Item-level monitoring, response style stability, and the hard-easy effect

Roy B. Clariana (RClariana@psu.edu)
Pennsylvania State University https://orcid.org/0000-0001-9374-0064
Eunsung Park
Pennsylvania State University University Park : Penn State

Research Article

Keywords: self-regulated learning, response confidence, response style stability, hysteresis, hard-easy effect, Coh-Metrix

Posted Date: February 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-256756/v1

Version of Record: A version of this preprint was published at Educational Technology Research and Development on March 26th, 2021. See the published version at https://doi.org/10.1007/s11423-021-09981-8.

Item-level monitoring, response style stability, and the hard-easy effect

Abstract

Cognitive and metacognitive processes during learning depend on accurate monitoring, this investigation examines the influence of immediate item-level knowledge of correct response feedback on cognition monitoring accuracy. In an optional end-of-course computer-based review lesson, participants (n = 68) were randomly assigned to groups to receive either immediate item-by-item feedback (IF) or no immediate feedback (NF). Item-by-item monitoring consisted of confidence self-reports. Two days later, participants completed a retention test (IF = NF, no significant difference). Monitoring accuracy during the review lesson was low, and contrary to expectations was significantly less with immediate feedback (IF < NF, Cohen's d = .62). Descriptive data shows that (1) monitoring accuracy can be attributed to cues beyond actual item difficulty, (2) a hard-easy effect was observed where item difficulty was related to confidence judgements as a non-monotonic function, (3) response confidence was predicted by the Coh-Metrix dimension *Word Concreteness* in both the IF and NF treatments, and (4) significant autocorrelations (hysteresis) for confidence measures were observed for NF but not for IF. It seems likely that monitoring is based on multiple and sometimes competing cues, the salience of each relates in some degree to content difficulty, but that the stability of individual response styles plays a substantive role in monitoring. This investigation shows the need for new applications of technology for monitoring multiple measures on the fly to better understand SRL processes to support all learners.

Keywords: self-regulated learning, response confidence, response style stability, hysteresis, hard-easy effect, Coh-Metrix

Item-level monitoring, response style stability, and the hard-easy effect

Jonassen (2000) has noted that metacognitive skills enable people to be strategic, and that these skills can be learned. Self-regulation is a critical metacognitive life skill that is central for individual wellbeing and for achieving success of both immediate and long-term goals (Bandura, 1991). Self-regulated learning (SRL) is an educationally important subset of self-regulation that includes setting and updating learning goals, ongoing planning, monitoring, occasional strategy-shifts, and ongoing progress evaluation (Butler & Winne, 1995; Panadero, 2017). Monitoring is a central aspect of SRL and is the emphasis of this experimental investigation, because improving technology-supported monitoring in SRL is important for designers, educators, and researchers (Brady, Rosenthal, Forest, & Hocevar, 2020; Kavousi, Miller, & Alexander, 2020; Reid, Morrison, & Bol, 2017; Zhu, Bonk, & Doo, 2020).

Panadero (2017) reviewed six actively researched models of SRL including those of Boekaerts; Efklides; Hadwin, Järvelä and Miller; Winne and Hadwin; Pintrich; and Zimmerman. Although these models differ in various ways, all rely on learners' ability to gauge their understanding, for example as judgements of learning (JOLs), in order to select and use appropriate cognitive and metacognitive strategies (Reid, Morrison, & Bol, 2017). But Nelson and Dunlosky (1991) note that, "… the nearly universal finding of the literature … is that accuracy of JOLs is far from perfect, in fact closer to nil than to perfect" (p. 267). Because of the fundamental relationship between monitoring accuracy and good decision making in SRL, monitoring accuracy is important and requires further research and thought (Bol & Hacker, 2012; Hacker, Bol, & Bahbahani, 2008).

A vital monitoring task asks, *how am I doing now*? Because some idea units are easier and some are more difficult, the actual and the perceived difficulty of lesson materials likely varies from moment-to-moment so collecting monitoring data more frequently, for example at the individual idea level, could improve monitoring accuracy (Butler & Winne,1995; Dunlosky & Lipko, 2007; Hartwig & Dunlosky, 2017).

Monitoring accuracy can be even better when accompanied by performance feedback that shows whether the monitoring judgement matches actual performance or not, and this feedback may then improve the subsequent monitoring judgements (Händel, Harder, & Dresel, 2020; Nietfeld, Cao, & Osborne, 2005, 2006). Dunlosky and Rawson (2015) reviewed the prior literature on SRL with feedback, with a focus on using test-like events during lesson review of previous course content. They report that feedback during review lessons improves memory and final course performance for that specific course content, transfers to comprehension and application outcomes, and *influences decision choices* in the lesson (Merriman, Clariana, & Bernardi, 2012).

A major goal of monitoring research is to find the basis of students' judgments by asking, "What factors are students making inferences about when they construct their judgments?" (Dunlosky & Thiede, 2013, p. 59). Monitoring as a manifestation of perceived difficulty may be inaccurate since it is likely to be based on multiple cues (Hertzog, Hines, & Touron, 2013) such as memory of past test performance (Finn & Metcalfe, 2007, 2008). Attention to multiple conflicting cues or to no cues is likely to establish monitoring inertia, a form of measurement bias referred to as the *stability of individual response styles* (SIRS, Javaras & Ripley, 2007; Weijters, Geuens, & Schillewaert, 2010). SIRS judgements may be an individual trait variable that remains fairly constant or stable over time "…despite external influences and direct

evidence to the contrary" (Kornell & Bjork, 2009, p. 460; also see Bjork, Dunlosky, & Kornell, 2013).

This current investigation examines the influence of immediate feedback on monitoring accuracy measured as item-by-item confidence self-reports (Dunlosky & Lipko, 2007). The central question is, *will immediate item-level feedback improve monitoring accuracy relative to a no immediate feedback control?* Most SRL studies only collect one or a few pieces of monitoring data typically at the start or at the end of a lesson or unit of instruction. Such sparse monitoring data does not allow for richer descriptive analyses, such as the extent of the relationship between the difficulty of an idea unit and learners' judgments of that idea unit, nor does it allow comparison of confidence across time or of the influence of previous confidence on future confidence (e.g., as hysteresis, the influence of a recent past outcome on a current outcome). Thus this investigation adds the following descriptive / exploratory questions, *is item level confidence non-monotonically related to item level difficulty (hard-easy effect)*? *Is an individual's item level confidence stable across responses within the review lesson (SIRS)*? *Do item level confidence responses influence follow-on confidence responses (hysteresis)? What immediate context cues influence confidence judgements (as Coh-Metrix item-level text features)?* Monitoring accuracy and the literature bases of each of these four exploratory questions are provided next.

Monitoring Accuracy in SRL

Monitoring accuracy is the degree of fit between a person's judgment of performance towards a goal and his or her actual performance related to that goal (Bol & Hacker, 2012; Keren, 1991). Fleming and Lau (2014) point out that cognitive

psychologists as early as 1885 were interested in how well people could monitor their own knowledge, and that confidence ratings were a mainstay of such analysis. Several broad areas can be monitored within the five SRL phases of setting learning goals, planning, monitoring, strategy-shifts, and progress evaluation (reflection) including regulation of cognition, of motivation, of behavior, and of context (Greene & Azevedo, 2007; Pintrich, Wolters, & Baxter, 2000). Even after many decades of research, Rutherford (2014) has noted, "although calibration is a growing area of research within Educational Psychology, unanswered questions remain about the nature of calibration: how it should be measured, its role as a dynamic aspect of metacognition, and how best to improve it." (p. xv). Of most interest in this investigation is cognition monitoring accuracy (i.e., or calibration).

Various interventions including providing guidance, practice, clear criterion of performance expectations, and feedback, have been proposed to reduce gaps between perception of performance versus actual performance (Burson, Larrick, & Klayman, 2006; Händel at al., 2020; Nietfeld et al., 2005, 2006; Stone, 2000). Bandura (1991) notes that the informativeness of performance feedback is necessary for students to have a clear idea of how they are doing (p. 251). Combining feedback and certitude-based JOLs have been actively researched for at least 40 years (Kulhavy, Yekovich, & Dyer, 1976; Vasilyeva, Pechenizkiy, & De Bra, 2008), but these studies have mainly considered how certitude can influence the effectiveness of feedback (Stock, Kulhavy, & Pridemore, 1992), but not the reverse, how feedback influences ongoing certitude and thus influences monitoring accuracy. Extending this to SRL settings, Nugteren, Jarodzka, Kesterm, and van Merriënboer (2018) propose that "Future studies could therefore incorporate feedback to improve the students' self assessments…." (p.375).

Content micro-to-macro level grain size is also an important consideration for cognition monitoring. It is not the same to ask a student *how well will you do in this course* or *how well will you do on this exam* (global monitoring) versus *how well will you do on this exam item* (local monitoring), and so on (Hartwig & Dunlosky, 2017). "In comparison to macro-level analysis, item-by-item analysis allows a more detailed, and likely more accurate, view of the process of forming metacognitive judgments" (Rutherford, 2014, p. 22). Butler and Winne (1995) comment that –

> in general, research investigating feedback and self-regulation has focused on behaviors at too large a grain size - for example, studying whole passages or answering sets of test items after studying is over - and has thereby collected data that fail to reflect the variance in behavior that is regulation. (p. 246)

Dunlosky and Lipko (2007) note, "Will even more specific judgments (e.g., at the level of individual idea units in each concept) serve to further reduce overconfidence and support even higher levels of accuracy?" (p. 231). Thus, it is not yet clear from the research base how item level monitoring with feedback influences monitoring accuracy.

Multiple cues that may influence monitoring accuracy beyond actual difficulty

Actual measured difficulty and perceived difficulty are not identical measures. Review of the literature point to several areas that consider perceived and actual difficulty. Prior familiarity could influence judgements, but when such information is incomplete, what other cues influence learner's perceptions of content difficulty? Specifically in this section, first, the hard-easy effect is presented to provide one account for past observed mixed and no difference findings for monitoring accuracy. Next, stability of individual response style (SIRS) observed in traditional survey research is extended here to item-level confidence self-report measures as an exploratory

and descriptive measure related to monitoring accuracy. Because multiple confidence measures are collected over time, the possible influence of confidence responses on subsequent confidence responses (hysteresis) is considered. And last, Coh-Metrix text features of the lesson items are examined as a context cue that can influence perceived difficulty and thus influence monitoring.

Hard-easy effect

A premise of this investigation and perhaps of all SRL investigations is that of a monotonic relationship between the observed difficulty of the lesson content based on actual performance scores and the learner's judgement of the difficulty of that content – that difficult material is perceived to be difficult and easy material is perceived to be easy. But previous SRL research shows that students tend to be overconfident on difficult things and under confident on easy things, the hard-easy effect (Arnold, Graham, & Hollingworth-Hughes, 2017; Juslin, Winman, & Olsson, 2000). So what is perceived difficulty? At a minimum, confidence perceptions of difficulty depend on multiple cues including familiarity, prior knowledge, characteristics of the text, characteristics of the items, guessing, and combinations of these (Dinsmore & Parkinson, 2013). The cues attended to will influence judgements of difficulty and confidence.

Reid et al. (2017) note that although most learners appear to be poorly calibrated, high-achieving students are generally more accurate than low achievers at estimating their performance, but also high performing students are more likely to be under confident than low performing students (Bol, Riggs, Hacker, & Nunnery, 2010; Grabowski, 2004; Kruger & Dunning, 1999; Kruger & Mueller, 2002; Rutherford, 2014; Stone, 2000). For example, Hacker, Bol, Horgan, and Rakow (2000) in a semester-long course placed students into 5 groups based on ability and asked them to make prediction and postdiction judgements of their performance on

successive multiple-choice exams. At exam 1 monotonic and nearly linear relationships between actual group-level performance and anticipated performance whether elicited before (predicted) or after completing the exam (postdicted) were observed (see Figure 1), but by the third exam several weeks later, predicted and postdicted judgements were non-monotonic, and postdicted judgements had become considerably more overconfident than predicted judgements. The lower ability groups clearly show substantial overconfidence with both predicted and postdicted judgements.

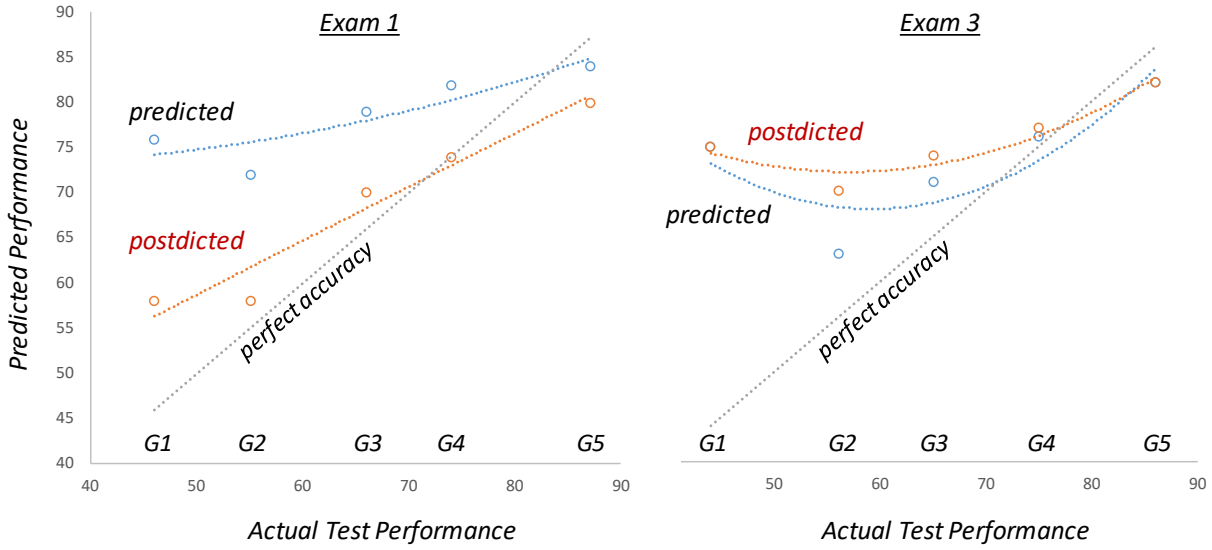

Figure 1. Group-level examination performance of five ability groupings (G1-G5) with predicted and postdicted performance judgements at two time periods (from Hacker at al., 2000).

Monitoring inertia due to stability of individual response style and hysteresis

Stability of individual response styles (SIRS) observed in survey research has been shown to be stable over the course of a single questionnaire administration (Javaras & Ripley, 2007) and is both a within survey as well as a longitudinal phenomenon.

Weijters, Geuens, and Schillewaert (2010) surveyed Belgian members of the public (n = 604) with non-overlapping surveys a year apart, the first survey asked items such as "Television is my primary form of entertainment" and "Air pollution is an important worldwide problem" while the second survey asked such questions as "I understand myself" and "The things I possess are not that important to me." Weijters et al. reported four Likert-scale response styles that to a large extent appear to be a stable individual characteristics, giving either mainly positive, negative, middling, or extreme responses. Monitoring judgements, such as predicting an exam score, are basically survey type responses, thus an individual's confidence responses may exhibit stability within and across trials (Bjork et al., 2013). In other words, individuals' confidence responses vary around their individual set point, and so these measures would be consistent across their own responses and thus consistently higher or lower relative to the group average. This is biologically plausible since in living systems there are numerous examples of such closed-loop control systems (Illingworth, 2011).

Similarly, Hacker et al. (2000) reported that judgments of performance were influenced by prior judgments, and so besides an individual response stability component (SIRS), confidence responses may also be influenced by recent previous confidence responses. This is also referred to as response history (Hertzog, Hines, & Touron, 2013), or hysteresis, defined as the dependence of the current state of a system on its recent history. For example, confidence response has been described as biased towards previous responses, where "recent confidence represents a mental shortcut (heuristic) which informs self-reflection when more relevant information is unavailable" (Benwell, Beyer, Wallington, & Ince, 2020, p. 18). In this present investigation, item-by-item confidence measures are examined for hysteresis using SPSS Forecasting-Autocorrelation of the confidence measures.

Coh-Metrix measures of test item cues

Besides actual measured item difficulty, confidence judgements can likely be influenced by the immediately available text features (Mills, D'Mello, & Kopp, 2015), for example Weaver and Bryant (1995) reported that metacognitive accuracy for reading comprehension depended on text readability level (Kelemen, Frost, & Weaver, 2000, p. 93). A multilevel theoretical framework of text features has been operationalized as Coh-Metrix that builds from and considerably extends the early readability measures such as Flesch, Dale-Chall, Gunning, and SMOG formulas (Dowell, Graesser, & Cai, 2016). The Coh-Metrix analysis tool is a well-established and well-researched computational tool that amalgamates a number of previously separate measures of linguistic and discourse features of a text. Coh-Metrix provides more than one hundred measures of a text in five discrete dimensions: Narrativity, Syntactic Simplicity, Word Concreteness, Referential Cohesion, and Deep Cohesion. Which of these measures and dimensions are salient here? Of these five dimensions, *Word Concreteness* is the only one that aligns with the review items in this investigation, because this lesson content does not have narrativity, the syntactic form of items has no variability because the item format is standard across all of the items, and the items are too brief to exhibit referential or deep cohesion.

Most Coh-Metrix investigations typically use long text portions, but a study by Walkington, Clinton, and Shivraj (2018) used Coh-Metrix measures of sentence-long mathematics word problems from 20 years of archived test data from the National Assessment of Educational Progress (NAEP). They used pilot studies to narrow down to four Coh-Metrix measures, specifically the *Word Concreteness* dimension plus three individual measures including word count, pronoun density, and presence of second person pronouns. These measures were related to performance in various ways, for

example students who were weaker in mathematics tended to benefit more from factors such as word concreteness that make the math problems easier to read (pp. 403-404). Following Walkington et al. (2018), this present investigation seeks to determine the influence of the Coh-Metrix item-level text measures on response confidence using two of these measures, word count and *Word Concreteness*. Note that pronoun density and presence of second person pronouns used by Walkington et al. could not be considered because the review items in this present investigation contain only three pronouns.

Purpose

This experimental investigation with random assignment to treatment examines the influence of item-by-item immediate knowledge of correct response feedback compared to no immediate feedback (control) on cognition monitoring accuracy in an end-of-course review lesson. There is only one question, does immediate feedback improve monitoring accuracy? But in addition, to reconcile past mixed findings, descriptive analysis (1) seeks evidence of non-monotonicity as the hard-easy effect, (2) of the influence of response stability (both SIRS and hysteresis), and (3) the possible role of selected Coh-Metrix text features on confidence responses.

Method

Participants

Students (n = 68) were a sample of convenience from an instructional design program of an eastern U.S. university who voluntarily participated in this experimental investigation, most of the participants self-reported as female (78%) and most were working professionals (78%, the remainder were full time students).

Materials and Procedures

This review lesson approach follows that of a series of SRL investigations by Dunlosky and colleagues who used definitions of key concepts in an introductory undergraduate psychology course that were provided as a computer-delivered review towards the end of the course (e.g., Dunlosky & Rawson, 2012, 2015). The content of this review lesson consisted of instructional design terminology from the course textbook. The lesson items were arranged in the order of primary occurrence in the textbook.

Students could drop in to any campus computer lab at any time to complete this unmonitored review lesson. After logging in to the lab computer, the software randomly assigned them to the IF or NF treatment group. Students under IF completed the end-of-course review lesson with immediate item-by-item feedback (n = 31), while students under NF received the same items in the same order but with no immediate feedback (n = 37). There was no participant mortality, the unequal group sizes are due to true random assignment by the software without regard to past assignment to group. Item responses included providing a confidence judgement and then selecting the term that matches the definition from four alternatives. Confidence judgements had 5 levels ranging from "I am just guessing" to "I am about 50% confident that I know the answer" to "I am certain I know the answer".

Both IF and NF here are classic knowledge of correct response (KCR) feedback that informs the learner of the correct answer to a specific problem by displaying the question and correct response with no additional information (Clariana, 1990; Clariana, Ross, & Morrison, 1991). For IF the correct answer was provided immediately after the response entry, while for the NF treatment, the correct answer feedback was given in

mass at the end of that lesson section (i.e., thus NF is actually delayed feedback). The review lesson also offered a second section that covered other course content, the second section is not included in this current analysis but that data is reported in a separate investigation by Follmer and Clariana (2020). Two days later, participants completed a paper-based retention test. The lesson and retention test data were matched for analysis using the user-created logins that were known only by the participants.

Results

The results are presented in two sections, the first section includes the individual-level cognitive monitoring accuracy data and the retention test performance. The second section presents the review lesson data descriptive analyses to consider relationships between confidence and performance (e.g., the hard-easy effect), the potential downstream influence of response confidence (hysteresis), and stepwise multiple regression to consider how Coh-Metrix measures of text features can predict lesson confidence responses beyond actual item difficulty.

Individual Participant-level Results

Cognitive Monitoring Accuracy

Participants' monitoring accuracy in the review lesson was calculated as the relationship between item-level confidence (as 1, 2, 3, 4, 5) and lesson-item performance (as 0, 1) data for each item. Because this confidence data is collected as ordinal-level data, there is a continuing debate regarding which parametric or non-parametric test may be used with ordinal data in these kinds of studies. Typical analysis uses Spearman *rho*, Goodman's *gamma*, or Receiver Operating Characteristic (ROC, as area under the curve). Following the analysis approached used by Bol and Hacker (2012) and Lin and Zabrucky (1998), this study utilized Spearman *rho*.

Individuals' relative monitoring accuracy data were calculated as the Spearman *rho* value of an individual's confidence (1-5) with their actual performance (0, 1) in the review lesson; thus each participant has 18 monitoring accuracy values, one for each item. The observed Cronbach's alpha reliability for this monitoring data is KR-20 = .96. As in most studies of this type, some individuals were excluded from the correlation analysis due to correlation indeterminacy, for example by having no variance in item confidence replies (e.g., gave a 4 for every item). Four participants were excluded due to indeterminacy, the individual-level analysis sample size for NF is reduced from $n = 37$ to $n = 35$ and for IF is reduced from $n = 31$ to $n = 29$. Because Spearman *rho* values are not interval level data, before averaging, these *rho* values were converted to Fisher's r-to-Z transformation values that are interval level measures. The averaged individual-level monitoring accuracy as Fisher's Z-transformation data are: NF (control) M = 0.27, SD = .30 and IF (treatment) M = .10, SD = .21. The data were analyzed by analysis of variance, with one between subjects factor Feedback (IF or NF). Levine's test was not significant, p = .126. Feedback was significant, $F(1,62) = 6.629$, MSe = .068, $p = .012$, eta square = .097. (i.e., NF > IF, Cohen's $d = 0.62$). Although the mean difference between NF and IF is significant, monitoring accuracy was very low for both NF and IF groups.

Retention Test Performance

The retention test given two days after the computer-based review lesson consisted of the same 18 four-alternative multiple-choice items as the review lesson, but the items were randomly ordered. The observed Cronbach's alpha reliability for the retention test is KR-20 = .77. Lesson and retention test performance for IF (n = 31) were M = 0.78 and M = 0.83, SD = .11, improvement ES = .45 ($p = .003$) and for NF (n = 37) were M = 0.79 and M = 0.86, SD = 0.12, improvement ES = .58 ($p = .005$), indicating

that students learned from both the IF and NF lesson and did well on the retention test. The performance data from the retention test were analyzed by analysis of variance with the between-subjects factor Treatment (NF or IF). Levine's test equality of error variances was not significant, p = .617. No main effect was observed, F(1,67) = 1.339, MS = .014, p = .251, eta square = .020. The IF and NF (delayed feedback) retention test means were statistically equivalent, which a normal finding in the literature (Clariana et al., 1991).

Item-level Descriptive Analyses of Lesson Data

The Hard-Easy Effect

For descriptive purposes, confidence rank data is treated as close approximations of interval data (from Tekin & Roediger, 2017). To convert item-level confidence ranks to interval data, we estimated as follows: for the 1-5 scale with these descriptions where 1 is "I am guessing", 3 is "I am about 50% confident that I know the answer", and 5 is "I am certain I know the answer", then 1 = 0%, 2 = 25%, 3 = 50%, 4 = 75%, and 5 = 100% (see Table 1).

The item-level data scatterplot of item difficulty (item *P*) and item confidence were prepared to check for linearity / monotonicity (Laerd, 2018; see Figure 2). The calibration ideal is represented by the perfect accuracy diagonal line, the classic hard-easy effect is apparent for both the NF and IF treatments, with overconfidence for difficult items (left side of Figure 2) and under confidence with easy items (right side of Figure 2).

Table 1. Review lesson item means and Coh-Metrix measures.

Lesson		NF (n = 37)		IF (n = 31)		Word	Word
Order	Vocabulary	Conf.	Perf.	Conf.	Perf.	Concrete.	Count
1	System	0.80	0.86	0.82	0.97	4.492	11
2	Systems approach	0.72	0.59	0.72	0.61	1.215	22
3	Model	0.74	0.95	0.76	0.94	4.192	18
4	Needs assessment	0.83	0.89	0.86	0.94	-3.299	15
5	Need	0.84	0.92	0.89	0.94	-0.567	13
6	Intellectual skill	0.72	0.81	0.76	0.81	3.281	20
7	Hierarchical analysis	0.61	0.68	0.65	0.61	4.139	24
8	Subordinate skill	0.85	0.43	0.77	0.61	-0.419	21
9	Instructional analysis	0.50	0.68	0.65	0.74	5.422	28
10	Learner analysis	0.89	0.95	0.85	0.97	-0.166	33
11	Target population	0.78	0.86	0.83	0.87	2.559	11
12	Learning context	0.72	0.41	0.73	0.32	-0.536	17
13	Terminal objective	0.75	0.92	0.81	0.90	-1.079	31
14	Subordinate objective	0.68	0.95	0.72	0.81	-2.042	20
15	Posttest	0.92	0.95	0.84	0.94	-1.287	27
16	Preinstructional activities	0.75	0.86	0.75	0.81	0.976	43
17	Media	0.77	0.54	0.73	0.39	1.452	16
18	Formative	0.77	0.92	0.76	0.87	4.89	24

Conf. is item confidence elf-report, Perf. is measured difficulty as item P

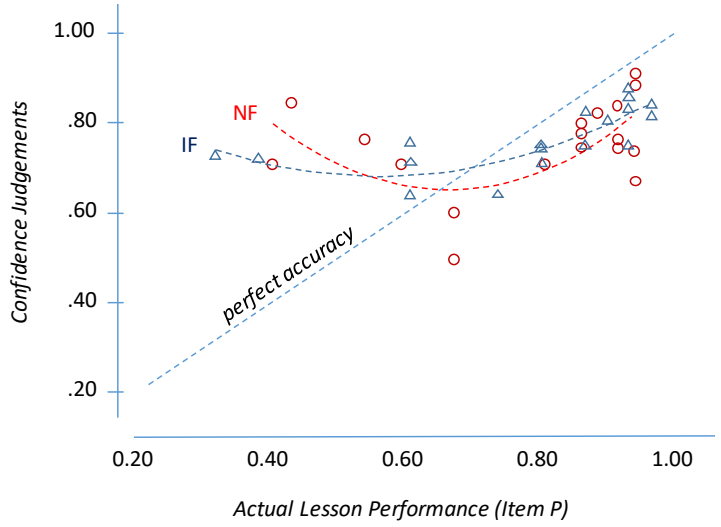

Figure 2. Scatterplots of the actual lesson performance average lesson item difficulties (item *P) and* item confidence for the NF and IF treatments.

The inflection points of the two curves are consistent with Lichtenstein and Fischhoff (1977) who reported overconfidence for relatively more difficult items below .75 item difficulty and under confidence for relatively easier items above .75 item difficulty. Notice the fairly linear relationship between confidence judgements and item difficulty for easy items greater than .75 item difficulty. A nonlinear, non-monotonic relationship between confidence and item difficulty is statistically problematic for determining cognition monitoring accuracy. Perhaps the fairly low values observed here for monitoring accuracy and in previous studies can be attributed to item difficulty of the lesson content, which may be a common phenomenon for studies that use content than spans a range of difficulty.

Descriptive Charts of the Lesson Data

Sequence charts of the item-level performance and confidence self-reports are provided in Figure 3. There is substantial similarity between the IF and NF groups' item-level confidence (IF-NF confidence, Pearson correlation r = .85) as well as their item performance (IF-NF item P, Pearson correlation r = .91). Inspection of the averaged item-level difficulty and item confidence measures indicates that the IF treatment (see top of Figure 3) and the NF treatment (see bottom of Figure 3) are quite alike. An item that was difficult for the NF group was also difficult for the IF group, and in the same way a low confidence response for an item by the NF group was also a low confident response for the IF group. Also within both IF and NF treatments there is a noticeable pattern between confidence and performance, with the confidence measures being less extreme than the performance measures (i.e., confidence nearer to the .76 average confidence value horizontal line).

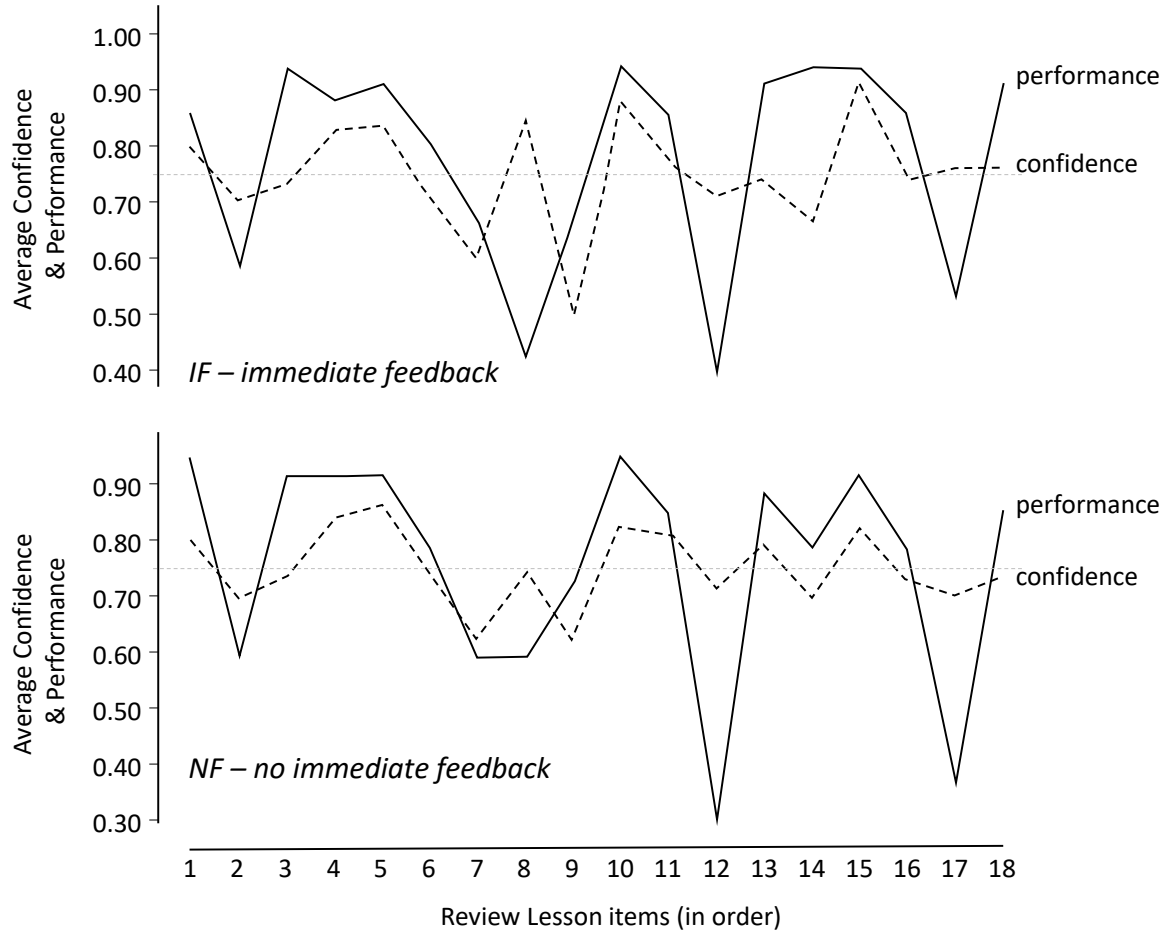

Figure 3. IF (top) and NF (bottom) treatments' review lesson average item-level difficulty (solid lines), item-level confidence (dashed lines).

Hysteresis and Autocorrelation

Does confidence at a response moment influence confidence for later items, and if yes, how far downstream (Hacker et al., 2000)? Since the 18 items occurred in a sequence, the average confidence measures for each item displayed in Figure 3 are a sequence chart. Separate autocorrelation analyses of item *P* data and of item confidence data were conducted with SPSS version 25 using Forecasting analyses. After inspection of the sequence charts, it was determined to use the typical analysis assumptions of 0 difference (i.e., no trend detected) and not to use the natural log transform. Analysis of

the ACF and PACF curves and the autocorrelation values for up to 4 lags indicate *no autocorrelation* for any item difficulty (*P*) series. But for the item confidence data, a significant Lag 1 was observed for the NF treatment (Lag 1, $r = -0.45$, $p = .03$) but not for the IF treatment (see Table 2).

Table 2. Autocorrelation analysis of confidence self-report measures for each treatment.

	NF (control)					IF (immediate feedback)				
			Box-Ljung Statistic					Box-Ljung Statistic		
Lag	*r**	SE	Value	df	Sig.	*r**	SE	Value	df	Sig.
1	**-.45**	**.22**	**4.37**	**1**	**.03**	-.07	.22	.10	1	.74
2	.10	.21	4.63	2	.09	-.14	.21	.57	2	.75
3	-.04	.20	4.68	3	.19	-.23	.20	1.82	3	.61
4	-.16	.19	5.31	4	.25	-.26	.19	3.51	4	.47

Note: r autocorrelations*

These autocorrelation values support that the item confidence exhibited a downstream influence under no immediate feedback (NF), but not with immediate feedback (IF). The negative Lag 1 autocorrelation for the NF (control) treatment suggests confidence response stability around an average setpoint where a confidence response influences the next confidence response. But with immediate feedback (IF), there is no observed measurable lag (hysteresis) in the ongoing confidence responses.

Coh-Metrix Text Features and Item-level Confidence

Stepwise multiple regression was used to predict the role of actual observed item difficulty (item *P*) and of selected text features on confidence of response. The web-based tool Coh-Metrix (version 3.0, http://tool.cohmetrix.com/) was used to calculate the text features of each of the review lesson items. Then the group-level lesson item difficulty (*P*) along with the two selected text features, *Word Concreteness* (a dimension) and word count (a measure), were

used to predict item-level confidence (see Table 1). Two separate stepwise multiple regressions were conducted, one for the IF treatment and one for the NF treatment.

For the IF treatment, two predictor variables entered the regression equation that significantly predicted item-level confidence (2,15) = 11.734, $p < .001$. As would be expected, the actual lesson item difficulty (item *P*) entered the analysis at Step 1 ($R^2 = .380$). Next, *Word Concreteness* z-score entered at Step 2 adding a substantial amount of accounted variance ($R^2 = .610$). Item word count did not enter the regression model. The regression equations for predicting item-level confidence for both immediate feedback (IF) and no feedback (NF) are:

IF confidence = (.217 x Item Difficulty) + (-.013 x Word Concreteness) + .618

NF confidence = (-.019 x Word Concreteness) + .783

For the NF (control) treatment, only one predictor variable entered the regression equation, *Word Concreteness* z-score ($R^2 = .258$) that significantly predicted item-level confidence (1,16) = 5.573, $p = .031$, but contrary to expectations, the actual NF lesson item difficulty (item *P*) *did not* enter the regression (t = 1.215, $p = .243$). The *Word Concreteness* dimension regression component was included and was negative in both models but for NF, text surface features are a better predictor of confidence response that was item difficulty.

Discussion

This investigation examined item-level monitoring with immediate item-level feedback that allows for local idea-level reflection and recalibration that should improve monitoring accuracy judgements. Contrary to expectations, immediate feedback did not improve monitoring accuracy; monitoring accuracy was low with or without immediate feedback, and monitoring accuracy with immediate feedback was significantly less accurate than with no feedback (IF < NF, Cohen's $d = .62$, $p = .012$). A non-monotonic

relationship between item performance and confidence was observed (hard-easy effect, Arnold et al., 2017; Juslin et al., 2000) that attenuates the findings here for relative monitoring accuracy (see Figure 2).

The item-level descriptive data show that cognitive monitoring accuracy in this investigation depends on both internal cues (i.e., recent monitoring responses) and external cues (i.e., text features, feedback). It is reasonable and even likely that item-level monitoring is based on multiple and sometimes competing cues, the salience of each cue relates in some degree to content difficulty as well as to text features of the materials. Also stability of individual response styles plays a substantive role in confidence measures.

Descriptive Findings

A substantial correlation between the NF and IF groups' *item difficulties* (*P*) were observed, $r = .91$ and also for the NF and IF *item confidence self-report* means, $r = .85$, indicating that difficult items were difficult for both groups, and that low confidence responses were also low for both groups, and vice versa (see Figure 3). This may be a generally under-reported but obvious finding that group average item-level confidence measures, similar to group average item-level difficulty (*P*) measures, are similar across treatment groups. The correlation values observed between the group-level confidence measures of the two groups seems like an important probative finding, and so should be considered in future SRL investigations.

Internal Cues

Confidence judgments inversely influenced follow-on confidence judgements when performance feedback was not provided (see Table 2). Specifically, at the item level (micro level), confidence self-reports exhibited hysteresis as autocorrelation for no feedback, where a confidence response influenced the next confidence response inversely at Lag 1. Since most

learning happens without immediate feedback, future SRL investigations should consider the influence of confidence responses on follow-on confidence responses (hysteresis).

The hard-easy effect that was observed for item difficulty and confidence is consistent with past investigations that individuals are often overconfident on difficult items and under confident with easy items. This can be parsimoniously explained by stability of individual response styles as a form of regression to the response confidence mean. This can also be thought of as an individual confidence setpoint. It would be prudent for future SRL investigations to report content difficulty estimates and also monotonicity between the monitoring and performance measures, at least as a scatter plot, as qualifying assumptions before reporting relative or absolute monitoring accuracy.

External Cues

Lesson confidence was predicted by the *Word Concreteness* dimension for both immediate and no feedback treatments. For both regression equations, the *Word Concreteness* Coh-Metrix dimension had a counter-intuitive *inverse* relationship with confidence. Word concreteness refers to a perceptible entity (i.e., a dog). Concrete words are easier to remember than abstract words (i.e., love) and are likely to be more familiar and thus "easier". So it is unclear why increased item concreteness decreases confidence self-report values for the item. Perhaps this is a 'feeling of knowing' or 'tip of the tongue' cue that can confound accurate monitoring. Further research should consider the influence of Coh-Metrix text features on monitoring with both new (unfamiliar) and review lesson (familiar) content.

Summary

The significant item-level confidence response autocorrelation for no feedback (hysteresis), the extraordinarily large reliability value for the item-level confidence measures (Cronbach's alpha of 0.96), and that item-level confidence responses tended to be less extreme than the item *P* values across the lesson (see Figure 3) taken together support the likeliness of stability of individual response styles for confidence judgements. This may be a form of stability bias (Kornell & Bjork, 2009); note that Bjork et al. (2013) have proposed the troubling implications of stability bias on monitoring in self-regulated learning (p. 429). Future investigations should consider the influence of stability of individual response styles on monitoring measures (Bjork et al. 2013; Javaras & Ripley, 2007). Perhaps stability bias can be statistically controlled in future SRL research.

Because micro-level monitoring depends on incomplete and even competing information, there is likely to be a monitoring trade-off between the bioenergetic commitment needed for constant diligence and monitoring accuracy (Rae et al, 2003), with the mental system favoring macro level monitoring and may surrender monitoring control to an external system when immediate feedback is available (both require less effort). Carver and Scheier (2000) note that for self-regulation in general, "It is pointless and maybe even counterproductive to plan too far ahead too fully … Thus, it makes sense to plan in general terms, chart a few steps, get there, reassess, and plan the next bits." (p. 67).

On the other hand, micro-level hypervigilance may pay off in some extreme settings. Thus there would be an advantage for human monitoring accuracy to be "just good enough", seeking a balance between micro-level accuracy, effort costs, and macro-level accuracy. The importance of explicit micro-level monitoring would likely be amplified in high stakes decisions.

Limitations of this study

Because content familiarity is likely strongly related to feeling of knowing, and both are suspected cues for explicit and implicit monitoring (Bjork et al., 2013; Kornell, & Bjork, 2009), then in terms of generalizability of these findings, these results here should only apply under conditions of familiar content (i.e., students in an end of course review of key vocabulary). In settings where content is unfamiliar or less familiar to the learners, pre-existing low familiarity is likely to be an important monitoring cue. Since *Word Concreteness* of words in recently studied material likely manifests differently than with unstudied materials, especially for technical vocabulary, because familiarity may falsely increase confidence as feeling of knowing, then these findings for *Word Concreteness* do not automatically apply when learning new or unfamiliar material.

The results observed here must be limited to recognition lesson tasks. In contrast, Pressley, Ghatala, Woloshyn, and Pirie (1990) report that short-answer questions can increase metacognitive accuracy relative to multiple-choice questions. Thus these findings should definitely not be extended beyond recognition tasks.

A ramification of this study is that it is possible that immediate feedback may deter some students from monitoring. Students may overly rely on the lesson feedback as explicit external monitoring rather than intuitive implicit monitoring (Azevedo & Hadwin, 2005; Corbett & Anderson, 2001). Future research should consider whether immediate feedback subverts self-regulation and whether this manifests as self versus system control of instructional strategy selection for some students.

Implications for Design and Theory

What seemed to be a simple data set of item performance and confidence self-report measures turns out to be interesting, complex, and difficult to interpret. We fully agree with Butler and Winne (1995), who note that:

> SRL is a process that unfolds step-by-step over time. It is also recursive; that is, internal monitoring of a current state in a task, the trigger for engaging SRL, generates feedback that, in turn, is input contributing to the learner's regulation of subsequent cognitive engagement. (p. 246)

New ways for measuring monitoring and strategy shifts on the fly, such as automated emotion recognition through facial expressions and bodily movements (Azevedo, 2014; López & Tucker, 2018), offer promise for measuring implicit, unconscious regulation that contrast with self-report measures (e.g., Kahneman, 2013; System 1 as fast, instinctive, and emotional and System 2 as slow, deliberative, and logical). Such new measurement approaches could provide abundant real-time data to better understand SRL processes. New data sources, both explicit and implicit, and more robust analysis approaches that capture activity across multiple scales as well as levels of these scales across time may soon drive SRL theory and research. So then there is a need now to refine and integrate multiple measures and methodologies to inform new theory building in SRL (Cascallar, Bockaerts, & Costigan, 2006).

References

Arnold, M. M., Graham, K., & Hollingworth-Hughes, S. (2017). What's context got to do with it? Comparative difficulty of test questions influences metacognition and corrected scores for formula-scored exams. *Applied Cognitive Psychology, 31*, 146-155.

Azevedo, R. (2014). Issues in dealing with sequential and temporal characteristics of self- and socially-regulated learning. *Metacognition and Learning, 9,* 217-228. https://doi.org/10.1007/s11409-014-9123-1

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition - Implications for the design of computer-based scaffolds. *Instructional Science, 33*(5-6), 367-379.

Bandura, A. (1991). Social Cognitive Theory of Self-Regulation. *Organizational Behavior and Human Decision Processes, 50*, 248-287.

Benwell, C. S. Y., Beyer, R., Wallington, F., & Ince, R. A. A. (2020). History biases reveal novel dissociations between perceptual and metacognitive decision-making. *bioRxiv*, online. https://doi.org/10.1101/737999

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444.

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in Psychology, 3*(229), 1-6.

Bol, L., Riggs, R., Hacker, D. J., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education, 21*, 81-96.

Brady, M., Rosenthal, J. L., Forest, C. P., & Hocevar, D. (2020). Anonymous versus public student feedback systems: metacognition and achievement with graduate learners. *Educational Technology Research and Development, 68*, 2853-2872. https://doi.org/10.1007/s11423-020-09800-6

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*(1), 60-77.

Butler, D. L. & Winne, P. H. (1995) Feedback and self-regulated learning: a theoretical synthesis. *Review of Educational Research, 65*(3), 245-281.

Carver, C. S., & Scheier, M. F. (2000). On the structure of behavioral self-regulation. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 41-84). San Diego, CA, US: Academic Press. http://dx.doi.org/10.1016/B978-012109890-2/50032-9

Cascallar, E., Bockaerts, M., & Costigan, T. (2006). Assessment in the evaluation of self-regulation as a process. *Educational Psychology Review, 18*, 297-306. DOI 10.1007/s10648-006-9023-2

Clariana, R.B. (1990). A comparison of answer until correct feedback and knowledge of correct response feedback under two conditions of contextualization. *Journal of Computer-Based Instruction, 17*, 125-129.

Clariana, R.B., Ross, S. L., & Morrison, G. R. (1991). The effects of different feedback strategies using computer-assisted multiple-choice questions as instruction. *Educational Technology Research and Development, 39*(2), 5-17.

Corbett, A. T., & Anderson, J. R. (2001, March). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 245-252). ACM.

Dinsmore, D. I., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4-14.

Dowell, N. M. M., Graesser, A. C., & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics, 3*(3), 72-95. http://dx.doi.org/10.18608/jla.2016.33.5

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16* (4), 228-232.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271-280. doi:10.1016/j.learninstruc.2011.08.003.

Dunlosky, J., & Rawson, K. A. (2015). Do students use testing and feedback while learning? A focus on key concept definitions and learning to criterion. *Learning and Instruction, 39*, 32-44.

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction, 24*, 58-61.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 238-244.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*, 19-34.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 1-9.

Follmer, D. J., & Clariana, R. B. (2020). Predictors of adults' metacognitive monitoring ability: The roles of task and item characteristics. *Journal of Experimental Education*, Online First. https://doi.org/10.1080/00220973.2020.1783193

Grabowski, B. L. (2004). Generative learning contributions to the design of instruction and learning. In D. H. Jonassen (Ed.), *Handbook of Research on Educational Communications and Technology, 2nd edition (pp. 719-743). Mahwah, NJ: Erlbaum.*

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's Model of Self-Regulated Learning: new perspectives and directions. *Review of Educational Research, 77*(3), 334-372. DOI: 10.3102/003465430303953

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*(2), 101-121.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160-170.

Händel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgement training over and above repeated testing. *Learning and Instruction, 65*, 101245. https://doi.org/10.1016/j.learninstruc.2019.101245

Hartwig, M. K., & Dunlosky, J. (2017). Category learning judgments in the classroom: Can students judge how well they know course topics? *Contemporary Educational Psychology, 49*, 80-90.

Hertzog, C., Hines, J. C., & Touron, D. R. (2013). Judgments of learning are influenced by multiple cues in addition to memory for past test accuracy. *Archive of Scientific Psychology, 1*(1), 23-32.

Illingworth, J. (2011). *Control Systems 2003* (html lecture notes). Retrieved from the University of Leeds website: http://www.bmb.leeds.ac.uk/illingworth/control/

Javaras, K. N., & Ripley, B. D. (2007). An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102*, 454-463.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*, 63-85.

Juslin, P., Winman, A., & Olsson, H. (2000). Naïve empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107*, 384-396.

Kahneman, D. (2013). *Thinking, fast and slow*. New York, NY: Farrar, Straus, and Girox.

Kavousi, S., Miller, P. A., & Alexander, P. A. (2020). The role of metacognition in the first-year design lab. *Educational Technology Research and Development, 68*, 3471-3494. https://doi.org/10.1007/s11423-020-09848-4

Kelemen, W. L., Frost, P. J., & Weaver, C. A. III (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory and Cognition, 28*, 92-107.

Keren, G. (1991). Calibration and probability judgments: conceptual and methodological issues. *Acta Psycholica, 77*, 217-273. https://www.sciencedirect.com/science/article/pii/000169189190036Y

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*, 449-468.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121-1134.

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68*(5), 522-528.

Laerd Statistics (2018). *Spearman's Rank-Order Correlation using SPSS Statistics*. Retrieved from https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159-183.

Lin, L.-M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*, 345-391. doi:10.1006/ceps.1998.0972

López, C. E., & Tucker, C. S. (2018). *Towards personalized performance feedback: mining the dynamics of facial keypoint data in engineering lab environments*. A paper presented at the 2018 ASEE Mid-Atlantic Spring Conference, April 6-7, 2018 - University of the District of Columbia. Retrieved from https://peer.asee.org/towards-personalized-performance-feedback-mining-the-dynamics-of-facial-keypoint-data-in-engineering-lab-environments.pdf

Merriman, K.A., Clariana, R.B., & Bernardi, R.J. (2012). Goal orientation and feedback congruence: effects on discretionary effort and achievement. *Journal of Applied Social Psychology, 42* (11), 2776-2796.

Mills, C., D'Mello, S. K., & Kopp, K. (2015). The influence of consequence value and text difficulty on affect, attention, and learning while reading instructional texts. *Learning and Instruction, 40*, 9-20.

Nelson, T. O., & Dunlosky. J. (1991). When people's judgements of learning (JOL)s) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*(4), 267-270.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*, 7-28.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning*, *1*(2), 159-179.

Nugteren, M. L., Jarodzka, H., Kester. L., & van Merriënboer, J. J. G. (2018). Self-regulation of secondary school students: self assessments are inaccurate and insufficiently used for learning-task selection. *Instructional Science, 46*, 357-381.

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

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 43-97). Lincoln: Buros Institute of Mental Measurements.

Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the main idea and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly, 25*, 232-249.

Rae, C., Scott, R. B., Lee, M., Simpson, J. M., Hines, N., Paul, C., … Radd, G. K. (2003). Brain bioenergetics and cognitive ability. *Developmental Neuroscience, 25*, 324-331. DOI: 10.1159/000073509

Reid, A. J., Morrison, G. R., & Bol, L. (2017). Knowing what you know: improving metacomprehension and calibration accuracy in digital text. *Educational Technology Research and Development, 65*, 29-45.

Rutherford, T. (2014). *Calibration of confidence judgments in elementary mathematics: Measurement, development, and improvement.* (Doctoral dissertation). Retrieved from UC Irvine Electronic Theses and Dissertations: http://escholarship.org/uc/item/99z17038

Stock, W. A., Kulhavy, R. V., & Pridemore, D. R. (1992). Responding to feedback after multiple-choice answers: the influence of response confidence. *Quarterly Journal of Experimental Psychology, 45*, 649-667.

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review*, *12*(4), 437-475.

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*(1), 49- 61.

Vasilyeva E., Pechenizkiy, M., & De Bra, P. (2008). Tailoring of feedback in web-based learning: The role of response certitude in the assessment. In Woolf B.P., Aïmeur E., Nkambou R., Lajoie S. (Eds.), *Intelligent Tutoring Systems. ITS 2008. Lecture Notes in Computer Science*, vol. 5091. Berlin, Heidelberg, GR: Springer. https://doi.org/10.1007/978-3-540-69132-7_104

Walkington, C., Clinton, V., & Shivraj, P. (2018). How readability factors are differentially associated with performance for students of different backgrounds when solving mathematics word problems. *American Educational Research Journal, 55*(2), 362-414.

Weaver, C. A., III, & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory and Cognition, 23*, 12-22.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods, 15*, 96-110.

Zhu, M., Bonk, C. J., & Doo, M. Y. (2020). Self-directed learning in MOOCs: exploring the relationships among motivation, self-monitoring, and self-management. *Educational Technology Research and Development, 68*, 2073-2093. https://doi.org/10.1007/s11423-020-09747-8

Figures

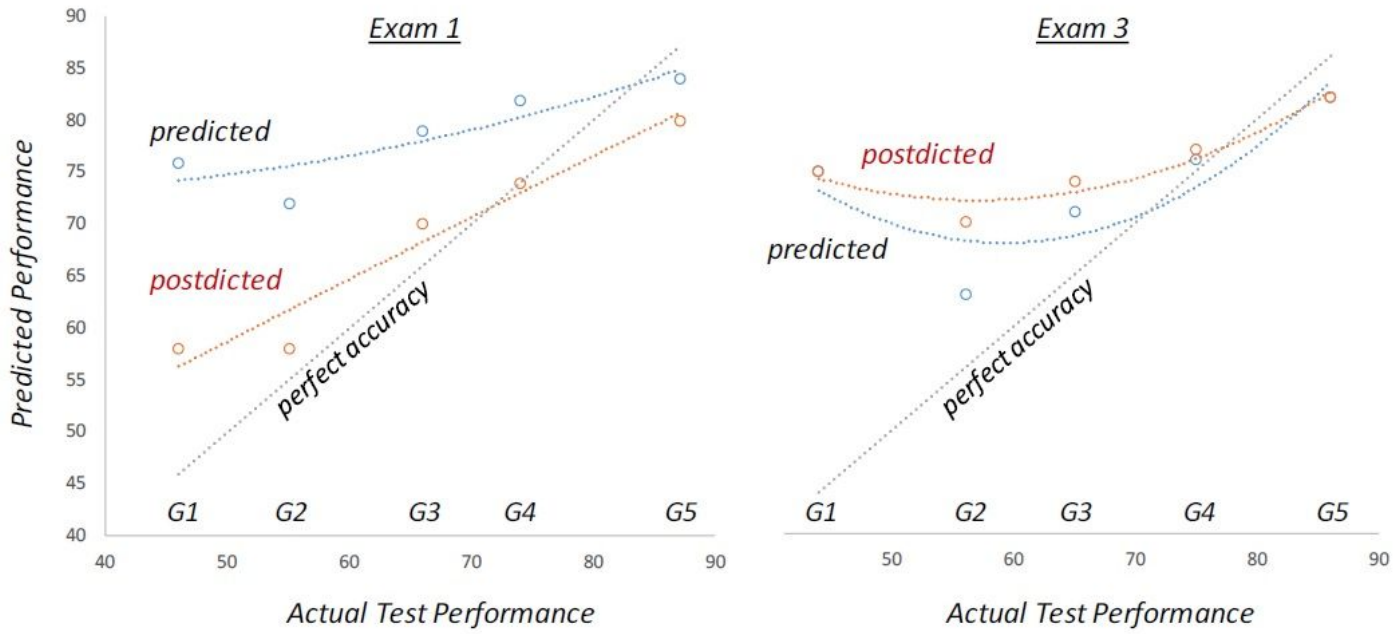

Figure 1

Group-level examination performance of five ability groupings (G1-G5) with predicted and postdicted performance judgements at two time periods (from Hacker at al., 2000).

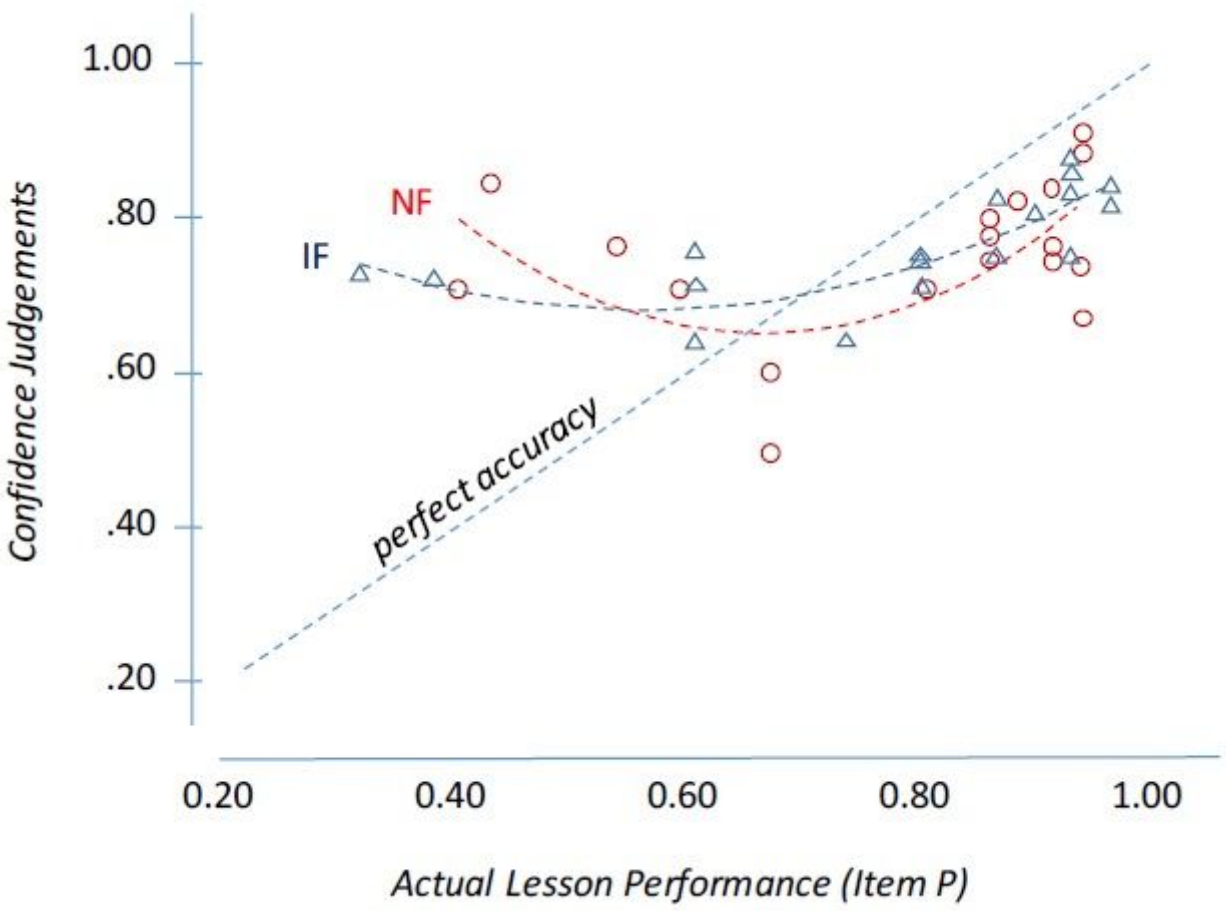

Figure 2

Scatterplots of the actual lesson performance average lesson item difficulties (item P) and item confidence for the NF and IF treatments.

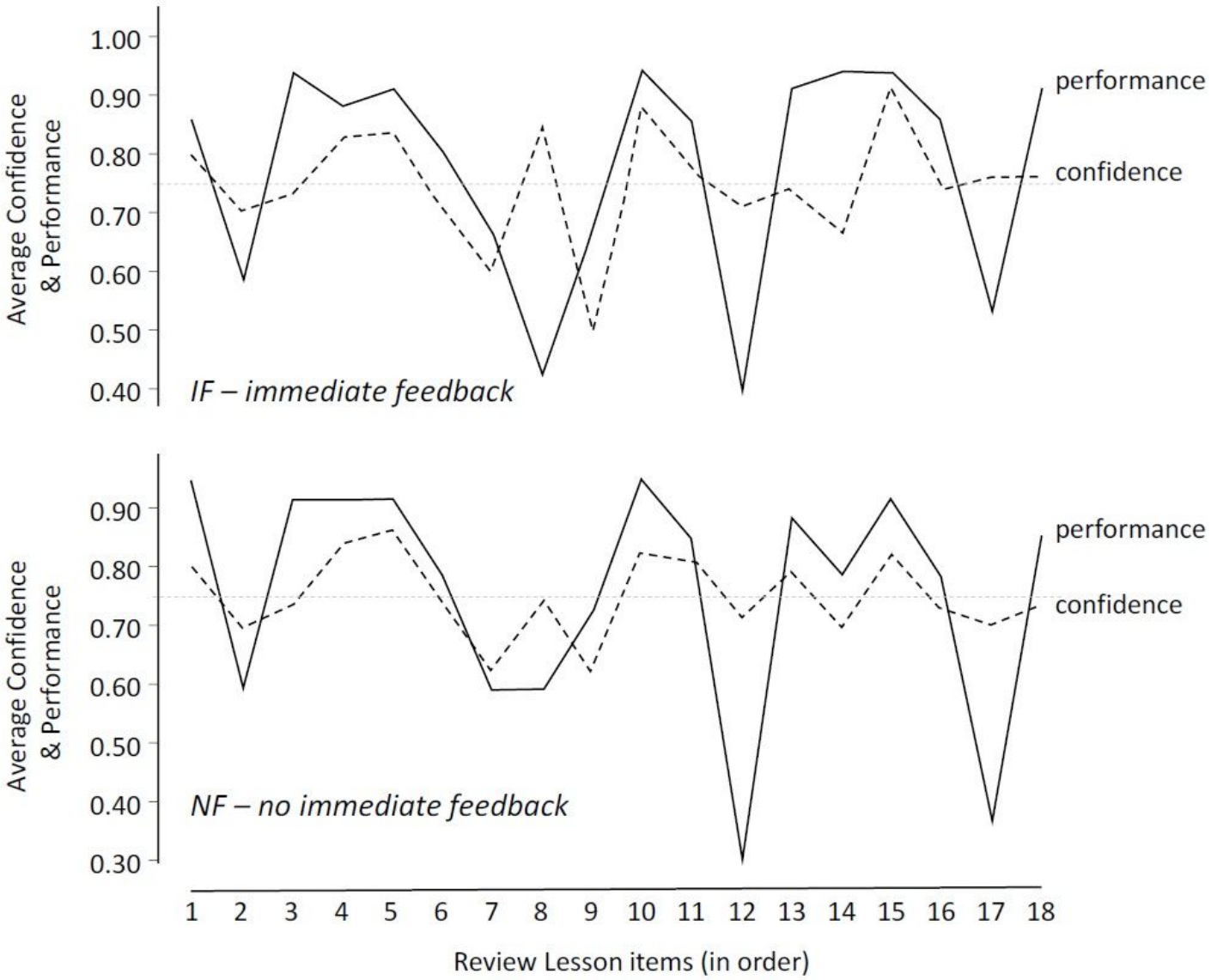

Figure 3

IF (top) and NF (bottom) treatments' review lesson average item-level difficulty (solid lines), item-level confidence (dashed lines).