Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

‘I know that I clicked but not if I read’: An Exploratory Study Comparing Data Traces and Self-Reports on Feedback Engagement

Eva-Maria TERNBLAD[a*], Agneta GULZ[a,b] & Betty TÄRNING[a]
[a] *Dept of Cognitive Science, Lund University, Sweden*
[b] *Dept of Computer and Information Science, Linköping University, Sweden*
*eva-maria.ternblad@lucs.lu.se

Abstract: The use of educational digital tools – both within and outside the classroom – has opened up for novel ways to investigate and pursue research on learning. For instance, the built-in intelligence that many applications have may also log the learner's actual activities, transforming them into models and statistics. Research reveals that such data-traces sometimes point in other directions than traditional self-reports, where students are asked about their learning activities or their use of a certain application or system. We wished to pursue this line of inquiry and explore if there are some types of estimations that correlate with objective measures, and hence if some types of estimations are better carried out through objective measures than through self-reports. Our study compared data-logged activities (clicks and eye-tracking measures) with self-reports on game-behavior for two types of estimations (requests for more feedback and reading feedback). To our knowledge, this kind of inquiry, comparing different types of data on feedback management, is rare. The results reveal that even if the students were quite good at estimating and reporting their requests for more feedback (in terms of clicks), they were substantially poorer at evaluating the extent of feedback messages they had read. These findings suggest that estimations of tasks that require more complex inner cognitive processes (e.g. reading) are more difficult for students themselves to report appropriately – whereas more procedural processes or concrete behaviors (i. e. clicking) are easier. These results are further discussed together with possible limitations in the measures used.

Keywords: Educational Games, Feedback, Reading behavior, Log files, Self-reports

1. Introduction

The digitalization of learning environments during the last decades has transformed education and learning science in fundamental ways. Not only may intelligent systems individualize learning by, for instance, providing automatic feedback, hints, or tailored instructions to a learner. In addition, log files may be collected and transformed into fine-grained statistics that may reveal how learners engage in a certain task, what resources they use, or if they engage in self-regulated learning. Such data traces, it turns out, often differ from self-reports, where students are asked to make judgements – often retrospectively – about their own applied strategies, actions, and efforts (Perry & Winne, 2006; Winne & Perry, 2000).

In fact, the validity of self-report instruments for evaluating learning activities is today broadly questioned, and in the field of self-regulated learning there are ongoing discussions of why self-report data may not be reliable indicators of the tactics learners actually use while they are studying (Hadwin, Nesbit, Jamieson-Noel, Code & Winne, 2007; Cho & Yoo, 2017). This might be particularly true for younger students, since they – due to less developed metacognitive capacities than those in youths and adults – have great difficulties when it comes to self-judgments (Demetriou & Kazi, 2006), and often strongly overestimate their own performances and abilities (Demetriou & Efklides, 1989). Younger children also tend to confuse the perceptions of academic ability with perceptions of appropriate social behavior (Paris & Newman, 1990). However, making proper judgements about one’s own learning strategies is considered important for gaining new knowledge (Zimmerman, 2001), and previous studies

show that metacognitive skills also often correlate with general performance (Bransford, Brown & Cocking, 2000).

One alternative to subjective reports is, of course, 'objective' ones. And, as stated above, today's digitized learning environments are developing rapidly, not at least in the field of learning analytics. Faster computers, better data storage together with sophisticated statistical tools make it possible to gather fine-grained and real-time digital patterns and feed this back to students, teachers, or researchers in a series of ways. And many researchers point to the possibility of not only evaluating students' active choices and strategies, but also their inner cognitive states and processes (Azevedo et al, 2013; Koedinger, d'Mello, McLaughlin, Pardos & Rosé, 2015) – all in the pursuit of optimizing teaching and learning situations and supporting students to become effective and self-regulated learners. One of the added values with gathering behavioral data in learning contexts is also to predict learning outcomes.

In this quest, it is of great importance to understand more about the pros and cons regarding objective versus subjective data, and to learn more about how such data best should be combined. However, before throwing ourselves into the big complex pond of learning analytics and student modelling, it may be of interest to start at a more fundamental level – not at least since many learning processes still remain to be explicitly defined. For instance, what are the relations between traceable interactive behaviors (such as clicking, reading) and the students' own thoughts about their learning processes (i. e. requesting/wanting information and/or processing it)? And what differences between self-reports and data logs may we expect when we focus on such processes and strategies? The purpose of this paper is to contribute to this line of research by looking at data from a specific kind of SRL-strategy, namely *feedback management*. By investigating two different types of behaviors – namely 'requesting' feedback and 'reading' feedback messages – we hope to spread some light over how self-reports and data traces may reveal different things, not only about the processes in general, but also about the student as a learner. To our knowledge, this kind of inquiry, comparing different types of data on feedback management, is rare.

1.1 Subjective and objective measures for evaluating SRL

When it comes to evaluating differences between self-reports and digital traces, most studies – at least those focusing on SRL – have mainly been concerned with the measures' appropriateness for modelling learning profiles or predicting learning gains. Here, Cho and Yoo (2017) found that subjective measures from the classical MSQL-questionnaire (Pintrich, Smith, Garcia & McKeachie, 1991) were less appropriate for predicting students' achievement levels (in form of final grades) than data traces from an online learning management system (Blackboard). However, van Halem, van Klaverena, Drachslerbc, Schmitzd, & Cornelisza (2020) received partially different results when comparing MSQL-reports with logs from another LMS (Canvas) accompanied by logs from a learning tool in statistics. Here, van Halem et al (ibid.) concluded that while some MSLQ-reports could explaine a substantial *proportion* of academic performance, other variables based on trace data instead might reveal a substantial *variation* in performance. And in another study by Ellis, Han and Pardo (2017), it was shown that the *combination* of self-reports and observational data provided a better predictive model of students' learning outcomes than any of the measures alone.

Fewer studies have made direct and straight-forward comparisons between self-report measures and data logs. However, Hadwin et al (2007) did an exploratory case study on eight students using gStudy, where they paired MSQL-reports with digital online behaviors (such as making questions, finding ideas, linking information, summarizing ideas, making lists and defining lack of knowledge). In general, self-reports correlated poorly with data logs, although they were slightly more comparable for some actions, such as summarizing important information. In another study, Kia, Hatala, Baker & Teasly (2021) added pop-up windows to an existing LMS (Canvas), where students answered multiple-choice questions about their ongoing behaviors (stated like 'which of the statements below best describes what you are doing on this assignment at this moment?'). These responses were then evaluated and compared to data logs categorizing different SRL-behaviors (task definition, planning, enactment, and adapting). It was shown that the correspondence between students' self-reports and categorized data logs varied between 32% and 73%. The most difficult behaviors to self-assess (or to interpret digitally, depending on what is thought of as the most correct) were 'planning' and 'adaption', while 'task definition' and 'enactment' (i.e., activities corresponding with working on the assignment) were easier.

Evidently, self-evaluating one's own behaviors is not straightforward, but even if self-reports have limitations, the weaknesses of data traces also need to be emphasized. As an example, a study by Tempelaar, Rienties & Nguyen (2020) confirms the individual biases in self-reports but lifts forth that the very biases also may be used as predictors for student performance. In addition, they point to the fact that data logged behaviors, such as a 'high level of learning activity' may have a series of different causes. A 'high level of activity' may signal an engaged and well performing student, but may also indicate a conscientious student, or a student with low proficiency in need of extra learning efforts. The authors conclude that especially trace data of 'process type' (e. g. is, number of attempts to solve an exercise, time on tasks, number of assignments completed) should be used with caution, whereas data traces of 'product type' (e. g. proportion of exercises correctly solved) could be better trusted, and especially so if complemented with self-report data. These conclusions are in line with other studies (Tempelaar, Rienties, Giesbers, 2015; Tempelaar, Rienties, Mittelmeier & Nguyen, 2018).

To summarize, even if some measures may be more reliable and/or useful than others for some specific purposes, it is not clear why and how subjective and objective measures of different phenomena tell us different things. However, in their review of 14 SRL-studies between 2003 and 2015, Rovers, Clarebout, Savelberg, de Bruin & van Merriënboer (2019) conclude that *granularity* is an important concept when comparing different measures of SRL, They also conclude that while self-report questionnaires may give a rather accurate insight into students' global level of self-regulation, behavioral objective measures are more accurate when evaluating specific SRL-strategies. This, on the other hand, seems like a simplification when considering the findings of Tempelaar et. al (2020).

1.2 Subjective and objective measures of feedback engagement and reading

Students' engagement with feedback is an important aspect SRL. For efficient learning to occur, you should attend to feedback, read it, understand it and act on it properly. When it comes to measuring students' thoughts of and use of feedback, this has historically been done through surveys, interviews or observations (Kerr, 2017; Eriksson, Björklund Boistrup & Thornberg, 2022). A classical setup is here to ask students what kind of feedback they like or dislike, if and how they make use of the feedback and why, and how the feedback ought to be formulated. On the other hand, when it comes to interpreting data logs for analyzing feedback behavior – which has been done in a series of studies during the last decades – the students' subjective opinions about the feedback (and their handling of it) have been set aside. Instead, the goal with the measurements has been to gather 'true' feedback engagement by, for instance, counting requests for feedback (Ternblad & Tärning, 2020), looking at if the feedback has been used when revising tasks (Silvervarg, Wolf, Blair, Haake & Gulz, 2021) or when the students engage with it (Chen, Breslow, DeBoer, 2018). Such data traces have often been gathered through the use of specifically designed digital learning applications or educational games (Tärning, Lee, Andersson, Månsson, Gulz, & Haake, 2020; Biswas, Roscoe, Jeong & Sulcer, 2009), but are, as far as we know, rarely analyzed in more open-ended learning systems.

To our knowledge, subjective and objective measures of feedback engagement have never been compared. This might be due to the simple structures of the used variables (such as clicking for getting more feedback) and that the digital footprints in this case seem to speak for themselves. To ask the students if they requested more feedback or not does in this case appear superfluous (if they don't remember clicking, they are simply mistaken, since they obviously did click). However, the student's judgements about their own behavior (that is, if they think of themselves as individuals wanting feedback, or if they are convinced that they behave appropriately even if they do not) may still be of interest. And equally, it may also be of value to evaluate if such judgements differ in respect to what is measured. Perhaps some behaviors are easier to self-evaluate than others? This is certainly the case when it comes to more complex measures, as Hadwin et. al (2007) and Kia et. al (2021) have pointed out.

One important aspect of feedback engagement is to properly attend to and read the feedback. To evaluate if and how students have read a text is mostly done by asking questions about its content. That means equating reading comprehension with reading (which may, of course, be the same thing). However, paying attention to a text and *trying* to read it, but without fully understanding it, is not the same thing as not attending to it at all. And if looking at these two behaviors in an SRL-context, not paying enough attention is far worse. So, how to measure and assess 'reading' without confounding it with 'understanding'?

One way of evaluating students' reading behaviors is, of course, to plainly ask them. This was typically done for evaluating reading habits and reading comprehension strategies during the late 1900's, revealing that students in all age groups have difficulties to self-report how and what they actually have read (Baker & Cerro, 2000). Not only do self-report measures concerning reading strategies lack a theoretical basis, but they are also highly context- task- and goal dependent (Hadwin et. al, 2007). Moreover, Cromley and Azvedo (2006) conclude, studying ninth graders readings of texts on American history, that one of the problems with self-reports is that they often are collected outside of the learning context (after of before the actual reading takes place). They compared prospective self-reports on specific reading behaviors with think-aloud protocols and a concurrent control measure of actual strategy use (gathered by multiple choice-questions). When trying to relate these measures to one another – as well as to a test on reading comprehension – they stated that the prospective self-reports were the only ones not correlating with any of the other measures.

A different, and today more common approach is to evaluate students' reading behaviors by using eye-tracking measurements. Eye-tracking is a non-intrusive measurement technique based on the Eye-mind hypothesis originating from Just and Carpenter (1980), which states that visual fixations on objects (or words) is strongly correlated with cognitively processing them. And even if it's important to bear in mind that the detection of eye-movements does not *directly* correspond to underlying cognitive processes (we may, for instance, attend to something that is not in our visual focus), it still constitutes a reliable instrument for evaluating capacities like information processing and visual attention (Carter & Luke, 2020). When detecting reading behaviors in eye-tracking studies, scanpath patterns are normally used together with durations of fixations and recordings of blinks and saccades. By using well proven algorithms (often based on Hidden Markov Models) it is then possible to predict whether a given sequence of events reflects a reading or not (see for instance Kollmorgen & Holmqvist, 2007; Simola, Salojärvi & Kojo, 2008; Wengelin et. al, 2009). And it is of course also possible to evaluate if some specific areas or objects have been attended to by merely looking at fixations. To our knowledge, eye-tracking measures capturing reading have rarely been compared to self-reports. Instead, they have been evaluated through reading comprehension tests and free recall.

1.3 The scope of the present study

This study has two purposes: First, to contribute to the line of research comparing subjective and objective measures important for SRL, and see if we, with our specific dataset and our specific subject (that is, feedback engagement), get similar results as previous studies, and if not, discuss why. Second, to illustrate how subjective and objective measures on feedback engagement may differ due to what is specifically measured. For this paper we looked at two different estimations students did after playing an educational game in history. On the one hand, we investigated fifth grader's retrospective self-evaluations of (i) how large a proportion of automatically presented feedback messages they read while playing an educational game, and (ii) how large a proportion of optional feedback messages they did request while playing. These self-reports were then compared to the students' gaze behavior (using eye-tracking) and logged actions (in terms of clicks), searching for correlations, deviations or similarities.

Of special interest to us was whether the subjective report of a more cognitive demanding and also more ambiguous task like reading (ambiguous in that way that it can mean different things to different people) would correlate less well with data-logging (i.e., eye-tracking measures) whereas self-reports for a more concrete task (estimating the number of clicks) would correlate better with the objective measures (i.e. data-logging). If such a discrepancy turned up, we were interested to learn what it could tell us about a) the different content and representability of the measures, and b) the students' metacognitive capacities of judging their own behavior. More specifically, our research questions were:

i) How good are middle-school students at judging their own 'clicking' and 'reading' behavior? And are there any differences between estimating 'reading' and estimating 'clicking' when comparing self-reports with data traces?

ii) Can potential differences between self-reports and data traces be used to predict performance?

Notably, the research questions specifically target students' engagement with feedback messages. And since engaging with feedback is an important part of self-regulated learning, we were also interested in discussing the possible benefits of using both types of measures and if and how comparisons between them may be used for evaluating students' SRL-skills.

2. Method

2.1 Participants, procedure and materials

45 5th graders (24 girls and 22 boys) from two classes from a school in southern Sweden participated in a study on feedback behavior by playing an educational game in history. In the game, students visited historical scenes and persons, searched for text-based information and solved tasks (either by constructing a concept map, categorizing statements, or pairing historical figures and events on a timeline) (see also Tärning et al, 2020). After having completed a task and having it corrected by a "correction machine", the students also received informative text-based feedback on selected errors and hints on how to correct them (see Figure 1 for a time-line task and examples of feedback on possible errors). The feedback was designed to be critical and constructive, scaffolding the students to do better when they revised the task. After the first feedback message, the students were offered to click on other marked errors to receive additional pieces of feedback. The students had to be at least 80% correct on a task to be allowed to start on a new one, and many students needed many trials to be able to progress in the game.

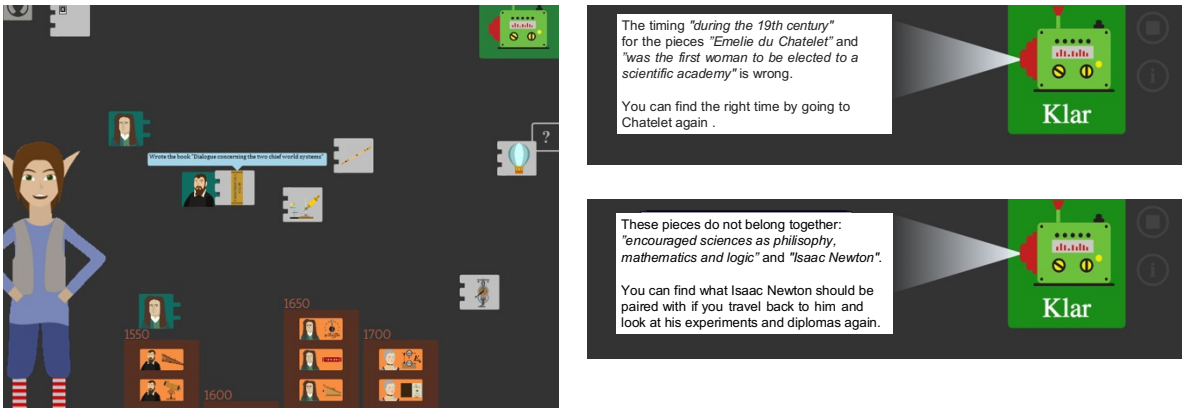

Figure 1. Timy constructing a time-line with historical persons and their scientific contributions (left), and examples of feedback from the correction machine (right).

In total, the students worked with the game during three one hour sessions. During the third session (the data from which we base the present analysis on), the students used computers with integrated eye-trackers situated in a classroom-like laboratory at Lund University. The experiment ended with the students filling in a questionnaire with questions regarding their efforts and behavior when playing. One of these questions targeted to what extent the students had read the feedback provided to them, while another asked to what extent they had clicked to receive more feedback on some other errors or mistakes. For both questions there were five possible answers: "always", "most often", "sometimes", "not very often" and "never".

2.2 Eye-tracking measurements and data logs

All screen activities were logged together with eye movement events (fixations, saccades and blinks) using an integrated SMI REDm eye tracking camera with SMI iViewX and SMI BeGaze 3.6 software with a sampling rate of 120 Hz. The eye-tracking recordings were used for evaluating if the students *noticed* the feedback or not, and if they *read* the feedback or not. For measuring if the feedback message was noticed, we used the *fixation based AOI hit*, which states for a fixation that its coordinate value is inside the area of interest (AOI) (Holmqvist et al., 2011). An AOI is defined as a region in the stimuli

in which the researcher is interested in gathering data. Consequently, in the present study, the areas containing feedback boxes were defined as AOIs. After verifying an AOI hit, every feedback instance was categorized as either 'noticed' or 'ignored', and the proportion of noticed feedback messages was calculated for each participant.

'Reading behavior' was identified by a learning algorithm in form of a support vector machine (SVM) (Lee, 2017). This was calibrated for each participant, using three eye movement measures: duration of fixation, saccadic amplitude, and regression. Every feedback text was then classified as 'read' or 'not-read' based on an intrinsic threshold determined in a pilot study (Lee, 2017). The reason for not using existing standardized models (which are based on adult readers) was due to the students' age and large variability in gaze behavior. Consequently, by simplifying the eye-tracking measure and adjusting it to each individual student, we ensured capturing a sufficient degree of 'reading-like' events.

Finally, we also calculated an in-game *performance measure* by using the logged scores (percent correct) for each task trial and averaging these values over all tasks and trials.

3. Results

To be able to compare self-reports with eye-tracking data, the five levels of 'clicked' and 'read' feedback from the questionnaires were translated into numbers: 100%, 75%, 50%, 25% and 0%. Data from 42 students was included in the final analysis, and the results are described in the passages below. It should be noted, that even if statistical significance levels are calculated and presented, no corrections for multiple analyses have been performed. The analysis is explorative – thus the results should be interpreted as indicative, mainly pointing in directions that could be further investigated.

3.1 Comparison between self-reports and data traces targeting feedback clicks

We start with an evaluation of the students' requests for additional feedback (here referred to as 'clicks'). The self-reported extent of feedback clicks (35% in average) was similar to the logged one (38% in average). A Mann-Whitney U-test revealed no significant difference between measures ($W=855$, $p = 0.67$). Studying the relationship between individual values by calculating Pearson's correlation coefficient, the objective and subjective measures were found to be significantly positively correlated: $r(34) = 0.40$, $p < 0.01$ (see Figure 2).

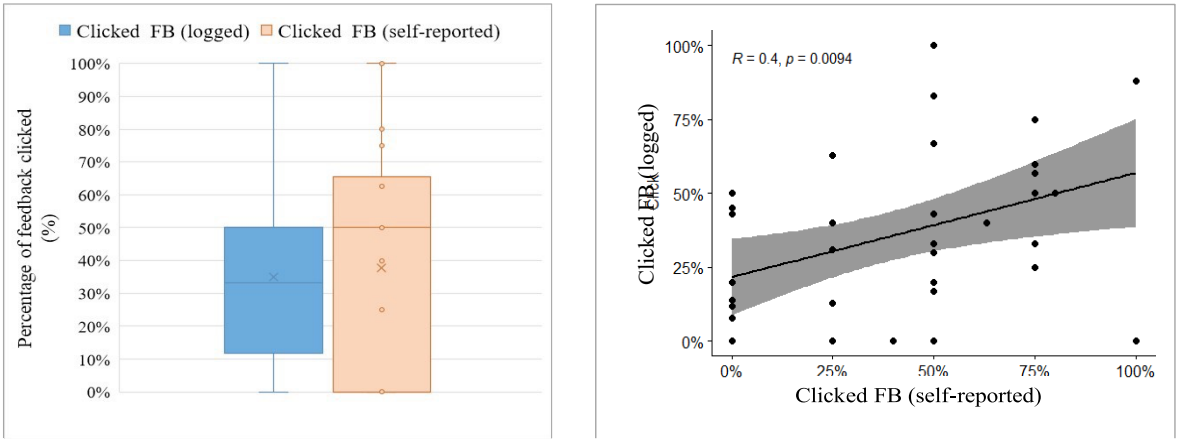

Figure 2. Comparison between logged and self-reported feedback requests (clicks). Left: Boxplot. Right: Scatterplot with Pearson's Correlation.

3.2 Comparison between self-reports and data traces targeting readings of feedback

When it comes to the students' reports of feedback readings, the self-reported extent of feedback read (59% in average) was larger than the one based on eye-tracking measurements (47% in average). A Mann-Whitney U-test here revealed a significant difference between the two measures ($W=486$, $p< 0.01$), indicating that the students, in general, overestimated the extent to which they read the messages.

In this case, the subjective and objective measures were not found to be correlated: r(34) = .05, p = .77. See Figure 3 for visualizations of the results.

To evaluate whether the students confused reading the feedback with merely 'glancing at it', the students' self-reported extent of feedback read (59%) was also compared to the logs of feedback 'noticed' (72%). A Mann-Whitney U-test revealed a weak but still significant difference between the two measures (W=777, p = 0.02), indicating that the students, in general, only self-reported a part of the noticed feedback messages as 'read'. These two measures were not found to be correlated either: r(34) = .04, p = .81 (see Figure 4).

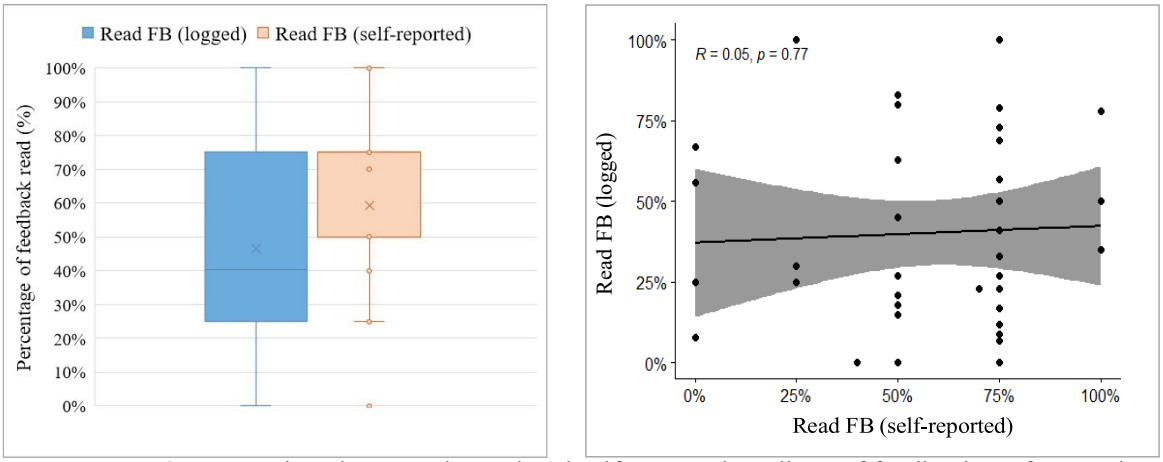

Figure 3. Comparison between logged and self-reported readings of feedback. Left: Boxplot. Right: Scatterplot with Pearson's Correlation.

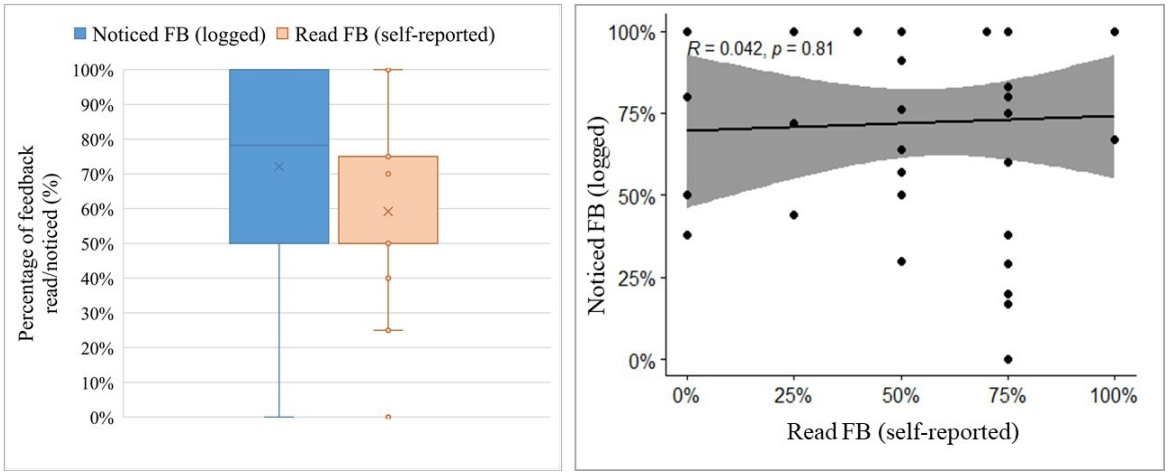

Figure 4. Comparison between self-reported readings of feedback and logs of feedback noticed. Left: Boxplot. Right: Scatterplot with Pearson's Correlation.

One might expect that the self-reported readings of feedback messages would lie somewhere in between the logs of 'noticed' and 'read' messages, and this was also true for 47% of the participants. However, 31% underestimated their reading, reporting to have read a lower extent of the messages than what was logged as 'read'. On the other hand 22% reported to have read a higher extent of messages than what was logged as 'noticed' – clearly making an improper judgement about their reading behavior, since it is impossible to read what you don't notice.

3.3 Comparisons of (mis)judgments and performance

As stated above, metacognitive capacities may correlate with academic performance. We therefore decided to investigate if the incorrectness of the individual student's judgement would correlate with their performance in the game, here focusing on feedback readings (since these were difficult to estimate). We could expect a negative correlation between this incorrectness (measured as the absolute

difference between the self-reports and logged feedback readings) and the average performance level (measured as % correct over all tasks and trials). The result is presented in Figure 5 below, revealing that the inaccuracy in the self-reports did not have any significant correlation with the students' average performance levels: r(34) = .17, p = .33.

As of yet, the self-reports on how much feedback the students had read haven't corresponded to any other measure, and the value of asking students about their reading behavior appears, so far, to be questionable. However, one final analysis, addressing if the self-reports could reflect other desirable and productive strategies (such as really trying to solve the tasks or performing well) is in its place. Hence, we decided to do a comparison between the self-reports on feedback readings and in-game performance. The result is presented in Figure 5 below, showing that the two measures had a positive and significant correlation: r(34)= 0.41, p=0.01.

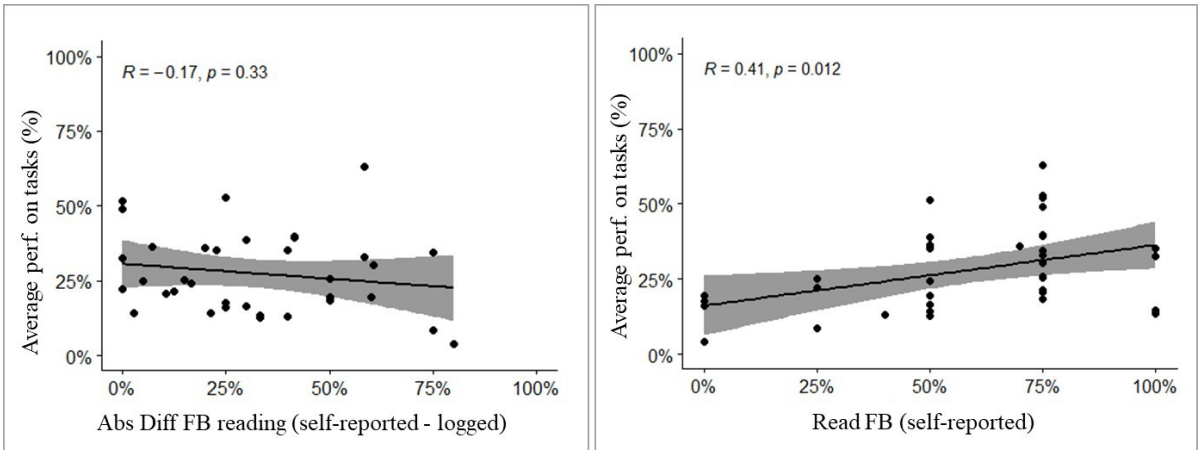

Figure 5. Scatterplots comparing in-game performance levels with the absolute differences between self-reports and data traces (left) and with the self-reports on feedback readings (right).

4. Discussion

Evidently, in the present study the students' ability to correctly judge to what extent they had 'read' a feedback text was substantially weaker than their ability to estimate to what extent they had made requests for additional feedback texts (by clicking). Even though some students overestimated (but also underestimated) their inclination to 'click', this behavior was less difficult to evaluate correctly compared to the cognitive, and quite complex, behavior of 'reading'. This is in line with previous research, showing that reading is difficult to self-assess (Baker & Cerro, 2000; Hadwin et. al, 2007). The difficulty in observing one's own reading behavior could be due to the fact that attention directed towards a text does not necessarily involve deeper engagement. The text could be 'noticed', 'glimpsed at' or 'skimmed' – without any proper decoding or understanding. However, the self-reports on feedback readings didn't correspond to the logged eye-tracking measures of having 'noticed' the feedback either. Indeed, self-evaluating this process appears to be hard.

Since reading is an important aspect of self-regulated learning, judging it properly is essential. However, many automatic processes that learners apply, and reading is one of them, often operate beneath the threshold of attention, making them hard to retrieve from memory (Perry & Winne, 2006). The self-report concerning 'reading the feedback' may therefore instead be an after-construction, related to performance, effort, or desirable behavior in school (a similar discussion regarding relations between retrospective-self-reports and achievement is done by Veenman and van Cleef (2019)). Hence, depending on if the students' feedback engagement led to a success or failure, the interpretation of 'having read the feedback' might differ.

In addition, the students' ability to correctly judge how much feedback they read (by looking at the differences between self-reports and data traces from eye-tracking measurements) did not correlate with in-game performance, questioning the actual benefits with comparing these two types of measures. Taken together, the findings suggests that 'reading' might be a typical process suitable for measuring objectively, without involving the student's own (and often biased) beliefs and wishes. Still, another

possibility is that the students only remembered reading (or not reading) the *latest* of the presented feedback messages, and thereby did not evaluate their behavior for the entire session. If this is the case, the problem is not the lack of accurate self-esteem, but flaws in memory retrieval. Consequently, questions about students' learning strategies should probably be posed when the students in fact are studying, and not after (or before, as Cromley and Azvedo (2006), also pointed out).

Before too strongly advocating eye-tracking for evaluating reading processes, the accuracy of the 'reading'-measure used in this study can, of course, also be questioned. Setting a different threshold would most certainly have led to a different outcome. And it might also be, that 'partly glancing at' or 'partly reading' the feedback messages could be more beneficial than not noticing them at all, and that 'reading' shouldn't be classified in a binary way. Unfortunately, at the same time as inner cognitive processes and operations are difficult to self-evaluate, they are also the most difficult ones to catch and quantify by intelligent algorithms and measurements. This is especially true for children, who behave not only differently, but also in a less stable and predictable way than adults.

It should finally be emphasized that students' self-reports may bring invaluable information about the student's own perceptions, attitudes, and beliefs about studying and learning. Hence, it might not be a question about which measures to use, but about *how* and *when* to register them and/or combine them. Gaining more knowledge about differences and similarities between measures is therefore essential, not at least before conducting large surveys about students' learning habits, or before turning data traces into advanced student models using sophisticated data mining techniques.

Acknowledgements

The presented research has been funded by Marianne and Marcus Wallenberg Foundation.

5. References

Azevedo, R., Harley, J., Trevors, G., Duffy, M., Feyzi-Behnagh, R., Bouchet, F. & Landis, R. (2013). Using Trace Data to Examine the Complex Roles of Cognitive, Metacognitive, and Emotional Self-Regulatory Processes During Learning with Multi-agent Systems. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 427-449). New York, NY: Springer.

Baker, L., & Cerro, L. C. (2000). Assessing Metacognition in Children and Adults. In G. Schraw & J. C. Impara (Eds.), *Issues in the Measurement of Metacognition* (pp. 99-145). Lincoln, NE: Buros Institute of Mental Measurements.

Biswas, G., Roscoe, R., Jeong, H. & Sulcer, B. (2009). Promoting Self-regulated Learning Skills in Agent-based learning Environments. In S. C.,Kong, H. Ogata, H. C.Arnseth, C. K. K. Chan, T. Hirashima, F. Klett, J. H. M. Lee, C. C. Liu, C. K. Looi, M. Milrad, A. Mitrovic, K, Nakabayashi, S. L. Wong, S. J. H. Yang, (Eds.) (2009). *Proceedings of the 17th International Conference on Computers in Education*. (pp. 67-74). Hong Kong: Asia-Pacific Society for Computers in Education.

Bransford, J.D., Brown, A.L. & Cocking, R.R. (2000). *How People Learn; Brain, Mind, Experience, and School (Exp. ed.). Washington, D.C:* National Academy Press.

Carter, B. T. & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology, 155*, 49-62,

Chen, X., Breslow, L., DeBoer, J. (2018). Analyzing productive learning behaviors for students using immediate corrective feedback in a blended learning environment. *Computers & Education, 117*, 59-74.

Cho, M-H., Yoo, J. S. (2017). Exploring online students' self-regulated learning with self-reported surveys and log files: a data mining approach. *Interactive Learning Environments, 25*(8), 970–982.

Cromley, J. G. & Azevedo, R. (2006) Self-report of reading comprehension strategies: What are we measuring? *Metacognition Learning, 1*, 229–247

Demetriou, A., & Kazi, S. (2006). Self-awareness in g (with processing efficiency and reasoning). *Intelligence, 34*(3), 297–317.

Demetriou, A., & Efklides, A. (1989). The person's conception of the structures of developing intellect: early adolescence to middle age. *Genetic, Social, and General Psychology Mono- graphs, 115*(3), 371–423.

Ellis., R. A., Han, F., Pardo, A. (2017). Improving Learning Analytics – Combining Observational and Self-Report Data on Student Learning. *Educational Technology & Society, 20*(3), 158–169.

Eriksson, E., Björklund Boistrup, L. & Thornberg, R. (2022). "You must learn something during a lesson": how primary students construct meaning from teacher feedback. *Educational Studies*, *48*(3), 323-340.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition Learning, 2*, 107–124.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye Tracking: A comprehensive guide to methods and measures*. Oxford, Great Britain: Oxford University Press.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Kerr, K. (2017). Exploring student perceptions of verbal feedback. *Research Papers in Education, 32*(4), 444-462.

Kia, F. S., Hatala, M., Baker, R. S. & Teasley, S. D. (2021). Measuring Students' Self-Regulatory Phases in LMS with Behavior and Real-Time Self Report. *In LAK21: 11th International Learning Analytics and Knowledge Conference*, Virtual Conference.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A. & Rosé, C. P. (2015). *WIREs Cogn Sci*, 6:333– 353

Kollmorgen, S., & Holmqvist, K. (2007). *Automatically detecting reading in eye tracking data*. (Working Papers 144, Department of Cognitive Studies, Lund University.)

Lee, Y. J. (2017). Effects of teachable agents on children's noticing and reading feedback in a digital educational game (Document ID: 48039). [Master's thesis, University of Vienna]. E-Theses – Hochschulschriften- Service. http://othes.univie.ac.at/48039/

Paris, S. G. & Newman, R. S. (1990). Developmental aspects of learning. *Educational Psychologist 25*(1), 87-102.

Perry, N. E., Winne, P. H. (2006). Learning from learning kits: gStudy traces of students' self- regulated engagements with computerized content. *Educational Psychology Review, 18*(3), 211–228.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the motivated strategies for learning questionnaire*. MI: University of Michigan.

Rovers, S. F. E., Clarebout, G., Savelberg, H C. M., de Bruin, A. B. H & van Merriënboer, J. J. G. (2019). Granularity matters: comparing different ways of measuring self-regulated learning. *Metacognition and Learning 14*(1), 1–19.

Silvervarg, A., Wolf, R., Blair, K. P., Haake, M., & Gulz, A. (2021). How teachable agents influence students' responses to critical constructive feedback. *Journal of Research on Technology in Education*, *53*(1), 67-88.

Simola, J., Salojärvi, J. & Kojo. L. (2008). Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research, 9*(4), 237-251.

Tempelaar D., Rienties B., Giesbers B. (2015). In search for the most informative data for feedback generation: Learning Analytics in a data-rich context. *Computers in Human Behavior*, 47, 157–167.

Tempelaar D., Rienties B., Mittelmeier, J., Nguyen, Q. (2018). Student profiling in a dispositional learning analytics application using formative assessment. *Computers in Human Behavior*, 78, 408–420.

Tempelaar, D., Rienties, B., Nguyen, Q. (2020). Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application. *PLoS ONE 15*(6): e0233977.

Ternblad, E-M., & Tärning, B. (2020). Far from success – far from feedback acceptance? The influence of game performance on young students' willingness to accept critical constructive feedback during play. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), Artificial Intelligence in Education. Lecture notes in computer science: Vol. 12163. (pp. 537-548). Springer, Cham.

Tärning, B., Lee, Y. J., Andersson, R., Månsson, K., Gulz, A., & Haake, M. (2020). Assessing the black box of feedback neglect in a digital educational game for elementary school. *Journal of the Learning Sciences, 29*(4-5), 511-549.

Van Halem, N., van Klaverena, C., Drachslerbc, H., Schmitzd, M., Cornelisza, I. (2020). Tracking Patterns in Self-Regulated Learning Using Students' Self-Reports and Online Trace Data. *Frontline Learning Research 8*(3), 140–163.

Veenman, M.V. J., van Cleef, D. (2019). Measuring metacognitive skills for mathematics: students' self-reports versus on-line assessment methods. *ZDM Mathematics Education 51*(4), 691–701.

Wengelin, Å., Torrance, M., Holmqvist, K., Simpson, S., Galbraith, D., … & Johansson, R. (2009). Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior Research Methods, 41*(2), 337-351.

Winne, P. H., Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531-566). San Diego, CA: Academic Press.

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives,* (pp. 1–37). Mahwah, NJ: Erlbaum.