TESL-EJ Teaching English as a Second or Foreign Language
The Electronic Journal for English as a Second Language

February 2018 – Volume 21, Number 4

Inaccurate Metacognitive Monitoring and its Effects on Metacognitive Control and Task Outcomes in Self-Regulated L2 Learning

Jim Ranalli
Iowa State University
<jranalli@iastate.edu>

Abstract

Accurate metacognitive monitoring of one's own knowledge or performance is a precondition for self-regulated learning; monitoring informs metacognitive control, which in turn affects task outcomes. Studies of monitoring accuracy and its connection to knowledge and performance are common in psychology and educational research but rare in instructed SLA. This paper describes two studies in which actual performance and self-evaluated performance were compared. In Study 1, 64 college-level ESL learners completed L2 vocabulary tasks that differed in complexity and familiarity. Wide discrepancies in monitoring accuracy were observed. In Study 2, the same sample was divided into two groups, and the more complex task from Study 1 was used as a pre- and post-test. One group was given strategy instruction to improve monitoring accuracy. Metacognitive control was operationalized as choice of dictionaries used during the task, and effects of control on task outcomes were operationalized as points achieved on the task that could be attributed to use of a particular dictionary type. Results showed that control decisions leading to poorer task outcomes were associated with lower monitoring accuracy.

Keywords

Metacognitive monitoring, Metacognitive control, Self-regulated learning, Learner strategies, Vocabulary learning

Introduction

In many learning-related tasks, learners need an accurate sense of their current state of knowledge or performance to regulate their behavior; that is, to decide whether to stay the course if the current approach seems to be closing the distance to the goal; to try something else if the current approach does not appear to be working; or to abandon the task altogether

if the costs of further engagement are deemed to outweigh the benefits. One's sense of one's current state of knowledge or performance originates in a process called *metacognitive monitoring*, while the enactment of decisions about maintaining, altering, or abandoning one's approach to learning represents another process called *metacognitive control*. Monitoring informs control. When monitoring is not accurate, control is based on suboptimal information, making desired task outcomes harder to achieve.

Psychologists have frequently investigated the accuracy of metacognitive monitoring, with more recent work linking monitoring accuracy to metacognitive control and learning outcomes, but such concerns have been largely absent from second language acquisition (SLA) research, perhaps because of researchers' primary concerns with unmonitored forms of language processing. To demonstrate the relevance of these issues to instructed SLA, I first review an influential model of monitoring and control and then compare treatments of these constructs across disciplines. Next, I describe two studies: the first showing how monitoring accuracy can vary considerably according to task complexity and familiarity, and the second demonstrating connections between monitoring accuracy, control, and task outcomes.

Monitoring and control as the basis of metacognition and self-regulated learning

The term metacognition to describe "thinking about thinking" was coined by Flavell, whose research showed that children begin to differentiate between mental representations of reality and reality itself at about age four. Flavell's descriptions of the conscious and purposeful nature of metacognition (e.g., Flavell, 1979) have been cited widely outside psychology, but it was a simple process model proposed by Nelson and Narens (1990) that first showed how monitoring and control interacted and further how adults could consistently but inaccurately evaluate their own memory and learning.

Nelson and Narens's "metamemory" model (1990) consists of an object level and a meta-level. The object level comprises cognitions related to objects and information in the outside world, including a person's own actions and behavior. The meta-level comprises cognitions about the cognitions at the object level. During monitoring, the object-level informs the present state of the meta-level's object level representation, whereas in control, the meta-level enacts changes in the current state of the object level. The model thus shows how potential discrepancies between the actual state of the object level and its representation at the meta-level can originate during monitoring and, in turn, affect conditions at the object level during control.

Monitoring and control form the basis of many theories of self-regulated learning (SRL), which is "a cyclical process in which students monitor the effectiveness of their learning methods or strategies and respond to this feedback in a variety of ways, ranging from covert changes in self-perception to overt changes in behavior" (Zimmerman, 2001, p. 5). In one influential SRL model proposed by Winne and Hadwin (1998), monitoring and control operate in all four stages of any learning event: task definition, planning/goal setting, task engagement, and post-task adaptations to goals, strategies, etc. While laboratory-based research into monitoring accuracy is abundant, naturalistic studies of monitoring accuracy's

effects on self-regulation and learning outcomes are rare (de Bruin & van Gog, 2012). However, recent special issues of the journal *Learning and Instruction* (Volumes 22/4 and Volume 24) attest to growing recognition of the need to investigate these connections.

Monitoring and control in L2 research

Monitoring and control are relevant to many areas of L2 learning and use insofar as they facilitate self-regulation of thought, strategic behavior, motivation, and affect. Through the lens of learner autonomy, control can be seen extended across broad domains of L2 learning, from cognitive processing to self-management of studying to selecting the content of learning (Benson, 2007). Technology has dramatically increased learners' opportunities to make decisions about where, when, and with what forms of the target language to engage (Kormos & Csizér, 2014; Lai, Shum, & Tian, 2014), while the trend toward online and distance education requires that more responsibility and thus control be devolved to learners (Hauck, 2005; Lai, 2013). In addition, L2 learning requires learners to regulate motivation and negative affect (Dörnyei, 2001; Dörnyei & Ryan, 2015).

Monitoring has received a great deal of attention in the L2 learner strategies literature, where it is usually characterized as a metacognitive strategy (along with planning and evaluation). Control can be seen collectively represented in the taxonomies of cognitive (i.e., lower-order) strategies proposed by researchers such as Oxford (1989, 2017) and Chamot (1987). Descriptive studies of strategy use have shown that monitoring is a hallmark of skilled listening (Vandergrift, 1997) and skilled writing (Cumming, 1989) while intervention studies have documented increased performance after training in metacognitive strategies, including monitoring (e.g., De Silva, 2014). However, one finds little if any discussion of monitoring accuracy or its effects on L2 strategy use.

One of the few L2 investigations of monitoring accuracy comes from the area of language testing. Phakiti (2005, 2016) framed the study in terms of *calibration*; that is, the extent to which learners' confidence in (or self-assessment[1] of) their performance aligns with an external measure. Phakiti had 295 EFL learners at a Thai university evaluate their performance on an English placement test using two confidence measures: (1) local confidence, which involved a self-assessment of every response on the test according to percentage scales (e.g., *0% confident, 25% confident,* etc.); and (2) global confidence, in which test-takers estimated their total number of correct answers out of 100 possible. (Confidence judgments differ from other measures of monitoring accuracy, such as feeling of knowing judgments, or FOKs, by being "post-dicted" after an item, task, or test, rather than predicted beforehand.)

A correlation of .54 between performance and local confidence was reported for the test as a whole. Phakiti (2005) also used a simple linear model to calculate calibration scores: $C = c - p$, where C is calibration, c is confidence and p is actual performance. A score of 0 indicates perfect calibration, while scores above 0 show overconfidence, and scores below 0 show underconfidence. Local confidence showed overestimation of +6.4% across four sections of the test. Calibration was best in reading, +2.64%, and worst in grammar, +9.00%. Global confidence for the sample was -.08%, although this does not mean individuals were well-

calibrated. A scatterplot of the data (Figure 1) shows roughly equal numbers of test-takers clustering above and below the identity line (representing perfect calibration), thus largely cancelling each other out. For reasons such as this, calibration studies typically involve visual analysis (Pieschl, 2009).

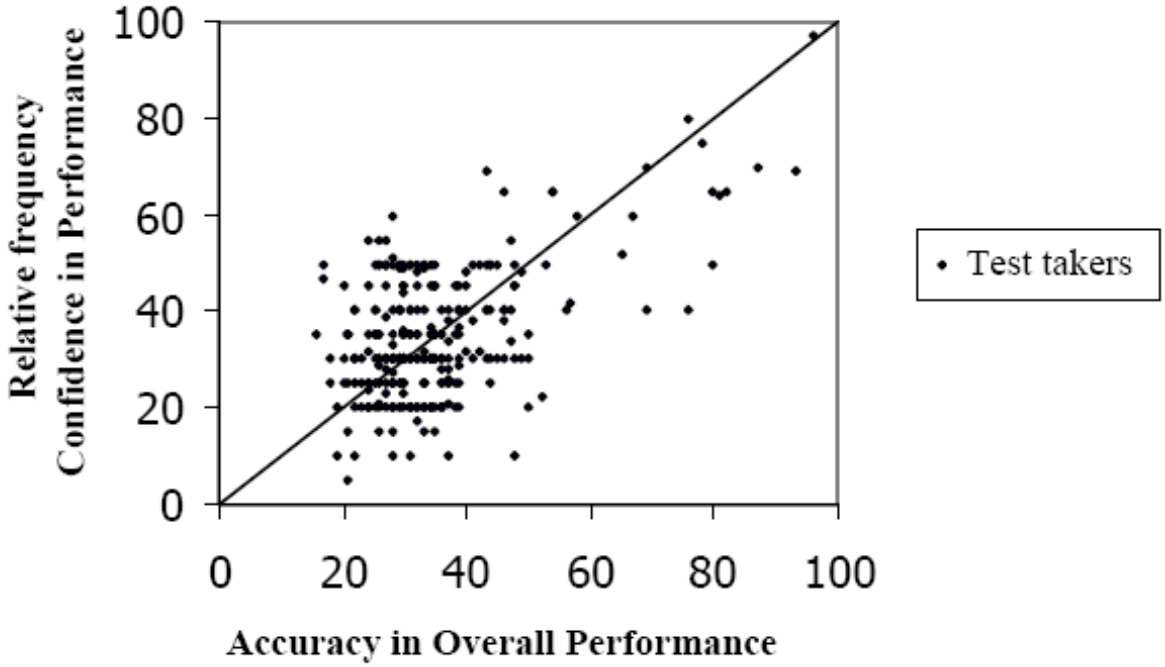

Figure 1: Global-measure calibration on the English Placement Test (N = 295) from Phakiti (2005, p. 43). From "An empirical investigation into the nature of and factors affecting test takers' calibration within the context of an English Placement Test (EPT)," by A. Phakiti, 2005, Spaan Fellow Working Papers in Second or Foreign Language Assessment, Volume 3, Copyright by English Language Institute, University of Michigan. Reprinted with permission.

In the more recent of two articles based on the same data set, Phakiti (2016) related participants' calibration to their use of metacognitive and cognitive test-taking strategies, which they reported via Likert-scale items on a questionnaire. An analysis based in structural equation modeling found test-takers' confidence to be only weakly related to metacognitive strategy use, which led Phakiti to conclude that "when test takers are inaccurate in their performance appraisals, they cannot use cognitive and metacognitive strategies to deal with test tasks successfully" (2016, p. 98). This remains speculation, however, insofar as metacognitive control was not directly measured in the study. To find more direct evidence connecting monitoring accuracy to control, one must look to psychology and other areas of education.

Another study by Trofimovich et al. (2016), focused on self-assessment of performance in the area of L2 pronunciation. L2 speakers ($N = 194$) from a number of L1 backgrounds were recorded performing an extemporaneous speaking task, which was then rated for accent and comprehensibility by each speaker him or herself as well as a panel of judges. The

relationships between actual and self-rated scores were found to be weak, with correlations of .06 for accentedness and .18 for comprehensibility. The study did not investigate relationships between monitoring accuracy and control.

Monitoring and control in psychology and other areas of education

In psychological research, many factors have been found to moderate the accuracy of self-assessments, including the amount of relevant prior knowledge possessed by participants (Nietfeld & Schraw, 2002); whether the knowledge assessed was general or domain specific (Glenberg & Epstein, 1987); the difficulty of items (Gigerenzer, Hoffrage, & Kleinbölting, 1991); the timing and type of self-assessment measure employed (Walczyk & Hall, 1989); and the stage of the learning process in which monitoring occurs (Pieger, Mengelkamp, & Bannert, 2016). The overall view that emerges from this research is that people's self-assessments "hold only a tenuous to modest relationship with their actual behavior and performance" (Dunning, Heath, & Suls, 2004, p. 69).

One strand of this research has focused on patterns of miscalibration along the scale of performance. Dunning and his associates (Kruger & Dunning, 1999; Ehrlinger & Dunning, 2003; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008) have shown that poorer performers tend to overestimate, often significantly, while top performers underestimate slightly. Whereas the former is attributed to lack of self-insight, the latter has been explained in terms of a "false consensus effect" (Ross, Greene, & House, 1977) whereby subjects overestimate the abilities of peers. Such slight underestimation can be seen among the top performers in Figure 1.

Surprisingly, few psychology studies have investigated the roles of task complexity and familiarity on monitoring accuracy, although such work can be found in the field of medical education, where self-assessment is of concern because inaccuracies can lead to suboptimal patient care. Fitzgerald, White, and Gruppen (2003) investigated the self-assessment accuracy of medical students over three years based on conventional, mostly multiple choice end-of-term exams as well as a culminating skills-based assessment addressing a variety of clinical tasks. The students showed consistent, slight underestimation on the conventional assessments but overestimation on the clinical exam, which the authors attributed to variations in task type and task familiarity.

Studies relating monitoring accuracy to both control and learning outcomes are a relatively recent development in psychology. This work assumes a "test preparation" scenario in which students monitor their grasp of previously learned material in anticipation of an external assessment and, in doing so, exercise control either by selecting which materials to (re)study (Thiede, Anderson, & Therriault, 2003; Dunlosky & Ariel, 2011), selecting which to devote more time to in (re)studying (Ariel, Dunlosky, & Bailey, 2009; Hines, Touron, & Hertzog, 2009; van Loon et al., 2013), or both (Mihalca, Mengelkamp, & Schnotz, 2017). These studies have demonstrated that learners indeed make poor decisions about the allocation of attention or study time based on inaccurate monitoring—in other words, monitoring influences self-regulation of learning (Ariel, Dunlosky, & Bailey, 2009; Mihalca, Mengelkamp, & Schnotz, 2017)—and, further, that flawed self-regulation negatively affects

learning outcomes (Hines, Touron, & Hertzog, 2009; Thiede, Anderson, & Therriault, 2003; van Loon et al., 2013).

However, the test-preparation paradigm has significant limitations regarding its ability to inform L2 research. The criterion tasks in these studies are usually of a single type, such as paired-associates learning or reading comprehension, and they vary greatly in complexity across studies, which not only makes it difficult to compare findings but has also contributed to gaps in the monitoring-accuracy literature regarding the influence of task complexity and familiarity (Pieschl, 2009). More importantly, the narrow conceptualization of learning reflected in psychological research (i.e., assimilating factual or conceptual information into long-term memory) contrasts with the broad range of activities thought to promote learning in instructed SLA, which encompasses conventional forms of studying (e.g., memorizing vocabulary lists) as well as a huge variety of language-use tasks. L2 learning will often be defined in larger units, such as a series of related activities, with learners making evaluations during task engagement or after the fact. Many L2 learning situations will also lack external forms of assessment to provide objective feedback, thus rendering post-dicted judgments more important. Research is therefore needed that incorporates tasks and measures more germane to L2 learning.

To sum up, monitoring and control are central to metacognition and self-regulated learning. Research has shown that monitoring tends to be inaccurate, particularly when some form of knowledge is lacking and, further, that inaccurate monitoring is connected to control decisions misaligned to study requirements or task demands. The influence of task characteristics on monitoring accuracy is also poorly understood. To address these gaps and to demonstrate the relevance of monitoring accuracy and control to instructed SLA, I conducted two studies focused on the following research questions.

Study 1: Does monitoring accuracy vary as a function of task complexity and task familiarity?

Study 2: Do variations in monitoring accuracy affect metacognitive control and resultant task outcomes?

The studies were based in the domain of L2 vocabulary learning because of the self-regulatory burden it puts on learners. My basis for this claim is that L2 curricula often lack a course devoted to vocabulary, addressing it instead through other subjects such as reading (Folse, 2010). The tacit assumption is that learners will acquire the lexis they need independently through incidental exposure in reading and listening as well as their own efforts at intentional learning.

Study 1

Methods

Participants. The participants were 64 international students (26 women and 38 men) enrolled in three sections of an English as a second language (ESL) composition course at a

large research university in the midwestern U.S. Most (76%) spoke Chinese as their first language (L1), with Arabic, Korean, and Malay making up most of the other L1s of the group.

Materials

Simple, familiar task. The simple, familiar task was the Vocabulary Levels Test, or VLT, (Nation, 1990; Schmitt, Schmitt, & Clapham, 2001), designed to test knowledge of basic form-meaning associations of English words in different frequency bands. The basic unit of the test is a cluster of three stems and six options (Figure 2). The test was adapted for online administration such that participants would key in the number of the correct word form in the appropriate field.

Figure 2: Portion of the Vocabulary Levels Test (Nation 1990; Schmitt, Schmitt, and Clapham 2001) adapted for online administration in the study

Four of the original five frequency-band levels of the VLT were employed: the 2000, 3000, 5000 levels and a level based on the Academic Word List (Coxhead, 2000). Each band was tested on a single web page that included 30 items, with one point automatically assigned for each correct answer for a maximum possible score of 120. One confidence judgment was collected for each of the four levels via web-forms displayed after each test page; that is, participants estimated their score out of 30 for each level, thus producing four confidence judgments.[2]

The VLT was deemed a simple task not unlike the paired-associate tests used in many calibration studies. Test-takers chose from a limited set of options with no manipulation or transformation of information required. They needed only to consult existing knowledge, and, in cases of unknown word forms, they could use the process of elimination to make educated guesses, basing the probability of correctness on the number of possible answers. Furthermore, the task was likely to be familiar to many ESL students from their previous schooling.

Complex, unfamiliar task. The complex, unfamiliar task, created as part of a larger study (Ranalli, 2012a), was called the Pattern Identification and Correction Task, or PICT. It was designed to assess an ESL student's' ability to consult pedagogical (or learner) dictionaries of English to find usage information for correcting errors of lexico-grammatical patterning.

The PICT was also completed online and is best described via example. Participants are presented with a web page containing a sentence such as *In the 1950s, American scientists warned people for the harmful effects of television* (Figure 3). In both written and video instructions, they have been told the sentence contains a content word (e.g., verb, noun, or adjective) used incorrectly. The problem – in this case, the pattern of the verb *warn* – must be identified and corrected using an online dictionary. Participants are told to copy and paste the sentence into the editing box to make their correction, and are told not to change the key word in any way but to focus on the words around it, adding, deleting or modifying them as needed but without altering the meaning of the sentence. A table of links to various online dictionaries is provided, but the participants are told they can use others if they wish.

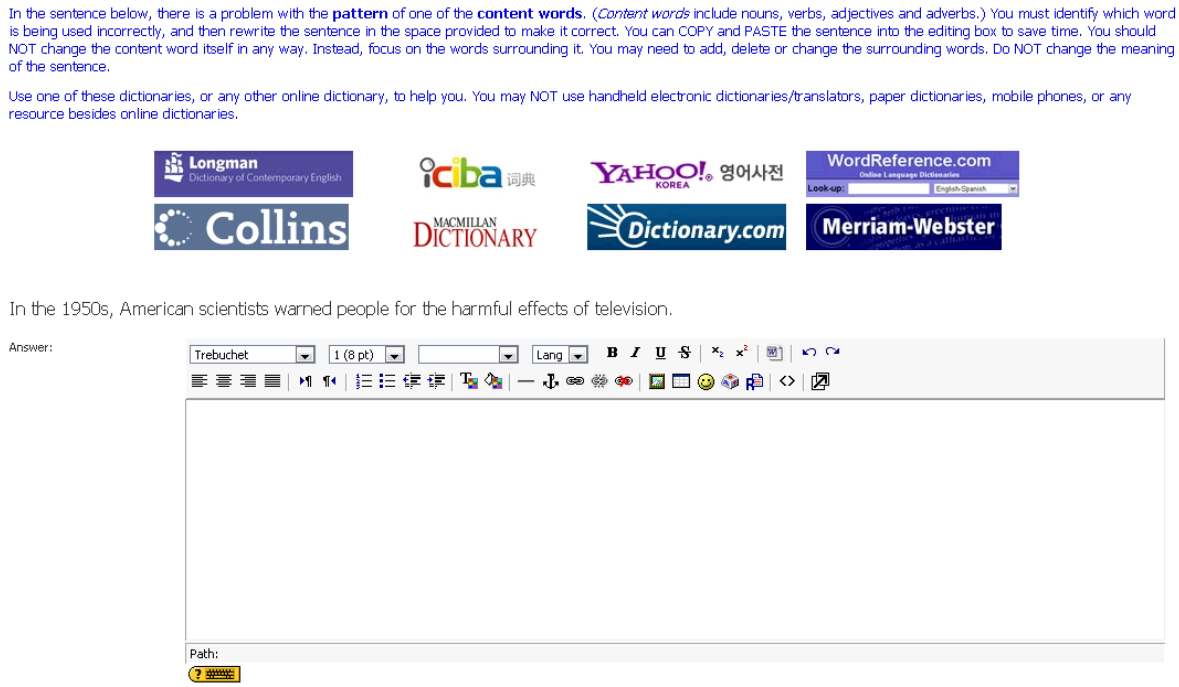

Figure 3: Item from the PICT showing instructions, links to online dictionaries, sentence containing a pattern-grammar error, and editing box

While participants could use existing knowledge, the sentences were selected to contain unfamiliar lexico-grammatical patterns; thus, the task typically requires use of a learner dictionary since bilingual or "native speaker" dictionaries do not reliably provide such information.[3] After finding the appropriate entry, participants must identify the sense of the word that corresponds to the usage in the sentence, select an appropriate pattern – in this case, *warn somebody about* or *of something* – and then return to the task page to make the required changes.

The PICT was deemed complex insofar as it requires:

- reading/decoding;
- possessing functional and conceptual knowledge of lexico-grammatical patterning;
- syntactic parsing of sentences to identify possible lexico-grammatical patterns;
- identifying the word class of sentence constituents;

- distinguishing among dictionary types and selecting the appropriate type for the purpose;
- finding relevant usage information in dictionary entries; and
- applying dictionary usage information to a new context, including modifications to form as needed.

It was also thought likely to be unfamiliar because lexico-grammatical patterns are taught in an ad hoc way if they are taught at all (Hunston, Francis, & Manning, 1997).

In scoring, one point was awarded if any change was made to the grammatical component of the pattern (i.e., to any part other than the lexical node word), which was interpreted as evidence that the error had been located. An additional point was awarded if the change matched an item in a list of acceptable responses compiled using information from pedagogical dictionaries and the Corpus of Contemporary American English (Davies, 2008). Thus, the maximum possible score was 20.

Confidence judgments on the PICT were collected via an online survey form, which briefly explained the scoring system and elicited a single post-dicted score out of a possible 20.

Procedure. Participants were recruited during regular weekly meetings of the composition classes in computer labs. After informed consent procedures, they were shown video instructions for the PICT and then had 20 minutes to complete it. Immediately afterward, they were directed to the post-diction form. One week later, again during a computer-lab session, they took the VLT, which they had 40 minutes to complete, and made their post-dictions after each frequency-band level.

Results and discussion

Reliability of the performance measures, determined using Cronbach's alpha, was excellent for the VLT as a whole (.96) and good or excellent across levels (2000 = .86; 3000 = .88; 5000 = .87, AWL = .92); reliability for the PICT was good (.74). Reliability of confidence judgements on the VLT, determined by a Spearman-Brown coefficient, was high (.94). Reliability was not calculated for confidence on the PICT because measurement included only a single item.

Descriptive statistics (Table 1) show overall performance as a percentage was much higher on the VLT than the PICT, while the margin of miscalibration on the VLT was much smaller. In addition, the negative VLT values indicate general underconfidence, as opposed to the general overconfidence on the PICT. The sample was reduced by one for the VLT because one participant neglected to provide confidence scores.

Table 1: Descriptive statistics for tasks

	VLT	PICT
	N = 63	*N* = 64
	Max. possible score = 120	Max. possible score = 20
Mean performance	89.27 (21.28)	6.44 (4.44)
Performance %	74.39%	32.20%
Mean confidence	83.38 (20.31)	12.45 (3.66)
Δ Confidence - performance	-5.89	6.01
Miscalibration %	-4.90%	30%

Note. Standard deviations in parentheses.

The contrasts are evident in the scatterplots in Figure 4. On the VLT, most performance scores occur above the 50% mark of the graph, while on the PICT, most performance scores occur below. Most VLT data points fall slightly below the identity line (representing perfect calibration) whereas most PICT data points occur above. Moreover, the regression line for the VLT data is nearly coincident with the identity line, which is not the case for the PICT data.

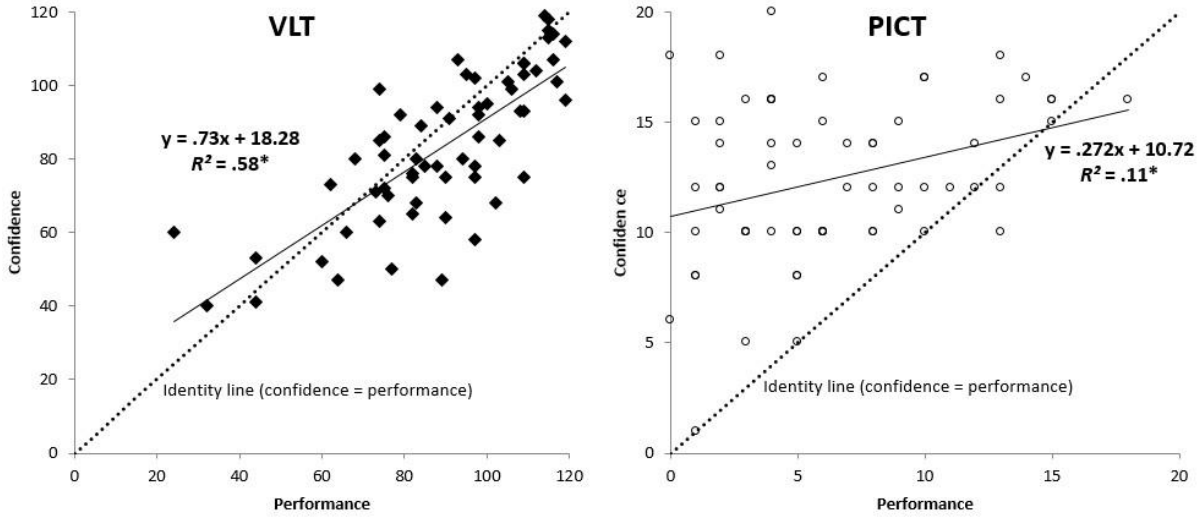

Figure 4: Performance and confidence from the VLT and PICT, with identity line, regression line and equation, and coefficient of determination

Pearson product-moment correlation showed a strong confidence-performance relationship on the VLT, $r(61) = .76$, $p < .001$, compared to the PICT, $r(62) = .33$, $p = .008$, with respective effect sizes that would be characterized as large, $R^2 = .58$, and small to medium, $R^2 = .11$, according to Cohen (1992).

A breakdown of the VLT data (Table 2) shows similar patterns of slight underestimation across frequency-band levels. A one-way ANOVA with difference score as the dependent variable and level as the independent variable found no statistical difference, $F(3,248) = .310$, $p = .818$, partial $\eta^2 = .004$. This means the sample exhibited similar levels of

monitoring accuracy in assessing their knowledge of lower-frequency, higher-frequency, and academic words despite variation in their actual knowledge of such words.

Table 2: Performance, confidence and calibration scores for VLT by level (N = 63)

	2000 level	3000 level	5000 level	AWL level
Mean performance	25.57 (4.43)	22.11 (5.94)	18.75 (6.25)	22.84 (6.11)
Performance %	85.24%	73.70%	62.49%	76.14%
Mean confidence	24.33 (4.52)	20.25 (5.47)	17.52 (5.59)	21.27 (6.14)
Δ Confidence - performance	-1.24	-1.86	-1.22	-1.57
Miscalibration %	-4.13%	-6.19%	-4.07%	-5.24%

Note. Standard deviations in parentheses. Max. possible score = 30 at each level. A*WL* = Academic Word List

The possibility that, on the VLT, miscalibration was stable while performance varied across frequency-band levels suggests monitoring accuracy is not necessarily a function of prior knowledge, as found in some previous research. SRL theory would in fact require high and relatively stable monitoring accuracy, operating independently of variable knowledge within a particular domain, as a facilitative condition for self-directed learning, since learners would need to be able to identify shortcomings in knowledge or performance in order to address them.

To sum up, Study 1 showed that monitoring accuracy can vary considerably according to task complexity and task familiarity. On the simple, familiar task, the sample was much better calibrated, trending toward slight underestimation like the high performers studied by Dunning and associates (e.g., Kruger & Dunning, 1999). On the complex, unfamiliar task, both performance and monitoring accuracy were generally low, with most participants demonstrating considerable overconfidence reminiscent of Dunning's poor performers despite being encouraged to use external resources containing all the necessary information.

The participants' choices regarding use of external resources, which allow inferences about metacognitive control, are taken up next in Study 2.

Study 2

To address the second research question regarding the effects of monitoring accuracy on metacognitive control and task outcomes, an experiment was conducted using the PICT as a pre- and post-test, with two treatment conditions administered in between. One condition included strategy instruction in the use of learner dictionaries with the goal of creating variation in monitoring accuracy among the sample, since strategy instruction can raise awareness of the need for, and free up cognitive resources for, monitoring (Nietfeld & Schraw, 2002). The second was a comparison condition.

Metacognitive control was operationalized as the participants' choice of dictionary type to use on the PICT. Control was linked to task outcomes by attributing points awarded on the PICT to a particular choice of dictionary type. Lexicographical research shows L2 learners generally prefer bilingual dictionaries (Atkins & Varantola, 1998), but, as described above, the PICT was designed to require use of learner dictionaries.

Methods

Participants. The same sample from Study 1 was divided into two groups: one designated the instructed group (n = 32) and the other the uninstructed group (n = 32). The groups were similar in gender balance, variety of L1s, and age (Table 3), while on average the instructed group had spent slightly more time in the U.S. Group size was more than sufficient to meet the target (n = 6) as determined by an *a priori* power analysis calculated in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) using an effect size of 1.32 (Cohen's f) obtained from pilot-study data, a desired power level of .95, and a significance level of .05.

Table 3: Biodata for participants

	Uninstructed (n = 32)	Instructed (n = 32)
Sex	12 Female	14 Female
	20 Male	18 Male
L1	Chinese 23	Chinese 26
	Arabic 4	Malay 2
	Korean 3	Hindi 1
	Thai 1	Spanish 1
	Malay 1	[a Bantu language] 1*
		Korean 1
Time in U.S. (months)	5.7 (8.4)	7.5 (14)
Age	20.1 (2.7)	21.6 (5.9)

Note. Standard deviations in parentheses. *Language not specified for confidentiality reasons.
Materials. The same form of the PICT and web form for post-dictions described in Study 1 were used in Study 2.

Online strategy instruction materials were used to train the instructed group. The materials, consisting of 10 hours of multimedia tutorials and accompanying text-based exercises with immediate feedback, addressed the knowledge and skills involved in using learner dictionaries to identify and correct lexico-grammatical patterning errors. The materials are described in detail in Ranalli (2013a).

An online vocabulary database activity was assigned to the uninstructed group, featuring a web form for inputting new words. The form included fields for several types of lexical information, including "usage information." The database, which was part of the learning

management system (LMS) for the course, also featured an index page listing all words input by the user, which were hyperlinked to individual pages showing the entry for each word. Video instructions accompanying the activity encouraged use of learner dictionaries, but no training or feedback was provided about lexico-grammatical patterning.

Procedures

Matched random assignment was used to control for receptive vocabulary knowledge, which had been found to correlate with strategy instruction outcomes in a pilot study. Participants were ranked according to their VLT scores, and pairs were created based on adjacent rankings. Members of each pair were then randomly assigned to the instructed or uninstructed group. Separate groupings were created in the LMS such that each group saw only the relevant materials.

The administration of the PICT described in Study 1 constituted the pre-test for Study 2. Following the pre-test, the participants worked through their respective materials over five weeks. Then, all participants took the post-test administration of the PICT on a regularly scheduled computer-lab day.

In both administrations of the PICT, screen capture software was used to record video of the participants' use of dictionaries. The video data was coded and transformed into two measures of metacognitive control:

1. *Lookup counts* of the number of individual dictionary searches performed by each participant, divided into two categories: *learner-dictionary lookups* (i.e., lookups in pedagogical dictionaries for learners of English) and *non-learner dictionary lookups* (i.e., lookups in bilingual dictionaries, "native speaker" dictionaries, or non-dictionary tools such as search engines). When footage showed a particular word, phrase, or sentence being searched, a lookup was recorded in the appropriate category, except in cases where the same search had already been conducted in the same resource in the same administration of the PICT.

2. *Lookup points* achieved on the pre- or post-test; that is, points attributable to use of a learner dictionary, another type of reference, or else to pre-existing knowledge, intuition, or guesswork. If footage showed explicit information from any type of dictionary used in a correct or partially correct response, the resulting points were coded either as *learner dictionary lookup points* or *non-learner dictionary lookup points*. Responses that could not be traced to use of an online resource were coded as *null lookup points*.

Results and discussion

For the post-test administration of the PICT ($N = 64$), reliability was excellent, $\alpha = .86$. Tests of the assumptions of equality of covariance matrices and error variances were satisfactory, but histogram and Q-Q plot analyses indicated non-normality in the post-test performance data. A base-10 logarithmic transformation, which is recommended in cases of substantial

positive skew (Tabachnick & Fidell, 2012), failed to resolve these issues, so the analysis was conducted with untransformed data while recognizing the potential for loss in statistical power.

Monitoring accuracy as represented by calibration on the PICT. Descriptive statistics (Table 4) showed low scores for both groups on the pre-test as well as overconfidence in excess of 30%. On the post-test, actual performance for both groups increased, by more than 100% in the case of the instructed group but only by about 20% for the uninstructed group. Miscalibration still took the form of overconfidence for both groups, but while it decreased for the instructed group, a slight increase was seen among the uninstructed group.

Table 4: Descriptive statistics for pre- and post-test administrations of the PICT

	Pre		Post	
	Uninstructed	Instructed	Uninstructed	Instructed
	$n = 32$	$n = 32$	$n = 32$	$n = 32$
Mean performance	6.03 (3.8)	6.84 (5.03)	7.22 (4.94)	13.88 (4.96)
Performance %	30.2%	34.2%	36.1%	69.4%
Mean confidence	12.22 (3.89)	12.69 (3.45)	14.06 (3.65)	15.25 (3.69)
Δ Confidence - performance	6.19	5.84	6.84	1.38
Miscalibration %	30.9%	29.2%	34.2%	6.9%

Note. Standard deviations in parentheses. Max. possible score = 20.

To test for changes in the uninstructed group's performance and confidence, their difference scores were isolated and submitted to paired-samples t-tests. The first test, comparing pre-test performance ($M = 6.03$, $SD = 3.80$) to post-test performance ($M = 7.22$, $SD = 4.94$), was not statistically significant, $t(31) = -1.536$, $p = .135$, two-tailed, $d = .27$. The second test comparing calibration at pre-test ($M = 6.19$, $SD = 4.53$) to calibration at post-test ($M = 6.84$, $SD = 4.78$) also showed no difference, $t(31) = -.799$, $p = .430$, two-tailed, $d = .139$.

In the visual analysis (Figure 5), the instructed group data clustered above 50% on the performance scale and closer to the identity line, while much of the uninstructed group data occurred below the 50% mark and higher above the identity line. Correlational analysis showed a stronger performance-confidence relationship among the instructed group $r(30) = .605$, $p < .001$, compared to the uninstructed group, $r(30) = .41$, $p = .019$, with effect sizes that would be characterized as large ($R^2 = .37$) and medium ($R^2 = .17$), respectively.

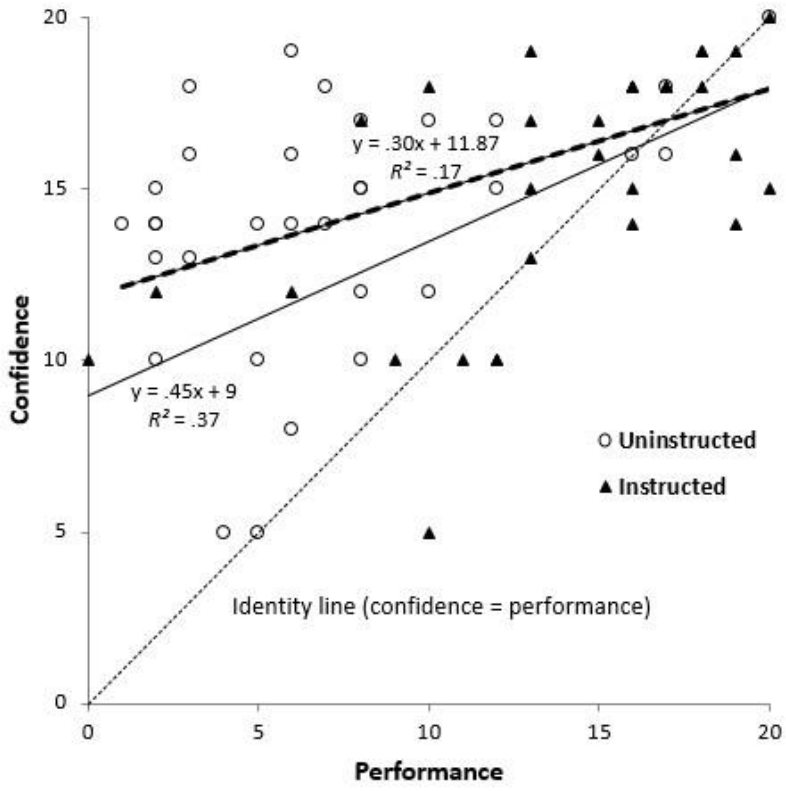

Figure 5: Performance and confidence for both groups at post-test, with identity line, regression line and equation, and coefficient of determination

Metacognitive control as represented by dictionary lookups on the PICT. Descriptive statistics for lookup counts (Table 5) showed similar usage patterns between groups at pre-test, with both groups performing less than three lookups in learner dictionaries on average. At post-test, the uninstructed group showed a slight increase in learner dictionary lookups and a slight decrease in non-learner dictionary lookups, whereas the instructed group heavily favored learner dictionary lookups.

Table 5: Descriptive statistics for dictionary lookup counts

	Pre		Post	
	Uninstructed	Instructed	Uninstructed	Instructed
	n = 32	*n* = 32	*n* = 32	*n* = 32
Learner dictionary lookups	2.75 (3.94)	2.63 (3.91)	3.34 (4.71)	9.53 (4.6)
Non-learner dictionary lookups	3.69 (4.9)	3.97 (4.74)	2.84 (3.73)	0.16 (0.52)

Note. Standard deviations in parentheses

To test for significance, the lookup data were modeled using a Poisson regression to represent the mean rate at which participants performed lookups in learner dictionaries, with learner

dictionary lookups as the response variable, group and time as predictors, and VLT scores as a covariate. Poisson regressions are appropriate when data represent counts and large count outcomes are rare (Kutner, Nachtsheim, Neter, & Li, 2005). Pairwise comparisons showed that the mean difference in learner dictionary lookups from pre-test to post-test was not significantly different for the uninstructed group (M = -.58, 95% CI: -1.92,3.09, d = .14), in contrast to the instructed group (M = -6.79, 95% CI: -10.34,-3.24, d = 1.62).

Control's effects on task outcomes as represented by dictionary lookup points

Descriptive statistics (Table 6) for dictionary lookup points showed both groups scored most points at pre-test without using references. The uninstructed group's averages showed a slight change at post-test, with more points from learner dictionary lookups and fewer from non-learner dictionary and null lookups, while the instructed group's points from learner dictionary lookups increased six fold.

Table 6: Descriptive statistics for dictionary lookup points (max. possible score = 20)

	Pre		Post	
	Uninstructed n = 32	Instructed n = 32	Uninstructed n = 32	Instructed n = 32
Learner dictionary lookup points	1.16 (2.69)	1.78 (3.63)	2.94 (5.3)	11.06 (6.02)
Non-learner dictionary lookup points	0.44 (1.04)	0.59 (1.39)	0.41 (1.01)	0 (0)
Null lookup points	4.44 (2.72)	4.5 (3.54)	3.88 (2.65)	2.81 (3.54)

Note. Standard deviations in parentheses

To test differences in the uninstructed group's lookup points for statistical significance, I used Hotelling's T^2, a generalization of Student's t for multivariate data used when there are multiple dependent variables likely to be correlated with each other and the independent variable has only two levels (Tabachnick & Fidell, 2012). A two-sample Hotelling's T^2 test was run using learner dictionary lookup points, non-learner dictionary lookup points and null lookup points as dependent variables and time as the independent variable. No difference was found, $T^2(3,60) = 1.03$, $p = .385$.

It is notable that, despite the difference in magnitude, post-test gains for both groups are attributable to increased exploitation of learner dictionaries. Examination of the raw data showed that 78% of the uninstructed group's post-test learner dictionary lookup points were contributed by only six individuals, whose average total post-test score, M = 15.67, SD = 3.14, greatly exceeded that of the remainder of the group, M = 5.27, SD = 2.67. This cohort performed most of the learner dictionary lookups at post-test, M = 10.8, SD = 1.7, versus M = 1.62, SD = 3.2 for the rest of the group. Furthermore, their calibration scores also showed

better monitoring accuracy, M = +1.33, SD = 2.25 versus M = +8.12, SD = 4.28, compared to their groupmates.

How did this group manage to improve?[4] It is unlikely this resulted from monitoring and adaptation while engaged in the PICT at post-test. Rather, changes probably occurred between pre- and post-test during the vocabulary database activity. A review of their entries confirmed that these participants used learner dictionaries to research usage information as they had been instructed to, which appears to have informed their understanding of the task at post-test as well as the evaluative criteria they used to monitor their task engagement. This explanation accords with a model of self-regulated learning that posits broader forms of monitoring following or between tasks that address the coordination of information across learning events, with control realized in "large scale adjustments to the student's understandings about the task, goals, plans, and tactical engagement" (Winne & Hadwin, 1998, p. 285).

To summarize, Study 2 showed that inaccurate monitoring is associated with metacognitive control that is misaligned to task demands, resulting in suboptimal performance. Among the instructed group participants, both performance and monitoring accuracy improved following strategy instruction, with performance gains traceable to the participants' strategic choices regarding dictionary use. In addition, a small cohort of uninstructed group participants who performed well also demonstrated higher monitoring accuracy and similar patterns of metacognitive control. In contrast, most uninstructed group participants showed no improvements at post-test in either performance or monitoring accuracy and no substantive changes in control. In other words, their monitoring was poorly calibrated to their actual performance, and this was reflected in their choice of strategies unsuited to the requirements of the task.

General discussion

The purpose of this paper was to corroborate previous findings regarding connections among metacognitive monitoring, metacognitive control, and learning outcomes in order to demonstrate their relevance to instructed SLA. To this end, two studies were undertaken using tasks and measures related to L2 vocabulary learning and use. Study 1 showed that monitoring accuracy varied significantly across tasks that differed in complexity and familiarity. On the simpler, familiar task, monitoring accuracy was independent of performance and characterized by slight underestimation. On the complex, unfamiliar task, performance and monitoring accuracy were largely uncorrelated; monitoring accuracy was generally poor and showed considerable overestimation. In Study 2, learners who performed poorly at both pre-test and post-test on the complex, unfamiliar task were also found to have monitored poorly, and to have selected strategies misaligned to task requirements. By contrast, those whose performance improved at post-test, either through strategy instruction or by engaging in repeated use of learner dictionaries, also showed higher monitoring accuracy and patterns of metacognitive control aligned to task demands.

The findings suggest two main conclusions. First, monitoring accuracy can vary considerably among L2 learners, even within the same domain of language knowledge, depending on

features of the task. Second, suboptimal task performance in which L2 learners employ inappropriate strategies may be attributable to inaccurate monitoring. Taken together, the conclusions point to the relevance of monitoring accuracy and its effects on metacognitive control to L2 learning and use.

These metacognitive phenomena can help explain previous L2 research that showed unsuccessful learners to be unaware of how their strategic choices were undermining their task performance (e.g., Macaro, 2014; Vann & Abraham, 1990). Learners who inaccurately assess their performance or understanding may not detect problems in their approach to learning and thus fail to enact control strategies better suited to task requirements. While contemporary learners have unprecedented opportunities to engage in self-directed L2 learning, such learning may be undermined by inaccurate monitoring in situations where external feedback is limited or unavailable.

Implications

The findings have implications for research, theory, and pedagogy. First, L2 research should give more attention to monitoring accuracy. Researchers should consider not only how learners perform on tasks, but how they evaluate themselves as having performed, and how this informs their strategic choices and achievement. Learners who are too confident in their ability may stop engaging in a learning task prematurely. Conversely, learners who are underconfident may not undertake learning tasks out of unfounded fears of failure.

The interplay of monitoring and control should also inform theorizing about metacognition and self-regulated learning in instructed SLA. Dörnyei and colleagues (Dörnyei & Ryan, 2015; Tseng, Dörnyei, & Schmitt, 2006) have proposed a model of self-regulation akin to volition that has inspired a number of L2 studies, but this model posits self-regulation as a trait underlying strategic behavior. A more dynamic, process-oriented view of self-regulation based on monitoring and control operating across all stages of a learning event (Winne & Hadwin, 1998) provides a potentially more powerful way of explaining and predicting aspects of L2 learners' engagement in tasks and the outcomes that result (see discussion in Ranalli, 2012b).

The most obvious pedagogical implications are in the area of learner autonomy, which is often cited by L2 researchers and practitioners as an unqualified good and the ultimate goal of instructed SLA. While researchers of learner autonomy have made frequent calls to have L2 learners take more control over their learning, particularly Benson (2001, 2010), empirically-supported understanding of when and how learners may be prepared to do so have lagged behind advocacy (Lewis & Vialleton, 2011). Because monitoring accuracy can be assessed relatively easily, as attested by the present investigation, it could provide a practical measure of preparation for autonomy in many domains of L2 learning as well as a solution for conundrums about the practicality and appropriateness of measuring autonomy directly (see discussion in Benson, 2010). Raising learners' awareness of monitoring accuracy could also benefit self-regulated learning. Performance and calibration information could be provided in combination as feedback about self-regulation, which some research suggests may be the most powerful form of feedback (Hattie & Timperley, 2007).

Limitations and future research

This study has focused on a single domain of L2 learning: vocabulary. Additional studies in other domains are needed to see if the findings transfer and under what conditions. The present study also relied on large-grained measures of monitoring accuracy in the form of global post-dictions of performance. Future studies should also employ finer-grained measures, such as local confidence judgments after each item on a task, to compare the accuracy, utility, and reliability of global versus local confidence judgements.

Importantly, this study has only documented issues with the accuracy of L2 learners' monitoring and not the factors that underlie inaccurate monitoring. Qualitative research making use of verbal protocols should be undertaken to determine what cues or factors contribute to learners' metacognitive judgments of performance or knowledge and the extent to which they originate in individual, contextual, or task characteristics.

In addition, our particular sample may have influenced the generalizability of the results. Some research has found Chinese-speaking participants to be more confident and overconfident than other ethnic groups (Trofimovich et al., 2016; Yates, Lee, & Shinotsuka, 1996). Future studies should be conducted with students from other backgrounds as well as educational levels, programs of study, and instructional contexts (e.g., second versus foreign language).

Conclusion

In an undertaking as lengthy, uncertain, and dependent on individual initiative as learning a second or foreign language, learners will be well served by an accurate internal compass to guide their decision-making and progress. Conversely, their time and efforts may be subverted without them even knowing it by inaccurate monitoring and the faulty strategic choices that result from it. More work is needed in applied linguistics to understand the threat that inaccurate monitoring poses to the L2 learning enterprise as well as the antecedents of inaccurate monitoring and the types of interventions that can mitigate it.

NOTES

1. Phakiti and other language testers typically use the term *self-assessment* to describe general evaluations of one's L2 abilities or level of achievement over a course of study, in contrast to monitoring judgments made regarding particular performances, which Phakiti calls *calibration*. In this paper, I use *self-assessment* to refer to these more specific judgments, reflecting usage of the term in the psychology and educational research that I cite.

2. Global confidence judgments (i.e., made after completion of a task or sections of a task) were used in this study instead of local confidence judgments (i.e., made after every item on the task) for the following reasons. First, global judgements have been found to be more accurate than local judgments (Schraw, 1994; Nietfeld, Cao, & Osbourne, 2005). Second, statistical calculation of local confidence on the VLT

would have been problematic insofar as the test format (three stems and six options in each set) meant confidence on any item would not be independent of the other two items in the set. And finally, holistic judgments appear to be more relevant to the forms of informal, self-directed L2 learning of concern here than item-by-item judgments, which are more germane to formal testing situations.

3. I confirmed with searches of all node words in the PICT that the lexico-grammatical patterning information needed to complete the task was only available in the learner dictionaries, as opposed to the other dictionaries to which links were provided. The learners were free to use any dictionary they wanted, including others not linked to from the PICT question pages. It is therefore possible that they may have been able to find pattern information in a non-learner dictionary or other online reference. However, a distinguishing feature of learner dictionaries is that they provide information about lexico-grammatical patterning that is indexed according to particular senses of a word (Ranalli & Nurmukhamedov, 2014). Without relevant information about the appropriate sense, pattern information may be of questionable value.

4. A reviewer questioned whether random selection of learner dictionaries on the PICT could have accounted for the six discrepant cases in the uninstructed group. This seems unlikely, as lexicographical research has shown that, on the basis of inaccurate self-assessments of their vocabulary knowledge, learners tend to ignore collocational information even when it is provided (Frankenberg-Garcia, 2011; Laufer, 2011) unless they have had language awareness raising about these lexical properties (Lew, 2011). Such awareness raising was included in the strategy instruction, as described in Ranalli (2013b).

References

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 138*(3), 432-447. doi: 10.1037/a0015928

Atkins, B. T. S., & Varantola, K. (1998). Language learners using dictionaries: The final report of the EURALEX/AILA research project on dictionary use. In B. T. S. Atkins (Ed.), *Using dictionaries: Studies of dictionary use by language learners and translators* (pp. 21-82). Tübingen: Max Niemeyer Verlag GmbH.

Benson, P. (2001). *Teaching and researching autonomy in language learning*. Harlow: Pearson.

Benson, P. (2007). Autonomy in language teaching and learning. *Language Teaching, 40*(01), 21-40. doi: doi:10.1017/S0261444806003958

Benson, P. (2010). Measuring autonomy: Should we put our ability to the test? In A. Paran & L. Sercu (Eds.), *New perspectives on language and education: Testing the untestable in language education* (pp. 77-97). Clevedon, GBR: Multilingual Matters.

Chamot, A. U. (1987). The learning strategies of ESL students. In A. L. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 71-83). Hemel Hempstead: Prentice Hall.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155

Cumming, A. (1989). Writing expertise and second-language proficiency. *Language Learning, 39*(1), 81-135. doi: 10.1111/j.1467-1770.1989.tb00592.x

Davies, M. (2008). The Corpus of Contemporary American English: 425 million words, 1990-present. Available online at http://corpus.byu.edu/coca/

de Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*(4), 245-252. doi: 10.1016/j.learninstruc.2012.01.003

De Silva, R. (2015). Writing strategy instruction: Its impact on writing in a second language for academic purposes. *Language Teaching Research*, *19*(3), 301-323. doi: 10.1177/1362168814541738

Dörnyei, Z. (2001). *Teaching and researching motivation*. Harlow: Longman.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. Mahwah, NJ: Lawrence Erlbaum Associates.

Dunlosky, J., & Ariel, R. (2011). The influence of agenda-based and habitual processes on item selection during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(4), 899-912. doi: 10.1037/a0023064

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69-106.

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84*(1), 5-17.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*(1), 98-121. doi: DOI: 10.1016/j.obhdp.2007.05.002

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fitzgerald, J. T., White, C. B., & Gruppen, L. D. (2003). A longitudinal study of self-assessment accuracy. *Medical Education, 37*(7), 645-649. doi: 10.1046/j.1365-2923.2003.01567.x

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *The American Psychologist, 34*(10), 906-911.

Folse, K. S. (2010). Is explicit vocabulary focus the reading teacher's job? *Reading in a Foreign Language, 22*(1), 139-160.

Frankenberg-Garcia, A. (2011). Beyond L1-L2 equivalents: Where do users of English as a foreign language turn for help? *International Journal of Lexicography, 24*(1), 97-123.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506-528.

Glenberg, A., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*(1), 84-93. doi: 10.3758/bf03197714

Hattie, J., & Timperley, H. S. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. doi: 10.3102/003465430298487

Hauck, M. (2005). Metacognitive knowledge, metacognitive strategies, and CALL. In J. L. Egbert & G. M. Petrie (Eds.), *CALL research perspectives* (pp. 65-86). Mahwah, NJ: Lawrence Erlbaum Associates.

Hines, J. C., Touron, D. R., & Hertzog, C. (2009). Metacognitive influences on study time allocation in an associative recognition task: An analysis of adult age differences. *Psychology and Aging, 24*(2), 462-475. doi: 10.1037/a0014417

Hunston, S., Francis, G., & Manning, E. (1997). Grammar and vocabulary: Showing the connections. *ELT Journal, 51*(3), 208-216.

Kormos, J., & Csizér, K. (2014). The interaction of motivation, self-regulatory strategies, and autonomous learning behavior in different learner groups. *TESOL Quarterly, 48*(2), 275-299. doi: 10.1002/tesq.129

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*(6), 1121-1134.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. Boston: McGraw-Hill.

Lai, C. (2013). A framework for developing self-directed technology use for language learning. *Language Learning & Technology*, *17*(2), 100-122.

Lai, C., Shum, M., & Tian, Y. (2014). Enhancing learners' self-directed use of technology for language learning: The effectiveness of an online training platform. *Computer Assisted Language Learning*, 1-21. doi: 10.1080/09588221.2014.889714

Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography, 24*(1), 29-49.

Lew, R. (2011). Studies in dictionary use: Recent developments. *International Journal of Lexicography, 24*(1), 1-4.

Lewis, T., & Vialleton, E. (2011). The notions of control and consciousness in learner autonomy and self-regulated learning: A comparison and critique. *Innovation in Language Learning and Teaching, 5*(2), 205-219. doi: 10.1080/17501229.2011.577535

Mihalca, L., Mengelkamp, C., & Schnotz, W. (2017). Accuracy of metacognitive judgments as a moderator of learner control effectiveness in problem-solving tasks. *Metacognition and Learning,* 1-23. doi: 10.1007/s11409-017-9173-2

Macaro, E. (2014). Reframing task performance. In H. Byrnes & R. Manchon (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 53-77). Amsterdam: John Benjamins.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In H. B. Gordon (Ed.), *Psychology of Learning and Motivation* (Vol. Volume 26, pp. 125-173): Academic Press.

Nietfeld, J. L., Cao, L., & Osbourne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*(1), 7-28.

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research, 95*(3), 131-142.

Oxford, R. (1989). Use of language learning strategies: A synthesis of studies with implications for strategy training. *System, 17*(2), 235-247.

Oxford, R. L. (2017). *Teaching and researching language learning strategies: Self-regulation in context* (2nd ed.). New York, NY: Routledge.

Phakiti, A. (2005). An empirical investigation into the nature of and factors affecting test takers' calibration within the context of an English Placement Test (EPT). *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 27-46.

Phakiti, A. (2016). Test takers' performance appraisals, appraisal calibration, and cognitive and metacognitive strategy use. *Language Assessment Quarterly*, *13*(2), 75-108. doi: 10.1080/15434303.2016.1154555

Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency – Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction, 44*(Supplement C), 31-40. doi: https://doi.org/10.1016/j.learninstruc.2016.01.012

Pieschl, S. (2009). Metacognitive calibration—An extended conceptualization and potential applications. *Metacognition and Learning, 4*(1), 3-31. doi: 10.1007/s11409-008-9030-4

Ranalli, J. (2012a). *The VVT Project: A web-based platform for strategy instruction and research into self-regulated learning of L2 vocabulary.* (PhD), Iowa State University, Ames, IA.

Ranalli, J. (2012b). Alternative models of self-regulation and implications for L2 strategy research. *Studies in Self-Access Learning Journal, 3*(4), 357-376.

Ranalli, J. (2013a). Designing online strategy instruction for integrated vocabulary depth of knowledge and web-based dictionary skills. *CALICO Journal, 30*(1), 16-43.

Ranalli, J. (2013b). Online strategy instruction for integrating dictionary skills and language awareness. *Language Learning & Technology, 17*(2), 75-99.

Ranalli, J., & Nurmukhadev, U. (2014). Learner dictionaries In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Malden, MA: Wiley-Blackwell.

Ross, L., Greene, D., & House, P. (1977). The false consensus phenomenon: An attributional bias in self-perception and social perception processes. *Journal of Experimental Social Psychology, 13*(3), 279-301.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19*(2), 143-154. doi: http://dx.doi.org/10.1006/ceps.1994.1013

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston: Allyn and Bacon.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66-73. doi: 10.1037/0022-0663.95.1.66

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition*, 19(1), 122-140. doi: 10.1017/s1366728914000832

Tseng, W. T., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. *Applied Linguistics, 27*(1), 78-102. doi: 10.1093/applin/ami046

van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merriënboer, J. J. G. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction, 24*, 15-25. doi: http://dx.doi.org/10.1016/j.learninstruc.2012.08.005

Vandergrift, L. (1997). The comprehension strategies of second language (French) listeners: A descriptive study. *Foreign Language Annals, 30*(3), 387-409. doi: 10.1111/j.1944-9720.1997.tb02362.x

Vann, R. J., & Abraham, R. G. (1990). Strategies of unsuccessful language learners. *TESOL Quarterly, 24*(2), 177-198.

Walczyk, J. J., & Hall, V. C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology, 81*(3), 435-437. doi: 10.1037/0022-0663.81.3.435

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *Metacognition in educational theory and practice.* (pp. 277-304): Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Yates, J. F., Lee, J. W., & Shinotsuka, H. (1996). Beliefs about overconfidence, including its cross-national variation. *Organizational Behavior and Human Decision Processes, 65*(2), 138-147. doi: 10.1006/obhd.1996.0012

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement* (pp. 1-38). Mahwah, NJ: Erlbaum.