743/2160-3715/2011.1085

Zumbach, D., & Bauer, P. C. (2021). deeplr: Interface to the 'DeepL' Translation API. https://CRAN.R-project.org/package=deeplr

How to cite this article: Moons, F., Holvoet, A., Klingbeil, K., & Vandervieren, E. (2024). Comparing reusable, atomic feedback with classic feedback on a linear equations task using text mining and qualitative techniques. *British Journal of Educational Technology*, *00*, 1–21. https://doi.org/10.1111/bjet.13447