Studies in Educational Evaluation 79 (2023) 101303

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

The role of reference frames in learners' internal feedback generation with a learning analytics dashboard

Lars de Vreugd [a,*,1], Renée Jansen [a,2], Anouschka van Leeuwen [b,3], Marieke van der Schaaf [a,4]

[a] Utrecht Center for Research and Development of Health Professions Education, University Medical Centre Utrecht, Heidelberglaan 100, 3508 GA Utrecht, the Netherlands
[b] Department of Education, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Self-regulated learning
Appraisal
Internal feedback
Learning analytics dashboard

ABSTRACT

Being able to self-regulate can positively impact learners' academic achievement. An inherent catalyst of Self-Regulated Learning (SRL) is internal feedback, the new knowledge which is generated when comparing current knowledge against reference information. Learners may not always generate internal feedback, hampering further SRL. Supporting SRL can be done with a Learning Analytics Dashboard (LAD), in which reference frames allow for comparisons and facilitate internal feedback generation. This study explores internal feedback generation using a LAD and the effect of reference frame availability. A multiple method design examined the interplay of reference frames, comparison processes, internal feedback generation and preparatory activities engagement. Differences between three conditions were explored using Bain ANOVA's. Results showed that reference frames almost exclude other external comparators and are used in parallel with an internal comparator. A peer reference frame leads to most verbalizations of internal feedback, and potentially to most verbalizations of preparatory activities.

1. Introduction

Being able to self-regulate positively impacts academic achievement (Richardson et al., 2012), and learners' self-regulation provides a sense of control resulting in more positive emotions (Pekrun, 2006). The generation of internal feedback is a catalyst for successful self-regulation (Nicol, 2021). Internal feedback generation contributes to competency development by comparing current knowledge, skills or attitudes to some frame of reference. Competencies can be myriad, e.g. musicians learning to be better musicians by comparing compositions or students learning to write better by comparing papers with peers (Nicol, 2021). Generated internal feedback informs students' further self-regulation of learning by discovering gaps in essential knowledge or obtaining awareness of insufficient study skills. Self-Regulated Learning (SRL) can be supported by providing learners with information about their learning process using a Learning Analytics Dashboard (LAD) (Marzouk et al., 2016). Learners may use this information to reflect on their strengths and discover potential points of improvement – i.e., generate internal feedback. For LADs several types of reference frames are advocated to support learners' interpretation of presented data (Wise & Vytasek, 2017). For example, a social reference frame shows an average of peers or top achievers, whereas a progress reference frame shows personal development over time (Jivet et al., 2017). These reference frames could be used as comparators and therefore facilitate internal feedback generation.

Problems may arise if learners do not generate internal feedback regarding their learning. Insufficient internal feedback may hinder meaningful self-regulation. If learners do generate internal feedback, they may not make accurate assessments automatically or use relevant sources of information as reference frames. This could lead to the generation of inaccurate internal feedback. If inaccurate internal feedback is generated, further self-regulation of learning may also be impaired. LADs may elicit and support internal feedback generation, but how internal feedback is generated when using a LAD is not yet clear.

* Corresponding author.
E-mail address: l.b.devreugd-2@umcutrecht.nl (L. de Vreugd).
[1] ORCID: 0000-0003-3486-0407.
[2] ORCID: 0000-0002-8385-8322.
[3] ORCID: 0000-0003-2970-1380.
[4] ORCID: 0000-0001-6555-5320.

https://doi.org/10.1016/j.stueduc.2023.101303
Received 9 November 2022; Received in revised form 1 April 2023; Accepted 18 August 2023
Available online 12 September 2023

Furthermore, reference frames in a LAD allow for comparisons, but how internal feedback generation is affected by presenting different types of reference frames also remains unclear.

This study therefore aims to gain insight in internal feedback generation and the potential support of a LAD for generating internal feedback. This study will provide insight into how learners' internal feedback generation and self-regulation of study behavior can be supported. These insights could inform LAD design.

1.1. Self-regulated learning

Generation of internal feedback is an inherent part of the appraisal phase in SRL (Nicol, 2021). This study follows the description of SRL by Panadero (2017), namely as a conceptual framework to understand cognitive, motivational, and emotional learning processes. SRL entails a cyclical process, often divided into three phases: a preparatory, performance, and appraisal phase (Panadero, 2017). The preparatory phase revolves around task analysis, setting goals, planning, self-motivating (Winne & Hadwin, 1998; Zimmerman, 2002), and activation of prior knowledge (Pintrich, 2000). In the subsequent performance phase, learners apply tactics and strategies (Winne & Hadwin, 1998), self-control (i.e., select and deploy specific strategies), self-observe (i.e., record events to reflect upon) (Zimmerman, 2002), and monitor themselves and their task performance (Pintrich, 2000). Finally, in the appraisal phase, learners attribute causal relations, self-react to their learning performance (Zimmerman, 2002), and they may receive performance feedback to evaluate (Boekaerts & Corno, 2005). A new self-regulation cycle may be instigated from conclusions or insights that arise in the appraisal phase, which may lead to a new preparatory phase in a SRL cycle. This is enabled when internal feedback is generated within the appraisal phase (Nicol, 2019).

1.2. Internal feedback within SRL

Internal feedback refers to "… *the new knowledge that students generate when they compare their current knowledge and competence against some reference information.*" (Nicol, 2021, p. 2). The new knowledge may relate to both content and process, e.g. informing learners whether or not the chosen approach to performance needs adjustment. In practice, generating internal feedback may happen when a learner reflects on a past performance and notes discrepancies between task performance and the intended outcome of that performance. These discrepancies may lead to setting (new) goals or choosing a new approach to improve future performance, inducing a new preparatory phase in SRL (plan to make adjustments for future tasks), and improved or altered performance.

The main mechanism for internal feedback generation is self-assessment, which is used interchangeably with (i.a.) evaluation or reflection in SRL (see Nicol, 2021). A key component in internal feedback generation is making comparisons based on sources of information (Nicol, 2021). For example, learners writing a paper generate internal feedback by comparing to a peer's paper, previous papers, or using assessment criteria. These comparators help assess the paper's quality and determine potential improvements. Learners may use multiple sources of information to make comparisons. These sources can be in the external information environment, for example a peer's paper or model paper to compare one's paper to. Also, sources of information can originate in the internal mental environment, for example internalized criteria about a paper's quality or feelings that arise when writing a paper (Nicol, 2021). Comparing to different sources of information may lead to new insights or knowledge (i.e., internal feedback), eliciting a new SRL cycle by setting new goals to improve future performance.

External comparators can be specified as being *analogical* or *analytical*. Analogical comparators are sources of information similar to performance, e.g. a peer paper as comparator for your paper. In general, they are concrete and understandable. Analytical comparators are more abstract sources of information, e.g. a rubric as comparator for a paper. They require more deliberation to use in a comparison (Nicol, 2021). Multiple comparators may be used in parallel to further calibrate internal feedback. Using multiple comparators (analogical and/or analytical) means that different perspectives are taken during internal feedback generation and may lead to differences in internal feedback (Nicol & McCallum, 2022).

All in all, learners are able to regulate their learning and develop self-regulatory abilities if appropriate comparisons are made (Nicol, 2021). What comparisons learners make partially depends on what comparators are available at the time. If appropriate comparisons can be made, learners may generate internal feedback (i.e., appraise) regarding (e.g.) their learning process, instigating a new SRL cycle. However, learners may not always consciously reflect on, or self-assess their study behavior. This possibly leads to insufficient or no generation of internal feedback, hampering initiation of a SRL cycle.

1.3. Learning analytics for internal feedback generation

Learning Analytics can elicit or support internal feedback generation (and subsequent SRL) (Roll & Winne, 2015; Wise et al., 2016). In Learning Analytics Dashboards (LADs), indicators about learners, their learning processes, and/or learning contexts are aggregated and visualized (Schwendimann et al., 2016). By presenting information that is derived from learning activities, LADs may stimulate awareness of and reflection on (i.e., appraise) learning activities (Verbert et al., 2014). It is crucial for learners to use information for accurate decision making whilst appraising their learning (Viberg et al., 2020). However, it is not clear how a LAD may support the generation of internal feedback, what processes dashboard users engage in, or how the information in the LAD is used. If no direct comparators are provided, making comparisons and generating internal feedback may be hampered. Other internal or external comparators may be used instead. In effect, there may be less generation of internal feedback and less subsequent preparatory activities. The first goal of this study is to explore this process of making comparisons, generating internal feedback, and preparatory activities engagement when using a LAD without additional comparators (i.e., reference frames).

In a LAD, *reference frames* are "*the comparison points which orient students' interpretation of analytics*" (Wise et al., 2016, p. 170). Reference frames can be provided as external comparator and allow for comparison and support data interpretation (Jivet et al., 2017; Wise & Vytasek, 2017). Within learning analytics literature, several reference frames are identified. First, a social reference frame allows users to compare their scores to that of peers, for example an average of peers or that of top achievers. Second, an achievement reference frame allows for comparison to a predetermined criterion, for example a criterion set by a teacher, or a goal set by the learner. Third, a progress reference frame allows users to compare to their earlier performance using historical data (Jivet et al., 2017; Wise & Vytasek, 2017).

In the current study, a peer reference frame (social) and criterion reference frame (achievement) are used. The peer reference frame represents the average fellow student and represents a concrete point of comparison. This type or reference frame can be perceived as an analogical comparator. It is quite similar in visual format as the presented score, is also expressed as a percentage, and pertains to an average student's score on that construct. A peer reference frame as an analogical comparator may easily be used when generating internal feedback and subsequent preparatory activities. A criterion reference frame may be perceived as an analytical comparator. It is also quite similar in visual format as the presented score and also expresses a percentage, but it represents a goal to reach determined by people unknown to the dashboard user (in contrast to peers, whom they know). Compared to the peer reference frame it is potentially more abstract and perhaps more of an analytical comparator. Internal feedback may be generated differently when the criterion reference frame is used as

comparator. It is unclear how similar the criterion reference frame is to the peer reference frame in supporting internal feedback generation and subsequent preparatory activities, or if using it as a comparator may be more difficult leading to different internal feedback generation.

Both reference frames may function as external comparators, but as they differ generated internal feedback may differ as well. The second goal of this study is therefore to explore the effect of providing a peer or criterion reference frame as external comparator on the generation of internal feedback and subsequent preparatory activities. Note that providing a progress reference frame requires availability of dashboard users' historical data. Prior engagement with a dashboard may influence participants' use and information interpretation. All participants in this study were first time LAD users, the effect of a progress reference frame was therefore not explored.

1.4. Current study

To summarize, generation of internal feedback is an inherent catalyst for self-regulation. Internal feedback is based on learners' reflection and self-assessment and requires making comparisons between current performance and a frame of reference. Learners may not automatically generate internal feedback regarding their study behavior. Providing information in a LAD may stimulate internal feedback generation and subsequent preparatory activities. To further stimulate internal feedback generation, several types of reference frames can be provided as comparator. This study aims to answer two research questions.

The first research question pertains to internal feedback generation when using a LAD without a reference frame, and what information is used as comparator in that situation. If no direct comparators are available in a LAD, other information must be used. This information could be internalized criteria such as the passing grade for an exam or the score for a different construct in the dashboard. The aim of this RQ is to explore this process, what comparators are used, and how internal feedback is generated. Thus, RQ1 is: *"To what extent and how is internal feedback generated whilst using a LAD when no reference frame is available?"*.

The second research question explores differences in internal feedback generation and making comparisons when a reference frame is presented in the LAD. If a peer reference frame is presented, making comparisons may be more frequent than when no reference frame is available. A criterion reference frame may be used in a similar way to a peer reference frame, but it may also be too analytical to use as a comparator. If no reference frame is available, dashboard users may have to use other comparators, be it external or internal. Therefore, RQ2 is: *"How does availability of no reference frame, a peer reference frame, or a criterion reference frame affect internal feedback generation and the comparison process?"*.

2. Method

2.1. Design

A multiple method experimental study with three conditions was performed. All data was collected during the 2021/2022 academic year at Utrecht University.

2.2. Participants

The sample consisted of 30 university students (20 female) from different study programs. All study programs had implemented the dashboard used in this study in their educational program (i.a. educational sciences, economics, and biomedical sciences). Participants were 1st (n = 20), 2nd year (n = 3), 3rd year (n = 1) bachelor students, and master (n = 6) students. Participants were recruited before or after a lecture or invited via email. Participants received €10 compensation. To divide participants randomly over the three conditions, they were assigned in the order in which they contacted the researcher (the 1st participant to condition 1, the 2nd to condition 2, and so on). In total, 37 participants were assigned to a condition, of which five dropped out, and two were excluded for already having used the dashboard. Each participant interacted with one of the three types of reference frame whilst thinking aloud and answering questions from the researcher.

This study is approved by the Ethics Committee of the Faculty of Social and Behavioral Sciences of Utrecht University under file number 21-0386.

2.3. Instruments

Thermos dashboard

To elicit and support students' appraisal of study behavior, the Thermos dashboard (Fig. 1) was developed. The dashboard aims to support students in reflecting on their own study behavior and to determine what construct they could or would like to improve. The aspects of study behavior in the dashboard are presented as generic, so not specified to a certain context (e.g., a specific course or study task). This allows students to use the dashboard during different phases of their study, helping them gain overarching insight in their study behavior. During their studies, they apply these insights in specific contexts. The dashboard offers suggestions for concrete actions to further improve that construct of study behavior. Data are gathered with a self-assessment questionnaire, which includes general info, the Motivation and Engagement Scale (MES, Martin, 2007), and the Group work Skills Questionnaire (GSQ, Cumming et al., 2015). The results are presented in graphs (Fig. 1, part 2 and 3). Furthermore, study progress data is retrieved from the university's data management system and presented (Fig. 1, part 4). Returning users can revisit earlier moments of dashboard use (Fig. 1, part 5). Feedback is presented in the feedback box if a user hovers over or clicks on one of the 13 constructs (Fig. 1, part 6). The feedback explains each construct's meaning, presents the user's score as a percentage, and informs the user why the construct is important for studying at university. Actionable feedback is available via the 'Prepare', 'Act', and 'Reflect' buttons, which show exercises to individually engage in. It also shows an 'Additional support' button, which offers suggestions for further support (e.g. a study coach).

2.3.1. Dashboard Reference Frames

In condition one (C1), no reference frame was available (Fig. 2). When hovering over a construct, participants in C1 saw their percentage in a tooltip: a small, see-through textbox showing (e.g.) Self-belief: 61 % (Fig. 1, part 7). In condition two (C2), a peer reference frame (i.e. average of peer students) was shown per construct as a line in the graphs, accompanied by a tooltip explanation, e.g. "Self-belief: 61 % (Peer reference: 82 %)". Reference frame percentages were based on MES guidelines (Martin, 2016) and aggregated data from one earlier cohort dashboard users. For condition three (C3) the same lines were shown in the graphs as in C2 but were labeled as a criterion reference frame, e.g. "Self-belief: 61 % (criterion reference: 82 %") (Fig. 2). The peer and criterion reference frame were further explained per construct in the feedback widget (Fig. 1, part 6), e.g. "For Self-belief, the average score of your peers (peer reference) is: 82 %." (C2), and "For Self-belief, the recommended percentage for studying successfully (criterion reference) is: 82 %" (C3). Reference scores for C2 and C3 were identical, to avoid variability in interpretation due to different distances of scores to the reference frame, and because there was no data available suitable to determine what is needed to 'study successfully'.

2.3.2. Interview protocol

A general interview protocol including thinking aloud and open questions was developed for all conditions, specified per condition where needed. The protocol consisted of a think aloud phase, guided interpretation phase, and a reflection phase. Before the think aloud phase, rapport was created between researcher and participant by

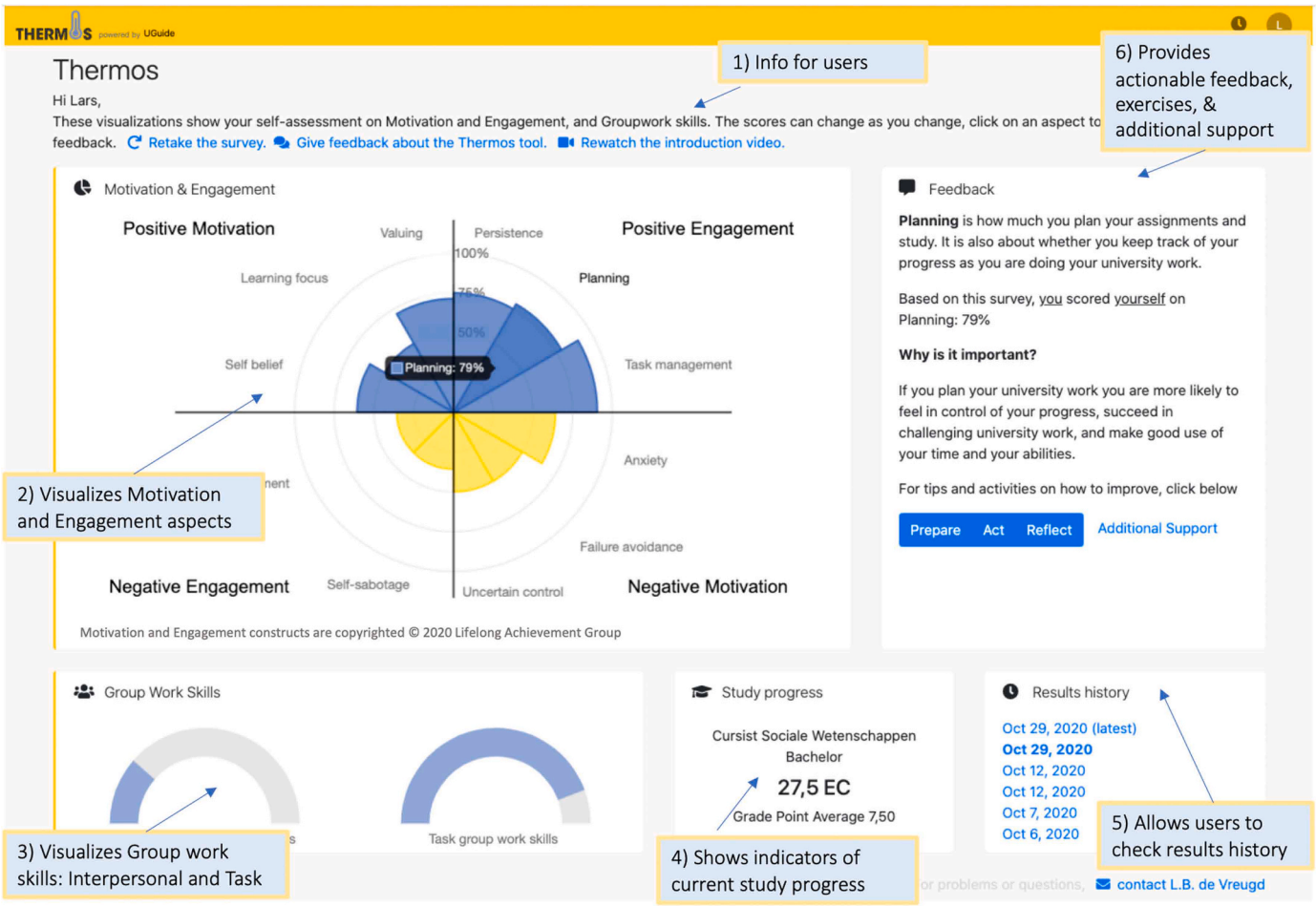

Fig. 1. The Thermos dashboard (without reference frames), with indicators and brief explanation of the different parts in it.

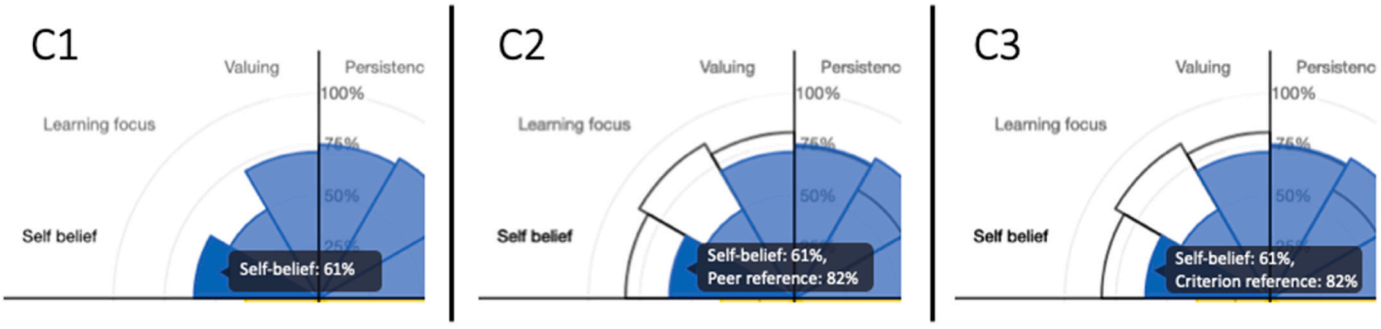

Fig. 2. Example of reference frame presentation and tool-tip information for condition 1 (C1), condition 2 (C2) and condition 3 (C3).

asking some introductory, unrelated questions (cf. Cohen et al., 2018). Participants performed two practice tasks (e.g. "Find out what the best restaurant in the Netherlands currently is, whilst thinking aloud") to get acquainted with thinking-aloud (Gibson, 1997; Miller-Young, 2013). During practice-tasks, participants were prompted by following fixed rules if (e.g.) a choice was not explained or if they were inaudible. After think-aloud practice, participants logged on to the dashboard and watched the video tutorial. In it, dashboard use instructions are provided, making sure participants have the information to work with the LAD. Participants then filled out the dashboard questionnaire.

In the think aloud phase, participants were asked to think aloud and freely interpret their scores on the LAD. The researcher turned off the camera during this phase to avoid distracting the participants or making them feel observed. Prompts were again given by fixed rules. The specific instruction participants received for the think-aloud phase was "*When you see your scores visualized, focus on interpreting and understanding them whilst thinking aloud*".

In the think-aloud phase, participants might not reflect on every construct within the LAD. Therefore, in the subsequent guided interpretation phase, participants were asked to reflect on scores of five constructs in the LAD (i.e., Self-belief, Planning, Failure Avoidance, Self-sabotage, and Interpersonal group work skills). The instruction was: *"Please explain what you think about your percentage on [Self-belief] and why you think that, and please tell me when you're finished"*.

In the final phase, participants were asked to reflect on the use of reference frames in the LAD. Participants in C1 were asked if they

thought having a reference frame would have helped them interpret their scores, and if so, what kind of reference frame they would prefer. Participants in C2 and C3 were asked how they used the reference frame available to them to interpret their scores. Then, participants in C1 were asked how they would feel if a reference frame was presented to help interpretation, while participants in C2 and C3 were asked how they felt when using the reference frame. Finally, participants in all conditions were asked if they had any last thoughts or remarks regarding their interpretation process.

The interview protocol was tested with three participants from a similar population, based on which minor adjustments were made. For example, the answer of the original second think-aloud practice task was perceived as too easy and did not offer a meaningful opportunity to practice thinking aloud. Therefore, it was replaced with a more difficult task.

2.4. Procedure

The entire procedure consisted of an instruction, interview, and debriefing section. Interviews were held via MS Teams in Dutch (25) or English (5), with the dashboard language set accordingly. Participants provided informed consent before participating. After joining the Teams meeting, participants could ask questions and were told they could do so again at the end. Participants closed all computer windows and disconnected any additional computer monitors to avoid unwanted sharing of personal information or distractions. Recording (audio and video) started, and participants provided general information (gender, study program, study year) and whether they were entitled to additional support for (e.g.) dyscalculia. No participants were excluded based on additional support. Next, the interview protocol was followed.

After data collection, participants in C2 and C3 were debriefed regarding the presented reference frames, explaining that scores below the reference frame did not mean they can't be successful students. Participants in C3 were also explained that criterion reference frame scores were actually those of peers, and why this deception was necessary (i.e., to ensure differences in reference frame scores would not interfere with the reference frame's effects on participants' internal feedback generation). Finally, participants could ask questions and were invited to ask questions later via email.

2.5. Data analysis

2.5.1. Coding verbal data

The interviews were transcribed and used for coding. The data were analyzed following Chi's (1997) verbal analysis approach. Segmenting and coding verbal data was done simultaneously, for which a coding scheme with 6 categories was developed. Category 1 contains researcher prompts and instructions, Category 2 pertains to dashboard use (e.g. clicks). Categories 3–6 were based on Nicol's (2021) framework. Category 3 (External Comparators) concerns external comparator use, for example reference frames or construct explanations in the dashboard. Category 4 (Internal Comparators) pertains to internal comparators use, e.g. a 5.5 (passing grade) or perception of construct proficiency. Category 5 (Internal Feedback) contains generated internal feedback. As this was of most interest, it was divided into three subcategories, 'Score judgment', 'Awareness', and 'Affective products'. Category 6 (Preparatory Activities) pertains to preparatory activities verbalization in SRL, for example goal setting. See Appendix A for the full coding scheme.

Next, the scheme was applied by three researchers independently (first, second, and third author) on a subset of excerpts from all conditions. Based on the first round, a 7th code category was added. This contained codes not included in Nicol's (2021) framework but deemed relevant for internal feedback generation. For example, a participant verbalized "*I don't believe this score*". The participant did not accept the presented score, potentially influencing internal feedback generation. The procedure of coding and revising was repeated in several rounds. Per round, Inter Coder Reliability (ICR) (O'Connor & Joffe, 2020) was calculated with Krippendorff's α as reliability measure (Hayes & Krippendorff, 2007) using SPSS v26.0.01. After the 5th round, ICR resulted in Krippendorf's $\alpha = 0.82$; values ≥ 0.80 are deemed reliable (Krippendorff, 2018). Excerpts from 15 transcripts were coded and discussed in total. The first author then independently coded remaining transcripts.

2.5.2. Internal feedback generation without reference frame

To answer RQ1, data from the no reference frame condition was sorted into a table in which code frequencies for every participant were listed. The first author inspected the table and the transcripts for patterns.

2.5.3. Influence of reference frames on internal feedback generation

To answer RQ2, Bayesian informative (Bain) hypotheses evaluation was used. Bain allows for evaluation and comparison of highly specific hypotheses concerning relationships between parameters (e.g. for 3 means $\mu1 > \mu2 > \mu3$ or $\mu1 = \mu2 > \mu3$), resulting in increased statistical power. Bain is an alternative to classical null hypothesis significance testing which has received frequent criticism (e.g., Cohen, 1995; Masson, 2011; Wagenmakers, 2007). Bain avoids p values and pre-determined significance levels, related issues of publication bias (see e.g. Simmons et al., 2016) and questionable research practices (see e.g. Wicherts et al., 2016).

For comparing two or more hypotheses in Bain, the Posterior Model Probability (PMP) is used. Values range from 0 to 1. Higher values indicate more support in the data for that specific hypothesis, in that set of evaluated hypotheses (Hoijtink et al., 2019, p.23). If hypotheses contain equality constraints (=), sensitivity analyses are needed to determine the influence of prior variance (see Hoijtink et al., 2019, p.29). As several hypotheses in this study contain equality constraints, sensitivity is examined by setting the fraction of information to 1, 2 or 3, and interpreting changes in PMP values. A null hypothesis can be evaluated by reporting PMP values for the null hypothesis ($\mu1 = \mu2 = \mu3$). A complement hypothesis (Hc) represents all other hypothesis combinations, μ1;μ2;μ3 (Hoijtink et al., 2019, p.10). There are no predetermined benchmarks for PMP value interpretation (Hoijtink et al., 2019, p.20). In this study we examine what hypothesis receives most support and how this differs from other hypotheses.

Bain ANOVA's were used to investigate differences on the amount of codes for six code categories that are relevant for internal feedback generation, i.e. External Comparators, Internal Comparators, Internal Feedback: Score Judgment, Internal Feedback: Awareness, Internal Feedback: Affective Products, and Preparatory Activities (Appendix A). For Bain ANOVA's, JASP (JASP Team, 2023) (v0.16.3) with Bain package was used.

For external comparators, we hypothesized that having no reference frame leads to the least external comparators used (H1a and H1b). Having a peer reference frame may lead to the most use of external comparators (H1b and H1c). The criterion reference frame may have a similar effect as the peer reference frame (H1a and H1d), or it may be too analytical and resemble the no reference frame condition (H1c and H1d). This leads to a set of four hypotheses and a complement:

H1a: μPeer RF = μCriterion RF > μNo RF

H1b: μPeer RF > μCriterion RF > μNo RF

H1c: μPeer RF > μCriterion RF = μNo RF

H1d: μPeer RF = μCriterion RF = μNo RF

Hc: μPeer RF, μCriterion RF, μNo RF

For internal comparators we hypothesize that having no reference frame leads to the most internal comparators used, as no external comparator is readily available (H2a and H2b). Having a peer reference

frame may lead to the least use of internal comparators (H2b and H2c). The criterion reference frame may again be similar to the peer reference frame (H2b and H2d) in the effect on making internal comparisons by students, or it may be too analytical and thus resemble the no reference frame condition (H2a and H2c). This leads to a set of four hypotheses and a complement:

H2a: μNo RF > μCriterion RF > μPeer RF

H2b: μNo RF > μCriterion RF = μPeer RF

H2c: μNo RF = μCriterion RF > μPeer RF

H2d: μNo RF = μCriterion RF = μPeer RF

Hc: μPeer RF, μCriterion RF, μNo RF

For internal feedback we hypothesize for all subcategories that having no reference frame leads to the least amount of internal feedback, as there is no comparator readily available (H3a and H3b). Having a peer reference frame may lead to the most internal feedback generation (H3a and H3c). The criterion reference frame may again be similar to the peer reference frame (H3b and H3d) or to the no reference frame condition (H3c and H3d). This leads to this set of four hypotheses and a complement:

H3a: μPeer RF > μCriterion RF > μNo RF

H3b: μPeer RF = μCriterion RF > μNo RF

H3c: μPeer RF > μCriterion RF = μNo RF

H3d: μPeer RF = μCriterion RF = μNo RF

Hc: μPeer RF, μCriterion RF, μNo RF

For preparatory activities we hypothesize that having no reference frame leads to the least amount of preparatory activities as there is no comparator readily available, leading to the least internal feedback generation (H4a and H4b). Having a peer reference frame may lead to the most internal feedback generation and therefore the most preparatory activities (H4a and H4c). The amount of preparatory activities when a criterion reference frame is available may again be similar to the peer reference frame (H4b and H4d) or to the no reference frame condition (H4c and H4d). This leads to this set of four hypotheses and a complement:

H4a: μPeer RF > μCriterion RF > μNo RF

H4b: μPeer RF = μCriterion RF > μNo RF

H4c: μPeer RF > μCriterion RF = μNo RF

H4d: μPeer RF = μCriterion RF = μNo RF

Hc: μPeer RF, μCriterion RF, μNo RF

3. Results

In total, 6242 segments were coded over all three conditions.

3.1. Internal feedback generation without reference frame

RQ1 concerns the no reference frame condition. Table 1 shows frequencies, mean, and standard deviation (SD) for codes and code categories.

3.1.1. External and internal comparators

In general, participants used three types of comparators. The primary internal comparator was participants' perception of proficiency of that construct ('Perception of construct proficiency'), for example participant 31 verbalized: *"…and I know I'm not good at planning"*. The most used external comparators are dashboard constructs in the LAD ('Dashboard constructs'), verbalized by participant 13 as "… *and that's comparable to the previous score"*. The second most used external comparator is percentages (e.g. 50 %) or proportions (e.g. 'half') ('Percentage/proportion'). This code was used by all participants apart from one. Participant 4 verbalized *"it's 50 %, so that's just in the middle of 100 %"*. Overall, participants used both internal and external comparators to generate internal feedback.

There were some differences between participants, which can be sorted in three patterns. Three participants primarily used the internal comparator 'Perception about construct proficiency comparator' and almost no external comparator (participant 1, 9, and 37). Eight participants used the internal comparator 'Perception about construct proficiency', and external comparators 'Dashboard constructs as comparator', and 'Percentage/proportion' almost equally (participant 10, 13, 16, 22, 25, 31, and 34). Participant 4 primarily used the external comparator 'Dashboard constructs as comparator' and almost no internal comparator.

Table 1
Number of comparators used, internal feedback, and preparatory activities, Mean and SD.

Participant	1	4	10	13	16	19	22	25	31	34	37	Mean	SD
External Comparators													
Other dashboard constructs	1	17	12	9	7	-	8	12	7	18	-	8.3	6.3
Percentage/proportion	-	4	7	7	15	1	1	4	2	6	2	4.5	4.3
New knowledge of construct relevance	-	-	1	-	1	-	-	-	-	-	1	0.3	0.5
Internal Comparator													
Internalized criterium: Grading system	-	-	-	2	2	-	2	-	-	1	1	0.7	0.9
Prior knowledge of construct relevance	-	-	-	-	1	-	-	2	-	-	1	0.4	0.7
Perception of construct proficiency	19	3	9	16	21	14	33	34	10	10	21	17.3	9.7
Perception about peers' proficiency	-	-	-	1	2	-	1	1	-	-	-	0.5	0.7
Feelings about construct	3	-	-	-	2	-	1	5	-	-	-	1.0	1.7
Internal Feedback													
Score judgment: Positive	1	5	11	19	11	6	11	10	17	8	7	9.6	5.2
Score judgment: Negative	1	3	2	2	7	5	1	2	4	11	-	3.5	3.2
Awareness	2	6	5	-	2	-	2	-	1	-	-	1.6	2.1
Affective product: positive	-	1	2	-	-	1	-	-	5	-	-	0.8	1.5
Affective product: negative	-	1	-	-	-	-	-	-	-	-	-	0.1	0.3
Preparatory Activities													
Setting a Goal	-	4	-	-	-	7	-	-	-	-	-	1.0	2.3
Deliberation of (not) using a tactic or strategy	-	3	7	-	-	2	-	-	-	2	4	1.6	2.3
Intention of (not) using a tactic or strategy	-	7	12	-	-	2	-	-	3	1	-	2.3	3.9
Total	27	54	68	56	71	38	60	70	49	57	37	53.4	14.4

Note. For readability purposes, "-" indicates a value of 0.

3.1.2. Internal feedback

Participants' score judgments were mostly positive, negative judgments were less prevalent. Awareness was verbalized by half of the participants and less frequent than score judgments. Affective products were verbalized the least, four participants had a few verbalizations of affective products.

3.1.3. Preparatory activities

Verbalizations of preparatory activities were infrequent. Five participants verbalized no preparatory activities whatsoever. There seems to be individual variance, as participant 4, 10 and 19 verbalized quite some preparatory activities. For example, participant 19 verbalized *"This is something I want to work on within the short term"*, and participant 10 verbalized *"I put the link to the mindfulness training on my to do list for a specific day"*.

Overall, participants from condition 1 used their perception of proficiency as dominant comparator, supplemented with other dashboard constructs and percentages/proportions. For Internal Feedback, Score Judgments were quite prevalent, but Awareness and Affective Products not so much. Some participants verbalized quite some Preparatory Activities, most did not verbalize any.

3.2. Influence of reference frames on internal feedback generation

First, to ensure code frequencies could be compared between conditions, equality of frequency amount was examined. Table 2 shows clear support for the null hypothesis, indicating that conditions are equal in average code frequency.

Then, the average number of codes per code category was calculated per participant using codes for that category. For example, the average for 'Preparatory Activities' was the average of the amount of 'Setting goals', 'Deliberation of using a …', and 'Intention of using a …' (see Appendix A).

Descriptive statistics for each condition on the code categories are shown in Table 3.

3.3. External comparators

For external comparators, Bain ANOVA (Table 4) showed that H1a was most supported. H1b gained support as the fraction increased. Hypotheses H1c and H1d were supported least. Stated otherwise, both H1a and H1b receive substantial support while H1c and H1d did not. This indicates that having a reference frame leads to more verbalizations of external comparators compared to having no reference frame, but it is unclear how a peer reference frame relates to a criterion reference frame.

Table 5 shows code frequencies for external comparators. Participants with the peer and criterion reference frame primarily used those, whereas participants without reference frame used dashboard constructs and percentage/proportions as external comparators.

3.4. Internal comparators

For internal comparators, Bain ANOVA showed most support for H2d, even with higher fractions (Table 6). H2a was supported least. There is some support for H2b and H2c. This indicates that conditions used internal comparators equally.

Table 2
Null hypothesis evaluation of code frequencies between conditions.

	PMP*	PMP**	PMP***
H0: μNo RF = μCrit. RF = μPeer RF	0.839	0.807	0.736
Hc: μNo RF, μCrit. RF, μPeer RF	0.107	0.193	0.264

Note. *, **, *** denotes Fraction of 1, 2, 3. PMPs are based on equal prior model probabilities.

Table 3
Mean (and SDs) for verbalizations per participant from all conditions on use of comparators, internal feedback generated, and preparatory activities.

	C1		C2		C3	
Mean codes per participant	212.72	(52.75)	223.88	(91.20)	192.92	(49.83)
External Comparators	13.00	(8.79)	22.75	(12.49)	20.73	(7.20)
Internal Comparators	19.82	(12.16)	23.25	(12.62)	18.00	(11.70)
Internal Feedback: Score Judgment	6.55	(5.25)	6.94	(6.26)	5.55	(4.96)
Internal Feedback: Awareness	1.64	(2.11)	5.25	(4.89)	1.91	(1.51)
Internal Feedback: Affective products	0.45	(1.14)	1.50	(2.53)	2.23	(3.90)
Preparatory Activities	04.91	(6.69)	10.00	(8.14)	07.27	(6.71)

Table 4
Bain ANOVA, condition on external comparators used.

	PMP*	PMP**	PMP***
H1a: μPeer RF = μCriterion RF > μNo RF	0.515	0.458	0.421
H1b: μPeer RF > μCriterion RF > μNo RF	0.306	0.385	0.433
H1c: μPeer RF > μCriterion RF = μNo RF	0.092	0.081	0.075
H1d: μPeer RF = μCriterion RF = μNo RF	0.056	0.035	0.026
Hc: μPeer RF, μCriterion RF, μNo RF	0.032	0.040	0.045

Note. *, **, *** denotes Fraction of 1, 2, 3. PMPs are based on equal prior model probabilities.

Table 5
Mean and SD for conditions and use of types of external comparators.

	C1		C2		C3	
Dashboard constructs as comparator	8.27	(6.26)	2.75	(3.54)	2.91	(1.70)
Percentage/proportion comparator	4.45	(4.27)	0.88	(1.13)	1.82	(2.99)
New knowledge of construct relevance comparator	0.27	(0.47)	0.25	(0.46)	0.18	(0.40)
Peer reference frame comparator	-	-	18.88	(9.48)	-	-
Criterion reference frame comparator	-	-	-	-	15.82	(5.98)

Table 6
Bain ANOVA, condition on internal comparators used.

	PMP*	PMP**	PMP***
H2a: μNo RF > μCriterion RF > μPeer RF	0.025	0.042	0.054
H2b: μNo RF > μCriterion RF = μPeer RF	0.153	0.179	0.190
H2c: μNo RF = μCriterion RF > μPeer RF	0.100	0.117	0.125
H2d: μNo RF = μCriterion RF = μPeer RF	0.646	0.535	0.465
Hc: μPeer RF, μCriterion RF, μNo RF	0.077	0.127	0.166

Note. *, **, *** denotes Fraction of 1, 2, 3. PMPs are based on equal prior model probabilities.

Table 7 shows code frequency for external comparators. Participants from all conditions used perception of construct proficiency the most.

3.5. Internal Feedback

For Internal Feedback: Score Judgment (Table 8), Bain ANOVA

Table 7
Mean and SD for conditions and use of types of internal comparators.

	C1		C2		C3	
Internalized criterium: Grading system	0.73	(0.90)	0.00	(0.00)	0.09	(0.30)
Prior knowledge of construct relevance	0.36	(0.67)	0.63	(0.92)	0.82	(1.47)
Perception of construct proficiency	17.27	(9.74)	18.25	(10.81)	16.00	(10.60)
Perception of peers' proficiency	0.45	(0.69)	2.50	(3.30)	0.27	(0.65)
Feelings about construct	1.00	(1.67)	1.88	(2.64)	0.82	(1.08)

Table 8
Bain ANOVA, condition on internal feedback subcategories.

Subcategory		PMP*	PMP**	PMP***
Score judgment	H3.1a: μPeer RF > μCriterion RF > μNo RF	0.048	0.077	0.099
	H3.1b: μPeer RF = μCriterion RF > μNo RF	0.111	0.126	0.132
	H3.1c: μPeer RF > μCriterion RF = μNo RF	0.247	0.280	0.292
	H3.1d: μPeer RF = μCriterion RF = μNo RF	0.542	0.435	0.371
	Hc: μPeer RF, μCriterion RF, μNo RF	0.052	0.083	0.106
Awareness	H3.2a: μPeer RF > μCriterion RF > μNo RF	0.271	0.341	0.385
	H3.2b: μPeer RF = μCriterion RF > μNo RF	0.029	0.026	0.024
	H3.2c: μPeer RF > μCriterion RF = μNo RF	0.639	0.569	0.524
	H3.2d: μPeer RF = μCriterion RF = μNo RF	0.020	0.012	0.009
	Hc: μPeer RF, μCriterion RF, μNo RF	0.041	0.052	0.058
Affective products	H3.3a: μPeer RF > μCriterion RF > μNo RF	0.114	0.160	0.190
	H3.3b: μPeer RF = μCriterion RF > μNo RF	0.491	0.485	0.470
	H3.3c: μPeer RF > μCriterion RF = μNo RF	0.075	0.075	0.072
	H3.3d: μPeer RF = μCriterion RF = μNo RF	0.238	0.166	0.132
	Hc: μPeer RF, μCriterion RF, μNo RF	0.082	0.114	0.136

Note. *, **, *** denotes Fraction of 1, 2, 3. PMPs are based on equal prior model probabilities.

showed most support for H3.1d. The data provided least support for H3.1a and H3.1b. H3.1c gained some support as the fraction increased. This indicates that there are most likely no differences of Score Judgments between conditions. For Internal Feedback: Awareness, Bain ANOVA showed most support for H3.2c. The data provided least support for H3.2b and H3.d. H3.2a gained support as the fraction increased. This indicates that participants with a peer reference frame verbalized most Awareness, but it is unclear how a criterion reference frame relates to no reference frame.

For Internal Feedback: Affective Products, Bain ANOVA showed most support for H3.3b. The data provided the least support for H3.3c. H3.3a and H3.3d had some support in the data. This indicates that a peer and criterion reference frame had the most Affective Products. Code frequencies are shown in Table 9.

Table 9
Mean and SD for conditions and Internal Feedback.

	C1		C2		C3	
Score judgment: Positive	9.64	(5.16)	8.50	(6.82)	8.64	(5.16)
Score judgment: Negative	3.45	(3.21)	5.38	(5.63)	2.45	(2.02)
Awareness	1.64	(2.11)	5.25	(4.89)	1.91	(1.51)
Affective product: Positive	0.82	(1.54)	2.50	(3.30)	3.27	(5.10)
Affective product: Negative	0.09	(0.30)	0.50	(0.76)	1.18	(1.89)

3.6. Preparatory activities

For preparatory activities, Bain ANOVA showed no clear support for a specific hypothesis (Table 10). H4b and H4c are continuously supported, H4a gains support with higher fractions, whereas H4d loses support. It may be that participants with a peer reference frame verbalized more preparatory activities than participants without a reference frame, but it's unclear how this relates to availability of a criterion reference frame. Table 11 shows code frequency for preparatory activities.

4. Discussion

In this study, we examined internal feedback generation with a LAD and how presenting reference frames affects this process. This sheds light on use of internal and external comparators, how presenting reference frames affect internal feedback generation, and subsequent preparatory activities. First, the research questions will be answered, followed by overarching insights and future research suggestions.

The first RQ was *"To what extent and how is internal feedback generated whilst using a LAD when no reference frame is available?"*. Participants who were not offered a reference frame (condition 1) mainly used their perception of proficiency as comparator. Other 'Internal mental environment' aspects (Nicol, 2021) were not or scarcely used. The external comparators 'other dashboard constructs' and 'percentages/proportions' were used in addition to the internal comparator. Participants' internal feedback consisted mainly of Score Judgments and were generally positive, Awareness and Affective Products were verbalized infrequently. Verbalizations of Preparatory Activities were scarce as well, although there was reasonable individual variance.

The second RQ was *"How does availability of no reference frame, a peer reference frame, or a criterion reference frame affect internal feedback generation and the comparison process?"*. Similar to the no reference frame condition, participants with a peer reference frame (condition 2) and with a criterion reference frame (condition 3) used their perception of construct proficiency as dominant internal comparator. However, presenting a reference frame greatly reduced the use of other external comparators which were used in the no reference frame condition ('other dashboard constructs' or 'percentages/proportions'). Providing a peer or criterion reference frame also led to more external comparator verbalizations. For internal feedback, the subcategory Score Judgments did not differ between conditions. Condition 2 had the most verbalizations of the subcategory Awareness. For the subcategory Affective products, participants from condition 2 and 3 verbalized these the most. It is unclear if there are differences between conditions for amount of preparatory activities, but there seems to be some evidence that a peer reference frame elicits more preparatory activities than having no reference frame.

Table 10
Bain ANOVA, condition on preparatory activities.

	PMP*	PMP**	PMP***
H4a: Peer RF > μCriterion RF > No RF	0.184	0.256	0.303
H4b: Peer RF = μCriterion RF > No RF	0.257	0.253	0.244
H4c: Peer RF > μCriterion RF = No RF	0.292	0.287	0.278
H4d: Peer RF = μCriterion RF = No RF	0.240	0.167	0.132
Hc: μPeer RF, μCriterion RF, μNo RF	0.027	0.037	0.044

Note. *, **, *** denotes Fraction of 1, 2, 3. PMPs are based on equal prior model probabilities.

Table 11
Mean and SD for conditions and preparatory activities.

	C1		C2		C3	
Setting a Goal	1.00	(2.32)	5.13	(4.88)	2.45	(4.93)
Deliberation of (not) using a tactic or strategy	1.64	(2.29)	1.63	(1.60)	1.09	(0.94)
Intention of (not) using a tactic or strategy	2.27	(3.88)	3.25	(4.43)	3.73	(4.69)

4.1. Influence of reference frames

The first insight pertains to the influence of (not) providing a reference frame in a LAD. If no reference frame is provided as external comparator, participants will use other external comparators instead. If a reference frame is provided, this becomes the dominant external comparator. For generation of internal feedback, the peer reference frame seems to be the favorable choice as it leads to most verbalizations of internal feedback. It possibly leads to most preparatory activities as well. The criterion reference frame may have been to analytical to be used as comparator, as it led to less internal feedback generation and preparatory activities.

An advantage of providing a reference frame is that it gives a clear and exact point of reference, making comparisons more straightforward for dashboard users. Several participants in the no reference frame condition used multiple external and/or internal comparators, for example 'other dashboard constructs', 'percentage/proportion', and 'perception of construct proficiency' (see Table 1). This may result in a less clear conclusion if the comparisons lead to different score judgements. For example, stating *"My 53 % score on Anxiety is well above the peer reference frame"* is a clear interpretation for which the peer reference frame was used. In contrast, stating *"My Anxiety is below Planning, but also above 50 %, and also better than I expected"* is a much less uniform conclusion, as a result of using 'other construct', 'percentage/proportion', and 'perception of proficiency' as comparators. The latter may result in less internal feedback as it is unclear what conclusion to draw from these multiple comparisons. This differs from Nicol and McCallum (2022), who argue that using multiple sources of information may positively affect internal feedback generation. This could be due to differing cognitive processes when making comparisons and generating internal feedback. In Nicol and McCallum (2022), participants explicated their generated internal feedback and how they learned from their comparisons. Participants answered several questions, e.g. "*Which essay is better and why?*" and "*What did you learn from reading the reviews from peers?*". Nicol and McCallum refer to research suggesting that explicating results of cognitive processes has beneficial effects on learning (e.g. Chiu & Chi, 2014; Bisra et al., 2018), and argue that making the output of comparison processes explicit certainly increased the quality of generated feedback for their participants. In the current study, participants were only asked to interpret their visualized scores and received no prompts or questions to explicate their learning from comparisons. It could well be that their cognitive processes were not elaborate enough to combine multiple comparators, leading to (e.g.) aforementioned inconclusive statements. Stimulating this process with questions (as in Nicol & McCallum, 2022) could elicit a more elaborate cognitive process to incorporate multiple comparators.

Also, the peer or criterion reference frames reduced the use of other external comparators. In less elaborate cognitive processes this may lead to clear interpretations but could also prevent the use of other relevant comparative information. For example, using acquired knowledge of construct relevance (Appendix A) alongside a peer reference frame could lead to a more nuanced interpretation, but it may require a more elaborate cognitive processes to combine these information sources. This in turn may lead to different internal feedback, potentially of better quality. When designing a LAD, determining what reference frame(s) to incorporate is complicated. Wise and Vytasek (2017) argue that dashboard users need an "appropriate reference frame" to determine the meaning of a score. Nicol (2021) argues that "appropriate comparisons" are needed for internal feedback generation. What "appropriate" entails is complex. In this study, participants without reference frames used other dashboard constructs as external comparator. In the Motivation and Engagement Wheel (Fig. 1, part 2), constructs pertain to an overarching category per quadrant. Using comparators in the same MEW quadrant (e.g. Planning for Persistence) may be "appropriate", whereas constructs from other quadrants (e.g. Uncertain Control for Persistence) may be less "appropriate". If provided a reference frame, participants use that external comparator to interpret each score separately, which can also be deemed as "appropriate". Dashboard designers should carefully consider what reference frame is "appropriate" for their audience and realize that providing no reference frame leads to use of other external comparators.

4.2. Dashboard relevance and credibility

The second insight pertains to participants' perception of dashboard relevance. Participants in all conditions scarcely used the internal comparator 'construct relevance'. Construct relevance is emphasized in the dashboard ('Why is it important?' Fig. 1, part 6), making this information available as external comparator as well. This may indicate participants' unawareness of constructs relevance, or not incorporating relevance in their thought process. This may explain the low frequencies of Affective Products and Preparatory Activities. Participants do verbalize negative Score Judgments frequently, but if constructs are perceived as irrelevant, these judgments may not lead to an urgency to engage in preparatory activities.

Apart from dashboard relevance, its credibility may also influence internal feedback generation. Multiple participants expressed doubts concerning their visualized scores, coded as 'not-accepting score' and 'questioning dashboard validity' (Appendix A, category 7). When participants' perception of construct proficiency differs from the score in the dashboard, they may retain their perception if the dashboard (and the score) is seen as invalid. Within the Technology Acceptance Model (TAM) (Davis, 1989), a technology user's 'attitude towards using' is a major determinant when it comes to using or rejecting the system (Granić & Marangunić, 2019). This may also affect internal feedback generation, as rejecting a system also means rejecting the feedback it provides. Conversely, if the dashboard and scores in it are deemed credible, this may induce internal feedback generation.

There is a potential role for tutors or student counselors in the process of internal feedback generation using a LAD. Wise et al. (2016) refer to the 'Dialogue/Audience' principle, arguing that having a dialogue with (e.g.) an instructor may have several benefits. A dialogue may increase students' commitment by clarifying construct relevance if students are unaware of this. An instructor may also help interpret scores, how these scores relate to a presented reference frame, generate internal feedback from this comparison, and elicit goal setting and engagement in preparatory activities. An instructor could also suggest using multiple comparators (e.g. 'knowledge of construct relevance') if students only focus on one comparator (e.g. the peer reference frame). Supporting the cognitive process of combining these comparators by asking questions (as in Nicol & McCallum) could be beneficial for students' internal feedback generation. Furthermore, if a reference frame is presented in a LAD, negative emotions may arise (e.g. Lim et al., 2019). Instructors may then help students by discussing these emotions and determining how to move forward.

4.3. Limitations and future research

In this exploratory study we focused on verbalization frequencies of comparators, generated internal feedback, and preparatory activities. A suggestion for future research is to explore possible sequences of internal feedback generation subprocesses. This may shed light on what

subprocesses follow one another in internal feedback generation. It could, for instance, be that 'not accepting' a score prevents further internal feedback generation and preparatory activity engagement, or that a 'score judgment' always precedes 'awareness'.

A second suggestion for future research is exploring the interplay of individual differences, the effects of a reference frame, and internal feedback generation. In this study, differences between conditions were examined, but there was reasonable variance for (i.a.) internal feedback generation as well. This may be due to participants' individual goal-orientation interacting with a reference frame's effect (Beheshitha et al., 2016). Furthermore, the effect of other reference frames on internal feedback generation can be explored, such as a progress reference frame and a top-achiever reference frame (Jivet et al., 2017).

5. Conclusions

This study showed that external comparators will be used regardless of being intended as such. Presenting a reference frame excludes the use of other external comparators. Learners in this study always supplemented an external comparator with an internal comparator, their perception of proficiency. In this study, the peer reference frame led to most internal feedback generation by learners and possibly assists the most in preparatory activities engagement.

Therefore, learners' appraisal in SRL can be supported with a LAD, as it can help generate internal feedback. As learners' SRL-skills may vary, they may need guidance when interpreting scores, generating internal feedback and engaging in preparatory activities. This could be a role for tutors or study counselors, whom may also play a part in clarifying the relevance of constructs within a LAD or offer further support.

Declarations of interest

None.

Appendix A. Coding scheme think aloud interviews

Category	Code	Explanation:	Example:
	Miscellaneous	Relevant to the study but 1) does not fit a category, or 2) is too vague.	"And… It's funny that this comes out"
	irrelevant	Irrelevant or too vague verbalization	"… I mean, that is… In itself just…"
1) Researcher			
	Researcher instruction	Researcher instructs participant	"Please keep thinking aloud"
2) Dashboard use			
	Reads construct explanation	Reading construct explanation.	
	Construct click	Clicking construct in a graph.	
	Construct hover	Hovering construct in a graph.	
	Feedback box hover	Hovering construct feedback box.	
	Scrolling	Scrolling up/down.	
	Hover study progress	Hovering study progress widget.	
	Hover results history	Hovering results history widget.	
	Construct focus	Verbalizing focusing on a construct or part of the dashboard.	
	Reads construct score	Verbalizing construct score	
	Hovers additional support	Hovering additional support.	
	Click additional support	Clicking additional support.	
	Hovers prepare exercise	Hovering prepare exercise.	
	Clicks prepare exercise	Clicking prepare exercise.	
	Hovers act exercise	Hovering act exercise.	
	Clicks act exercise	Clicking act exercise.	
	Hovers reflect exercise	Hovering reflect exercise.	
	Clicks reflect exercise	Clicking reflect exercise.	
3) External environment comparators			
	Peer reference frame	Uses peer reference.	"I see that I'm below average"
	Dashboard constructs	Uses dashboard construct(s).	"And here… planning is lower"
	Criterion reference frame	Uses criterium reference	"… criterium reference is 20 %"
	Percentage/proportion	Uses percentages or proportions as a reference frame, e.g. 50 %, 'the middle'.	"It's all about 75 %"
	New knowledge of construct relevance	Uses construct's relevance	"…it's really important."
4) Internal environment comparators			
Internalized criteria	Internalized criterium: Grading system	Uses Dutch grading system	"Well usually a 5.5. is a passing grade"
	Prior knowledge of construct relevance	Uses known construct (ir)relevance.	"I know this is important"
	Perception about construct proficiency	Uses perception about construct proficiency as a reference.	"I try my best when working in groups…"
	Perception about peers' proficiency	Uses the idea of peers' proficiency as a reference.	"…planning is lower for a lot of other students"
	Feelings about construct	Uses feelings of performing that construct as reference.	"…because I get frustrated in life when I don't get"
5) Internal feedback			
Score Judgment	Positive	Verbalizing positive judgment about score.	"… I think that's a good score."
	Negative	Verbalizing negative judgment about score.	"I think that score is low"
Awareness		Verbalizing awareness/insight related to themselves.	"Apparently I have strong beliefs about my abilities"
Affective Product	Positive	Verbalizing positive feeling.	"…seeing that makes me happy"
	Negative	Verbalizing negative feeling.	"Seeing that is a bit confronting"
6) Preparatory Activities			
	Setting a Goal	Verbalizing goal to achieve.	"…I need to get better at that"
	Deliberation of (not) using a tactic or strategy	Verbalizing deliberation of (not) using tactic/strategy.	"I want to click the prepare button."
	Intention of (not) using a tactic or strategy	Verbalizing intention of (not) using tactic/strategy.	"I'll spend more time at university."

(*continued on next page*)

(*continued*)

Category	Code	Explanation:	Example:
7) Other feedback relevant codes			
	Understanding construct	Understanding (part of) the construct.	"Planning is about planning to study and if you actually do it"
	Not-understanding construct	Not understanding (part of) the construct.	"I don't understand what is meant by valuing"
	accepting score	Verbalizing score acceptance or recognition	"I think this score's accurate. "
	not accepting score	Verbalizing not accepting scores, excuses, or not recognizing.	"This doesn't match me…"
	questioning dashboard validity	Verbalizing doubts regarding (part of) dashboard's validity.	"Maybe I didn't fill out the questionnaire in the right way"

References

Beheshitha, S. S., Hatala, M., Gašević, D., & Joksimović, S. (2016). April). The role of achievement goal orientations when studying effect of learning analytics visualizations. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 54–63).

Bisra, Kiran, Qing, Liu, Nesbit, John C., Salimi, Farimah, & Winne, Philip H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review, 30*(3), 703–725. https://doi.org/10.1007/s10648-018-9434-x

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology, 54*(2), 199–231. https://doi.org/10.1111/j.1464-0597.2005.00205.x

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*(3), 271–315. https://doi.org/10.1207/s15327809jls0603_1

Chiu, J. l, & Chi, M. T. H. (2014). Supporting self-explanation in the classroom. In A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the curriculum* (pp. 91–103). Washington, DC: American Psychological Association. ⟨https://psycnet.apa.org/record/2013-44868-008⟩.

Cohen, J. (1995). The earth is round (p <.05): Rejoinder. *American Psychologist, 50*(12), 1103. https://doi.org/10.1037/0003-066X.50.12.1103

Cohen, L., Manion, L., & Morrison, K. (2018). Methods of data collection: Interviews. p. 506 – 541. In *Research methods in education* (pp. 245–284). Routledge. https://doi.org/10.4324/9780203224342.

Cumming, J., Woodcock, C., Cooley, S. J., Holland, M. J., & Burns, V. E. (2015). Development and validation of the groupwork skills questionnaire (GSQ) for higher education. *Assessment & Evaluation in Higher Education, 40*(7), 988–1001. https://doi.org/10.1080/02602938.2014.957642

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340. https://doi.org/10.2307/249008

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology, 50*(5), 2572–2593. https://doi-org.proxy.library.uu.nl/10.1111/bjet.12864.

Gibson, B. (1997). Talking the test: Using verbal report data in looking at the processing of cloze tasks. *Edinburgh Working Papers In Applied Linguistics, 8*, 54–62.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539. https://doi.org/10.1037/met0000201

Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice. In *European conference on technology enhanced learning* (pp. 82–96). Cham: Springer. https://doi.org/10.1007/978-3-319-66610-5.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage Publications.

Lim, L., Dawson, S., Joksimovic, S., & Gašević, D. (2019). Exploring students' sensemaking of learning analytics dashboards: Does frame of reference make a difference? *Proceedings of the 9th International Conference on Learning analytics & Knowledge*, 250–259. https://doi.org/10.1145/3303772.3303804

Marzouk, Z., Rakovic, M., Liaqat, A., Vytasek, J., Samadi, D., Stewart-Alonso, J., & Nesbit, J. C. (2016). What if learning analytics were based on learning science? *Australasian Journal of Educational Technology, 32*(6). https://doi.org/10.14742/ajet.3058

Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology, 77*(2), 413–440. https://doi.org/10.1348/000709906X118036

JASP Team (2023). *JASP (Version 0.16.3) [Computer software]*. https://jasp-stats.org/faq/how-do-i-cite-jasp/.

Martin, A.J. (2016). *The Motivation and Engagement Workbook* (16th Edition). Sydney, Australia: Lifelong Achievement Group (www.lifelongachievement.com).

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*(3), 679–690. https://doi.org/10.3758/s13428-010-0049-5

Miller-Young, J. E. (2013). Calculations and expectations: How engineering students describe three-dimensional forces. *Canadian Journal for the Scholarship of Teaching and Learning, 4*(1), 4. https://doi.org/10/5206/cjsotl-rcacea.2013.1.4.

Nicol, D. (2019). Reconceptualising feedback as an internal not an external process. *Italian Journal of Educational Research*, 71–84. https://doi.org/10.7346/SIRD-1S2019-P71

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education, 46*(5), 756–778. https://doi.org/10.1080/02602938.2020.1823314

Nicol, D., & McCallum, S. (2022). Making internal feedback explicit: Exploiting the multiple comparisons that occur during peer review. *Assessment & Evaluation in Higher Education, 47*(3), 424–443. https://doi.org/10.1080/02602938.2021.1924620

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*. https://doi.org/10.1177/1609406919899220

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In *Handbook of self-regulation* (pp. 451–502). Academic Press. https://doi.org/10.1016/B978-012109890-2/50043-3.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353.

Roll, I., & Winne, P. H. (2015). Understanding, evaluating, and supporting self-regulated learning using learning analytics. *Journal of Learning Analytics, 2*(1), 7–12. https://doi.org/10.18608/jla.2015.21.2

Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., & Dillenbourg, P. (2016). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies, 10*(1), 30–41. https://doi.org/10.1109/TLT.2016.2599522

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2016). *False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.* ⟨https://doi.org/10.1037/14805-033⟩.

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., & Klerkx, J. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing, 18*(6), 1499–1514. https://doi.org/10.1007/s00779-013-0751-2

Viberg, O., Khalil, M., & Baars, M. (2020). Self-regulated learning and learning analytics in online learning environments: A review of empirical research. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 524–533. https://doi.org/10.1145/3375462.3375483

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wicherts, J. M., Veldkamp, L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, A. L. M. (2016). Degrees of freedom in planning, analyzing, and reporting psychological studies; A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wise, A. F., & Vytasek, J. (2017). Learning analytics implementation design. *Handbook of Learning Analytics, 1*, 151–160. https://doi.org/10.18608/hla17

Wise, A. F., Vytasek, J. M., Hausknecht, S., & Zhao, Y. (2016). Developing learning analytics design knowledge in the" middle space": The student tuning model and align design framework for learning analytics use. *Online Learning, 20*(2), 155–182.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Winne, P. H., and Hadwin, A. F. (1998). "Studying as self-regulated engagement in learning," in Metacognition in Educational Theory and Practice, eds D. Hacker, J. Dunlosky, and A. Graesser (Hillsdale, NJ: Erlbaum), 277–304.