Monitoring comprehension in a foreign language: Trait or skill?

Lilach Temelman-Yogev, Tami Katzir & Anat Prior

Metacognition and Learning

ISSN 1556-1623
Volume 15
Number 3

Metacognition Learning (2020)
15:343-365
DOI 10.1007/s11409-020-09245-5

Springer

Metacognition and Learning (2020) 15:343–365
https://doi.org/10.1007/s11409-020-09245-5

Monitoring comprehension in a foreign language: Trait or skill?

Lilach Temelman-Yogev[1] · Tami Katzir[1] · Anat Prior[1]

Received: 12 August 2019 / Accepted: 18 September 2020 / Published online: 2 October 2020

Abstract

Success in higher education is highly dependent on students' ability to efficiently read and comprehend large amounts of text in the speaker's first/native language (L1) and also in a Foreign Language (FL). Good text comprehension requires readers to implement a variety of metacognitive processes in order to self-regulate understanding. However, most readers are inaccurate when monitoring their own comprehension level, in the native language. Several studies have investigated FL comprehension monitoring, mostly using self-report measures. The current study further explored the relationship between L1 and FL comprehension monitoring through the paradigm of 'calibration of comprehension' (Glenberg and Epstein in *Journal of Experimental Psychology: Learning, Memory and Cognition, 11*, 702-718, 1985). Specifically, 145 university students read texts in each language, answered comprehension questions and rated their confidence. Absolute and relative monitoring accuracy was calculated (bias and resolution, respectively) to study whether comprehension monitoring processes are trait-oriented (shared across languages and domains) or skill-oriented (dependent on language proficiency level). Results suggested that absolute monitoring accuracy is both trait and skill oriented. On the one hand, confidence ratings and bias were significantly correlated across L1, FL and a non-verbal task, suggesting trait-orientation. On the other hand, only individuals who were highly proficient in the FL shared their absolute monitoring skills between the languages, supporting the notion of a skill orientation. Relative monitoring was not associated across tasks or languages. Theoretical and practical implications for effective instruction and learning methods are discussed.

Keywords Comprehension monitoring · Confidence ratings · Calibration of comprehension · English as a foreign language · Bilingualism · L2

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11409-020-09245-5) contains supplementary material, which is available to authorized users.

✉ Anat Prior
aprior@edu.haifa.ac.il

Extended author information available on the last page of the article

Reading and comprehending are active processes which require the introspective awareness of individuals to meaning construction, through which the reader builds a semantic network of notions to form a "text model of comprehension"(Grabe 2014; Kintsch 2012; RAND model; Snow 2002). This high order processing, called metacognition, was defined by Flavell (1979) as "cognition about cognition", referring to the *knowledge* students have about the cognitive processes involved during learning; the *regulation* procedures taken to control the learning process (such as planning, monitoring or evaluating); and the cognitive or affective *experiences* that occur while engaging in cognitive activity. The three components are intertwined so that metacognitive regulation is based on metacognitive knowledge and experiences; and metacognitive knowledge is modified by the inner feedback received from metacognitive regulation and experiences (Zhang and Zhang 2019). Thus, monitoring of reading comprehension consists of a range of skills and abilities, which enable readers to engage in different problem-solving procedures while reading (Grabe 2014). Indeed, comprehension monitoring abilities were found to be good predictors of text comprehension by their own right (e.g., Grabe and Stoller 2011; Hudson 2007; Kolić-Vehovec and Bajšanski 2007). Additionally, several strategies were found to positively influence comprehension monitoring accuracy, such as rereading (Rawson et al. 2000), summarizing (Thiede and Anderson 2003), delayed generation of keywords after reading (Thiede et al. 2003), and a self-explanation strategy (Griffin et al. 2008). Yet, the relationship between comprehension monitoring and successful learning is indirect. Specifically, accurate monitoring is supposed to produce more effective regulation, and regulation in turn produces better learning outcomes (Thiede et al. 2003).

However, research has demonstrated that comprehension monitoring accuracy is generally low (Dunlosky and Rawson 2012; Dunlosky and Lipko 2007; Maki 1998). Individual differences between poor and good comprehenders in comprehension monitoring have been widely documented, suggesting that good comprehenders are also good in monitoring their knowledge, as opposed to poor comprehenders who also frequently overestimate their comprehension performance as better than it actually is. This tendency has been observed in children (e.g., Ehrlich et al. 1999), in adolescents (Kleider-Tesler et al. 2019) as well as in adults (e.g., Maki and Berry 1984; Maki et al. 2005). However, since poor monitoring could be the cause or the result of poor comprehension, the causal relationship between comprehension monitoring and test performance cannot be directly inferred (Thiede et al. 2003).

All of the above-mentioned studies were conducted in students' native language (L1). Less research has focused on comprehension monitoring in English as a foreign language, as often practiced in our global world today. Previous research on the subject concentrated on self-report questionnaires of metacognitive knowledge or strategies (e.g., Kolić-Vehovec and Bajšanski 2007; Sarac and Tarhan 2009; Taki 2016; Van Gelderen et al. 2004; for a review see Zhang and Zhang 2019). Other studies used error detection or/and think aloud protocols to measure foreign language (FL) comprehension monitoring (e.g., Block 1992; Khonamri and Mahmoudi Kojidi 2011). However, none of these studies focus on online measures of comprehension monitoring (though see Silawi Shalhoub-Awwad & Prior, 2020). The current study investigated the relationship between L1 and FL comprehension monitoring through the paradigm of "calibration of comprehension" (Glenberg and Epstein 1985) usually used in native language (L1) contexts. Specifically, our aim was to explore whether comprehension monitoring is shared across languages (trait oriented) or is a language specific skill (skill oriented), dependent on proficiency?

Reading comprehension and monitoring across L1 and FL

Similarities in the course of reading development in L1 and FL are apparent in studies which examined early development of single word reading abilities in L1 and FL (for a review see, Lesaux et al. 2006), as well as parallels in low level linguistic skills such as phonological awareness, letter identification and rapid naming (e.g., Abu-Rabia and Siegel 2003; Chiappe et al. 2002; Kahn-Horwitz et al. 2005; Lesaux and Siegel 2003). However, along with the shared cognitive and linguistic component skills, L1 and FL reading abilities still differ from one another due to the simple fact that FL reading cannot rely on a broad infrastructure of spoken language, as the native language does (Grabe 2014; Grabe and Stoller 2011; Hudson 2007). This discrepancy is emphasized when high levels of text reading and comprehending are required. Students learning in a FL face a wide range of difficulties, such as lack of fluency in reading, lack of vocabulary and grammatical background knowledge, and limited experience with extensive reading (Geva and Farnia 2012; Grabe and Zhang 2013; Prior, Zeltsman-Kulick & Katzir, 2020). Specifically, most studies show that a critical mass of FL knowledge is important in order to achieve efficient reading and comprehension (Bernhardt and Kamil 1995; Laufer and Ravenhorst-Kalovski 2010; Lervåg and Aukrust 2010). Research has suggested that for 8th to 10th graders, vocabulary knowledge is the best predictor of foreign language reading comprehension. However, from grade 10 and onwards, the relative importance of vocabulary decreases, and metacognitive knowledge becomes more prominent (Hudson 2007). Thus, it seems that at higher levels of reading, reliance on more complex skills such as comprehension monitoring, knowledge of text structure and inference making, are required for proper understanding. These demands are common to both first and foreign languages (Hudson 2007).

Numerous studies have investigated metacognitive knowledge and strategy use in FL learning (e.g., Wenden 1999; for a recent review, Zhang and Zhang 2019), focusing especially on the relationship between L1 and FL metacognitive strategies. Some studies that compared strategy use between L1 and FL reading comprehension found that FL readers tend to use lower-level linguistic processing strategies (such as vocabulary meaning construction) rather than higher-level comprehension strategies (such as inference making and linkage to prior knowledge) (e.g., Stevenson et al. 2007). Other studies demonstrated that proficiency in the FL was associated with the type of strategy used, such that proficient FL readers presented a similar pattern of strategy use in both L1 and FL, but less proficient readers used different strategies in FL than in L1 and employed fewer strategies in FL than in L1 (e.g., Block 1992; Khonamri and Mahmoudi Kojidi 2011; Lin and Yu 2015; Tsai et al. 2010). Additionally, metacognitive knowledge (measured by questionnaires) was associated with improved FL reading comprehension in adolescents (e.g., Trapman et al. 2017; Van Gelderen et al. 2004) as well as in adults (Dabarera et al. 2014; Sheorey and Mokhtari 2001; Taki 2016).

Hence, the extant evidence on the use of metacognitive strategies and metacognitive knowledge across languages supports both trait (shared processes) and skill (language-specific processes) assumptions. Yet, less research was conducted on FL metacognitive regulation processes, and specifically on comprehension monitoring while reading (calibration of comprehension paradigm). Thus, in the current study we will elaborate on this issue.

Measures of comprehension monitoring

One of the most widely used paradigms to assess comprehension monitoring is *calibration of comprehension* (Glenberg and Epstein 1985), in which participants read texts, answer comprehension questions and evaluate their confidence in their answers. This paradigm is comprised of objective variables (performance accuracy), subjective variables (monitoring of comprehension and confidence ratings), and indices that are created by the relationship between them, such as bias and resolution (Ackerman and Leiser 2014). In the current study, the analysis focuses on monitoring of L1 and FL reading comprehension as subjectively evaluated by confidence ratings of performance and their corresponding monitoring accuracy, calculated by two main measures: absolute and relative (Nelson 1996). *Absolute* monitoring accuracy measures reflect the difference between the average of judgments for all test items and the average of performance on these same items, and is usually referred to as *bias* (Stankov et al. 2015). This calculation creates the distinction between overconfidence; when a person is more confident than correct, versus underconfidence; when a person is more correct than confident (Buratti et al. 2013). *Relative* monitoring accuracy measures are based on the discrimination between rated confidence on correct and incorrect items, also referred to as *resolution*. It correlates the judgments with the performance of each participant. When variables are of ordinal scale level, the gamma coefficient is used (e.g., Ackerman and Goldsmith 2011; Dinsmore and Parkinson 2013; Mengelkamp and Bannert 2010). A participant shows high resolution when they tend to have higher confidence judgements on items for which they give correct responses, and lower confidence judgements on items for which they give incorrect responses. Several studies have suggested that different measures of monitoring accuracy assess different underlying aspects of metacognitive monitoring (e.g. Koriat 2007; Maki 1998; Maki et al. 2005; Schraw 2009). For this reason, there might be a dissociation between absolute and relative monitoring scores. Namely, resolution can be very high while bias is very poor and vice versa (e.g., Griffin et al. 2009; Koriat et al. 2002; Maki et al. 2005).

Both measures of comprehension monitoring (absolute and relative) are based on participants' self-confidence ratings (Stankov et al. 2015) and can be submitted at two central time points during task performance; before the task is conducted, thus requiring prediction or judgment of learning (JOL); and during or after the task is conducted, which requires evaluation of confidence in performance after an item in a test has been answered, also referred to as post-diction judgments (Schraw 2009). Generally, post-dictions are more accurate than pre-dictions/JOL's (Pieschl 2009), due to the knowledge participants have gained while doing the task, helping them to evaluate performance (Leonesio and Nelson 1990; Maki 1998; Moore et al. 2005).

In the present study, we based our analysis on post-diction confidence ratings given after each question that was answered, hereafter referred to as confidence ratings.

Comprehension monitoring - trait or skill?

An ongoing debate in the field of cognitive and educational psychology concerns the generality and specificity of monitoring processes - whether they are trait oriented or skill oriented. A literature review reveals evidence to support both possibilities.

The *trait oriented approach* relies on the argument that confidence ratings and monitoring accuracy represent a general personality characteristic which is not influenced by task demands

(Kleitman and Stankov 2007; Stankov 1999; Stankov et al. 2012). This claim is supported by studies that calculated monitoring accuracy by either absolute (e.g., Pallier et al. 2002) or relative accuracy measures (Moore et al. 2005), and found that confidence ratings in a specific task do not necessarily correlate with performance accuracy on that task, but rather that confidence ratings and monitoring accuracy of an individual tend to be correlated across different tasks (Moore et al. 2005; Pallier et al. 2002). For instance, Pallier et al. (2002), in a confirmatory factor analysis study, tested individual differences in absolute monitoring accuracy across a wide range of cognitive, perceptual and personality tasks. They found significant correlations between the confidence ratings of participants (post-dictions) regardless of their performance accuracies across tasks. When the absolute accuracy of those confidence ratings was assessed (through the bias index), individual differences were stable across different tasks and a variety of difficulty levels. The decisive conclusion of the researchers was that "humans have a trait that mediates their ability to evaluate the accuracy of their responses" (Pallier et al. 2002, p. 37). In line with these conclusions, using a relative measure of monitoring, Moore et al. (2005) studied comprehension monitoring over different learning trials. Results showed that both pre- and post-diction resolution were significantly correlated with one another and across learning trials, indicating that the ability to monitor comprehension is stable across learning. Additionally, confidence ratings (both pre- and post- diction) by themselves were heavily dependent on general self-perceptions aggregated from past experiences, rather than on the actual performance on a specific task (Moore et al. 2005).

In the area of FL learning, the trait-oriented approach discussed above corresponds with the common underlying cognitive processes hypothesis (Geva and Ryan 1993), which argues that the relationship between the native language (L1) and the foreign language (FL) is based on "shared" cognitive and meta-cognitive abilities. Thus, correlations between parallel tasks in L1 and FL do not have to reflect the action of "transfer" from one language to the other, but rather indicate the sharing of cognitive processes which underlie performance in both languages (Chung et al. 2019). Hence, this framework explains how higher-level abilities, such as comprehension monitoring, which are considered as "language free" abilities, can be shared across the two languages, regardless of proficiency in each language.

Support for this hypothesis derives from three research directions. First, a recent study by Silawi et al. (2020) which examined calibration of comprehension in undergraduate trilinguals (L1-Arabic, L2-Hebrew, L3- English), revealed significant associations in monitoring accuracy of reading comprehension across L1, L2 and L3, regardless of language proficiency. Second, an intervention study targeting adolescent reading comprehension strategies in L1 (Spanish) and L2 (English) through a think-aloud paradigm (Jiménez et al. 1996), showed significant differences in strategy use between good and poor readers rather than between proficient and less proficient language users. Namely, students who developed good comprehension strategies in their first language, also implemented those strategies in their second language. Third, evidence for an opposite generalization of reading comprehension strategies from the less established language (FL) to the mother tongue (L1) has also been documented, in children (Abu-Rabia et al. 2013), and in adults (Schwartz et al. 2013). These indications for bidirectional sharing of comprehension strategies between L1 and FL support the notion that comprehension monitoring might act as a general ability and therefore can be shared between languages.

Alternatively, there is also evidence that supports the *skill oriented approach* of comprehension monitoring, coming from studies in which confidence ratings and monitoring accuracy measures were shown to be dependent on test characteristics or test difficulty (e.g., Maki

et al. 1990; Moore and Healy 2008; Schraw and Roedel 1994; Weaver and Bryant 1995). Thus, studies which calculated monitoring by both absolute and relative accuracy measures, reflected variability in monitoring accuracy according to the measure used. To illustrate, Maki et al. (2005) found that absolute monitoring accuracy in reading comprehension was dependent on students' verbal ability and text difficulty, such that in difficult tasks high verbal ability students were under-confident but low verbal ability students were overconfident. In contrast, relative monitoring accuracy (measured by the gamma coefficient) was not related to verbal ability or test difficulty. This pattern of results is also demonstrated in other studies comparing both relative and absolute accuracies in different tasks using either prediction accuracies (Kelemen et al. 2000) or post-diction accuracies (Mengelkamp and Bannert 2010). Kelemen et al. (2000) reported that the relative meta-memory accuracy was not consistent across different tasks and over time, whereas performance accuracy, confidence ratings and even absolute accuracy were stable (see Appendices B and C in Kelemen et al. 2000). These results were replicated in a monitoring comprehension learning trial study (Mengelkamp and Bannert 2010), indicating that absolute measures (bias and absolute-bias) were stable across learning sessions whereas the relative measures (gamma and *d*a) were not.

In FL learning, the skill-oriented approach can be related to the linguistic interdependence theory (Cummins 1979; Cummins 2012), according to which linguistic transfer from one language to the other is dependent on proficiency in both languages. Aligned with this theory, several studies which explored comprehension monitoring in university students using different measures such as "think aloud" paradigms (e.g., Block 1992; Lin and Yu 2015), error detection (e.g., Han 2012; Han and Stevenson 2008) or a combination of both (e.g., Khonamri and Mahmoudi Kojidi 2011), found that comprehension monitoring was modulated by proficiency. Students performed significantly better on comprehension monitoring tasks or used more advanced metacognitive strategies in L1 than in FL. Specifically, Han (2012) found that more proficient FL readers performed better in comprehension monitoring than less proficient readers. These findings support the hypothesis that FL reading proficiency level may influence comprehension monitoring processes.

Finally, there is also evidence supporting both assumptions of trait and skill (Kasperski and Katzir 2013), suggesting that fourth grade children's comprehension monitoring (using post-diction confidence ratings and indexed by bias) is influenced by both reading ability and personality factors. Low comprehenders had significantly lower confidence ratings and less accurate monitoring than average and high comprehenders, demonstrating that monitoring accuracy was specific to reading ability. However, in the same study individual differences in confidence were found within each level of reading ability.

In sum, it seems that different studies report a variety of findings regarding the orientation of monitoring processes, depending on the timing of the confidence ratings, the measure of accuracy used, age of participants, and the type and language of the task.

The current study

Keeping in mind both theoretical frameworks (Cummins 1979, 2012 vs. Geva and Ryan 1993), the current study was designed to investigate whether comprehension monitoring processes are directed by domain general processes and therefore shared across languages (i.e. trait oriented), or alternatively, language specific and thus dependent on language proficiency (i.e. skill-oriented). Comprehension monitoring was measured with the indices

of bias and resolution (via gamma coefficient) in order to capture the different information each measure provides. Additionally, to explore monitoring beyond the language domain, we used an additional non-verbal task, for a control comparison (Raven 1960). In this way, we can compare confidence ratings and calibration scores in the Raven test with the other two tests (L1 and FL reading comprehension) and compare performance monitoring between the non-verbal and verbal domains. The following research questions guided our study: 1. Do confidence ratings and monitoring accuracy (calculated by bias and resolution), generalize across languages (L1 and FL) and domains (verbal vs. non- verbal)? If confidence ratings and their accuracy generalize across languages and task domains, the trait-oriented approach will be supported. 2. Are confidence ratings and comprehension monitoring (calculated by bias and resolution) modulated by language proficiency (L1, FL)? If the level of proficiency in the language dictates the level of confidence and comprehension monitoring accuracy, the skill-oriented approach will be reinforced.

Method

Participants

The study included 145 undergraduate university students from different departments (mean age24 = .19 years, SD = 2.72, 112 females), native speakers of Hebrew (L1) who had studied English as a foreign language beginning at age 8, or the 3rd grade. All students were healthy, with no history of a learning disability and intact or corrected hearing and vision. All participants were recruited through advertisements, gave informed consent, and were compensated by course credits or payment. The study was approved by the University of Haifa Institutional Review Board (IRB).

Measures

Baseline measures

Language proficiency was assessed in five ways in order to cover various aspects of foreign language proficiency.

I. *Vocabulary knowledge – The Shipley Institute of Living scale* (part A) (Goodman et al. 1974). A 40-item multiple choice synonym test (English version, Shipley (1940) and Zachary (1986); Hebrew version, Gilboa (unpulbished); split half internal consistency reported as 0.87 by Zachary 1991). This scale (0–40) has been successfully used in previous studies with the same population (Prior 2012).
II. *Reading Fluency - Test of Word Reading Efficiency (TOWRE).* (Torgesen et al. 1999, internal consistency ranged from .86 to .98 for word reading efficiency in English and in Hebrew). TOWRE contains real words (104 words) of increasing level of difficulty arranged in four columns. The participant is required to read aloud as many words as possible within 45 s. In the analysis, separate scores were calculated for word reading efficiency in Hebrew and English. The parallel Hebrew version (Katzir et al. 2012; Cronbach's alpha = .95) presented real words without diacritics.

III. *The Language Experience and Proficiency Questionnaire (LEAP-Q).* (Marian et al. 2007; Cronbach's alpha for L1 components = .92, and for L2 components = .88). The parallel Hebrew version (Beznos and Prior 2009) was adapted to this study. The questionnaire includes questions regarding participants' history and context of acquiring the languages they know, present language use, language preference and proficiency.

IV. *English Vocabulary Size Test* (Nation and Beglar 2007) - A computerized version of the test which measures written receptive vocabulary size. The test samples from the most frequent 14,000 word families of English. It consists of 140 items, ten from each 1000 word level. Each word level consists of 10 multiple choice questions. The test words appear in a simple non-defining context and participants are required to choose the most accurate meaning for the test words. The test is terminated when a participant achieved less than 60% accuracy in four consecutive word levels (Beglar 2010; *Rasch item reliability = .96*).

V. *English scores from the Psychometric exam*, an entrance exam to the university (National Institute for Testing and Evaluation). Participants were requested to sign a release form allowing us to access their scores in the English section assessing reading comprehension and vocabulary. This score provides us an additional objective measure of a wider range of language related abilities (English section reliability, Kuder-Richardson-20 formula =. *92*).

Study measures

Reading comprehension: participants read expository texts in each language (L1 and FL) and answered five multiple-choice questions following each text. Prior to initiating the current study, 24 texts and questions in each language (Hebrew and English) were piloted on 128 undergraduate university students (who did not participate in the main study) to ensure that they are appropriate for the study population. Participants' reading comprehension was indexed by the number of questions they answered correctly. Based on the pilot data, 18 texts with 5 questions in each language were selected. The goal was to select texts with a medium level of accuracy, equated as much as possible across the two languages (mean accuracies of the texts were 69.8% for English, 71.6% for Hebrew). Difficulty levels were planned to be equal across languages since previous studies have suggested that absolute monitoring accuracy scores were linked to performance levels (Dunlosky et al. 2015; Pressley and Ghatala 1988). However, texts in English (the foreign language) were by necessity objectively less complex than texts in Hebrew (the native language) – because of the population differences in proficiency, matching performance level required selecting texts of different complexity level (see more on this point in the results section).

This large number of texts was prepared, because the results presented here came from a pre-test session of a large longitudinal study. Thus, the 18 texts in each language were divided into 6 collections of 3 texts each. These collections were roughly equated in difficulty level, length and content. In the pre-test session reported here, each participant was randomly assigned one collection, or 3 texts, in each language, Hebrew and English, followed by corresponding questions. Within each language, the three texts were presented in a random order to each participant.

The average length of the texts was 336 words in Hebrew (range: 157–490 words), and 275 words in English (range: 209–405 words). However, since each participant read three texts in Hebrew and three texts in English, the cumulative number of words in each language was averaged as well; The average length of Hebrew text collections was 1006 words (range: 901–1028). The average length of English texts collections was 824 words (range: 746–878).

The topics of the texts in both L1 and FL included: human behavior, human history and natural phenomena (examples of texts in each language are available as supplementary materials). Prior knowledge of these topics was not directly examined. However, the texts topics relate to general knowledge and were considered by the research team to be of similar familiarity.

Each text was followed by five multiple-choice questions with four possible answers, testing different levels of comprehension such as inference, summarizing, recall of details and vocabulary. The order of the questions and the answers was also randomly determined for each participant. Questions order was randomized to control for order effects in the confidence judgements. The texts remained available to participants while they responded to the questions, such that task performance was not based on memory skills.

Hebrew texts were adapted from the Psychometric exam (National Institute for Testing and Evaluation – NITE, *Cronbach's alpha* = *.88*). English texts were adapted from five different sources to diversify texts structure, content and language; nine texts were selected from the Psychometric and Amir exams (NITE), six texts from the Nelson Denny Reading Test (Brown et al. 1993), two texts from the SAT (Scholar Aptitude Test) (College Board), and one text from the Read Theory workbooks site (Read Theory) (*averaged English texts' Cronbach's alpha* = *.77*). The Hebrew and English texts were each written originally in that language and did not constitute translations of each other.

1. *Judgments of learning*: Immediately after reading the text and before answering the questions, participants were required to predict their future performance on comprehension questions on a continuous scale between 25%- 100%, by answering the question "what is the level of accuracy you predict to have on comprehension questions regarding the text you read?", by moving an indicator on an analog scale using the computer mouse (i.e. prediction).[1]
2. *Confidence ratings*: Following each comprehension question, participants were asked to use a scale ranging from 25% (chance) -100% (completely sure) to answer the question "How confident are you that you responded correctly?" (i.e. post-diction). This scale was chosen since questions had four possible answers; thus it was based on the chance level of a correct response (Pallier et al. 2002). It is acknowledged that when calculating absolute accuracy (bias) the scale of the confidence ratings should be equivalent to the scale of task performance (Dunlosky et al. 2015). However, since the measure of confidence used for the calculation of bias was the average ratings across question and texts in each language, this shortcoming was eliminated. Additionally, we ascertained that in the reading comprehension task in both languages there was no single participant who achieved a comprehension accuracy score below the rate of 25% correct. In the Raven task, 11 subjects had an average accuracy performance below 25% correct and were thus considered as outliers and were omitted from the relevant analysis.

[1] As described in the introduction, these predictive Judgements of Learning were not analyzed in the current manuscript.

Monitoring accuracy was calculated using two measures, absolute and relative:

3. *Absolute calibration accuracy, Bias,* the accuracy of confidence ratings was calculated by the measure of bias. Namely, the accuracy of performance of each text was subtracted from the average confidence rating the participant gave for all questions in that text. The average bias across texts was then calculated for each language.
4. *Relative calibration accuracy, Resolution* measures the extent to which metacognitive judgements distinguish correctly between correct and incorrect items; it is calculated using a within-participant Goodman–Kruskal gamma correlation (Nelson 1996) between item accuracy and confidence judgement.
5. *Bias and resolution in a non-verbal task*-In order to examine general monitoring ability, which is not related to language skills or reading level, five sets of items from three levels of difficulty (total 15 Matrices) were adapted from the *Raven Standard Progressive Matrices* (Raven 1960) (*Cronbach's alpha* = *.75*). The items were presented in an increasing level of difficulty. Following each item, participants were asked to answer the question "How confident are you that you responded correctly?". The same confidence rating scale used in the reading comprehension tasks was applied in this test. As mentioned above, 11 subjects had an average accuracy performance below 25% correct, were thus considered as outliers and were excluded from the relevant analysis. Bias and resolution were calculated as described in the reading comprehension tasks.

Design and procedure

The study entailed one session of 90 min, out of six study sessions of a training study. It was administered individually to each participant in a quiet room at a research lab at the university. Before the first session, participants completed the LEAP-Q questionnaire online. At the beginning of the session, each participant provided informed consent. Next, the reading comprehension texts were presented on a computer screen in each language (Hebrew or English), with language order counterbalanced across participants. Following each text, participants responded to the comprehension questions and rated their confidence level in each answer. While answering the questions, participants could go back and read the text again, but responses to the questions and confidence ratings could not be altered once they were entered by the participant. The presentation of the subsequent text was triggered by the participant. Each participant completed three reading comprehension texts in each language. After completing the reading comprehension task in each language, participants conducted a battery of baseline measures, including reading ability and language proficiency in the same language of the reading comprehension texts, administered in counterbalanced order across participants. Upon the completion of reading comprehension in the second language participants conducted language proficiency tasks and the Raven test (including confidence ratings), in random order.

The following data were recorded during the reading comprehension task: accuracy of responses to the questions and participants' confidence ratings (per text and per question). Presentation and response collection were controlled by E-prime (PST, www.pstnet.com, Schneider et al. 2012, version 3.0).

Results

As part of an initial examination of results, we calculated the mean scores of each participant in all base-line measures of language proficiency, reading comprehension and non-verbal task. Seven participants who had scores which were 2.5 standard deviations above or below the sample mean for comprehension accuracy were excluded from the sample. The final sample included 138 participants. Descriptive statistics for all linguistic measures collected are presented in Table 1. As seen and according to our expectations, students were significantly more proficient in their L1 (Hebrew) than in their FL (English). This pattern was observed in all language proficiency measures.

Table 2 presents descriptive statistics of study measures and their corresponding monitoring accuracies across languages (L1 and FL) and domains (verbal and non-verbal). With regard to comprehension levels across languages, in our pilot study text difficulty levels were well matched across languages. Yet, in the main study sample, differences in comprehension accuracy were detected between the two languages, so that higher accuracy was found in L1 than in FL. Therefore, we cannot rule out the possibility that the difficulty of the assignment is responsible for the effects presented between the languages. There is an inherent difference between L1 and FL, and test difficulty cannot be separated completely from the language in which the task is being transformed. Ercikan and Por (2020) argued that "in reality, variations across assessment versions or across language and sociocultural groups are inevitable." Thus, they suggest that when comparing a construct between languages, establishing the tolerable levels of differences should be defined through empirical evaluations, as was done in the pilot study of the texts.

1. Are confidence ratings and monitoring accuracy (calculated by resolution and bias) shared across languages (L1 and FL) and domains (verbal vs. non-verbal), supporting the trait-oriented approach?

To examine *the trait-oriented approach* suggesting that monitoring is shared across languages and domains, we explored the relationship between confidence ratings, absolute monitoring accuracy (bias) and relative monitoring accuracy (resolution) across the three tasks and languages (reading comprehension in L1 and FL, and the non-verbal Raven matrices), by Pearson correlations. Eleven subjects who performed at less than chance level (average accuracy below 25%) in the Raven task were excluded from this analysis. There were

Table 1 Means (SDs) of language proficiency measures across languages ($N = 138$)

Language proficiency measure	L1 (Hebrew)	FL (English)
LEAP-Q*	9.6 (0.7)	7.2 (1.3)
(scale- 0-10)	Range 5–10	Range 3–10
TOWRE*	80.9 (11.8)	72.54 (8.9)
(Scale- 0-140)	Range 54–103	Range 46–96
Shipley*	27.4 (6.2)	18.6 (3.9)
(Scale- 0-40)	Range 13–39	Range 10–36
English Psychometric scores		115.6 (17.1)
(Scale 0–150)		Range 59–146
Vocabulary Size Test		30.5 (18.1)
(Scale- 0-140)		Range 3–90

*$P < .05$

Table 2 Means (SDs) of study measures and metacognitive monitoring accuracy ($N = 127$)

	Reading comprehension in L1 (Hebrew)	Reading comprehension in FL (English)	Non-verbal (Raven)
Study measures			
Performance accuracy (%)	80.1 (12.4) Range 46–100	69.4 (15.5) Range 27–100	55.5 (20.2) Range 26–100
Confidence ratings (%)	87.4 (8.3) Range 59–100	78.4 (12.9) Range 36–99	67.9 (13.5) Range 26–97
Monitoring accuracy			
Absolute monitoring (bias) *	7.3 (12.1) [a] Range − 28-47	8.9 (14) [a] Range-27-38	12.3 (18) [b] Range − 28-58
Resolution - Relative monitoring (Gamma coefficient) **	0.45 (0.46) [a] Range − 1-1	0.44 (0.41) [a] Range − 1-1	0.01 (0.52) [b] Range − 1-1

*$p < .05$, ** $p < .01$. Means in the same row with different superscript letters are significantly different from each other

significant correlations between confidence ratings and bias scores. Resolution scores were not correlated across the three tasks (see Table 3).

Additional analysis of the relationship between confidence and performance accuracy in each task reveals moderate to high significant positive correlations in each task (L1, $r = .37$, $p < .01$; FL, $r = .53$, $p < .01$, Non-verbal, $r = .48$, $p < .01$), suggesting that students were generally calibrated - high performance accuracy was in accordance with high confidence and vice versa, especially in the FL and in the non-verbal task (Raven). Still, overconfidence was abundant in all three tasks, as evidenced by the fact that 70% of participants were overconfident` in L1, 73% in FL and 75% in the Non-verbal task (see Fig. 1).

2. Are confidence ratings and comprehension monitoring accuracy (calculated by resolution and bias) language specific (L1, FL), dependent on proficiency, supporting the skill-oriented approach?

Next, to explore *the skill-oriented approach* and examine whether the correspondence between monitoring accuracy in L1 and FL might be influenced by FL proficiency level, we conducted a moderated regression analysis to examine the possible moderating role of FL proficiency on the relationship between L1 and FL across three measures: 1. Confidence ratings; 2. Calibration bias; 3. Resolution. The moderation analysis was conducted using Hayes (2013) computational versatile tool of PROCESS MACRO version 3.4. We conducted three moderation models; **model 1** examined the relationship between confidence in L1 and confidence in FL with the moderation of FL proficiency; **model 2** examined the relationship between bias in L1 and bias in FL with the moderation of FL proficiency; **model 3** examined the relationship between resolution in L1 and resolution in FL with the moderation of FL proficiency.

Table 3 Pearson correlations across languages and tasks ($N = 127$)

Language/task	Confidence		Bias		Resolution	
	L1	FL	L1	FL	L1	FL
Non-verbal (Raven)	.44**	.40**	.31**	.34**	−.05	.08
Hebrew (L1)		.55**		.24**		.03

** $p < .01$

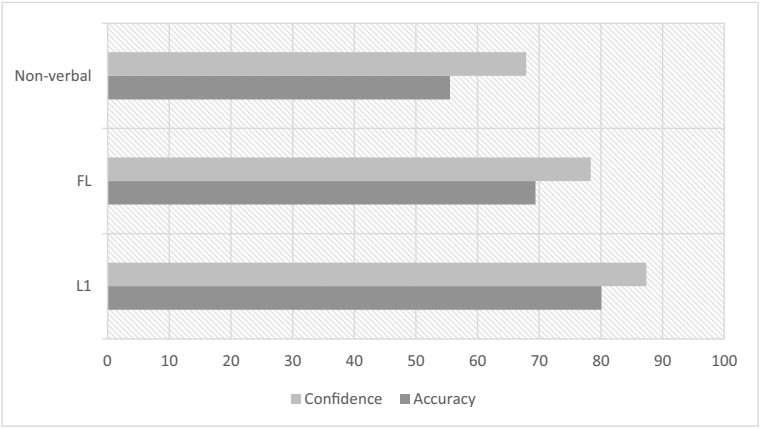

Fig. 1 Confidence and performance accuracy in the three tasks (L1 and FL reading comprehension and the non-verbal task) ($N = 127$)

Each model was conducted twice, once with the two independent variables (main model) and the second with the interaction between them (interaction model).

As described in the method section, FL Proficiency was assessed with five measures examining receptive vocabulary knowledge (2 tasks), reading fluency, self-reported proficiency and the English section of the Israeli psychometric exam. The raw scores in each test were converted into z scores and then, using principal components analysis, the scores of the five tasks were combined to create a single composite FL proficiency variable. The first principal component accounted for 51.1% of the variance and had similar weights for each of the five variables. All other variables in the model were also converted to Z scores for compatibility.

Model 1- confidence ratings

The first moderation model tested the relationship between L1 and FL confidence ratings with the moderating role of FL proficiency. Results indicated significant main effects for L1 confidence and FL proficiency, explaining together 50% of the variance in FL confidence, $R^2 = .504$, $F\ (2, 135) = 68.2$, $p < .001$. The interaction model predicted a significant portion of the variance, $R^2 = .502$, $F\ (3, 134) = 45.45$, $p < .001$, but this interaction effect did not significantly increase the variance explained by the main effects, $\Delta R^{101} = .003$, $F\ (1, 134) = .66$, $p = .417$, suggesting that each variable (L1 confidence and FL proficiency) significantly explained independent variance in FL confidence (see Table 4).

Table 4 Main and interaction effects of L1 confidence and FL proficiency on FL confidence ($N = 138$)

	Main model: Z- FL Confidence			Interaction model: Z- FL Confidence		
Predictors	*β*	*95%CI*	*p*	*β*	*95%CI*	*p*
(Intercept)	−0.00	−0.12 – 0.12	0.993	−0.01	−0.13 – 0.11	0.898
Z - L1 Confidence	0.47	0.35–0.59	<0.001	0.47	0.34–0.59	<0.001
Z - FL Proficiency	0.64	0.47–0.80	<0.001	0.65	0.47–0.82	<0.001
Z - L1 Confidence * Z - FL Proficiency				0.06	−0.09 – 0.21	0.417

Analysis of the slopes (beta) indicates a positive relationship between confidence in L1 and FL, namely, the higher the confidence is in L1, the higher the confidence is in FL. Additionally, FL proficiency was also positively associated with FL confidence - the higher the proficiency in the FL, the higher is the confidence in FL.

Model 2 - Bias

The second moderation model tested the relationship between L1 and FL absolute monitoring accuracy, calibration bias, with the moderating role of FL proficiency. Results indicated significant main effects for L1 bias and FL proficiency, explaining together 6.7% of the variance in FL bias, $R^2 = .067$, F (2, 135) = 4.81, $p < .001$. Inspection of the unique contribution of each independent variable indicates that only L1 bias positively and significantly predicted FL bias ($\beta = .23$, $p < .001$). FL proficiency did not add any unique contribution to explaining variance in FL bias ($\beta = -.11$, $p = .36$). Importantly, the model also identified a significant interaction effect, $R^2 = .093$, F (3, 134) = 4.56, $p < .001$. Specifically, FL proficiency did not influence FL bias by its own right, but FL proficiency did moderate the relationship between L1 and FL bias. The additive value of this interaction was marginally significant, $\Delta R^{101} = .03$, F (1, 134) = 3.87, $p = .051$; see Table 5.

Analysis of the simple slopes indicates that for individuals with low FL proficiency (−1 SD), L1 bias is unrelated to FL bias, $\beta = .08$, $p = .50$, 95%*CI* (−.15, .30). With medium levels of FL proficiency (set at the mean), L1 bias and FL bias are weakly associated, indicating that as L1 bias increases so does FL bias, $\beta = .24$, $p = .01$, 95%*CI* (.07, .41). Finally, at high levels of FL proficiency (+1 SD), the association between L1 bias and FL bias is of medium strength, again showing that as L1 bias increases so does FL bias, $\beta = .40$, $p < .001$, 95%*CI* (.16, .65); see Fig. 2.

Model 3 - resolution

The third moderation model tested the relationship between L1 and FL relative monitoring accuracy, resolution (gamma coefficient), with the moderating role of FL proficiency. Results indicated non-significant main effects, $R^2 = .002$, F (2, 135) = 0.13, $p = .776$, and a non-significant interaction effect, $R^2 = .016$, F (3, 134) = 0.71, $p = .549$. There was no significant association between L1 and FL resolution, and FL resolution was also unassociated with the moderator variable, FL proficiency, $\Delta R^{101} = .01$, F (1, 134) = 1.86, $p = .175$; see Table 6.

Table 5 Main and interaction effects of L1 bias and FL proficiency on FL bias ($N = 138$)

	Main model: Z- FL bias			Interaction model: Z- FL bias		
Predictors	β	*95%CI*	*p*	β	*95%CI*	*p*
(Intercept)	0.00	−0.16 – 0.16	0.999	0.04	−0.13 – 0.21	0.639
Z - L1 bias	0.23	0.06–0.40	0.009	0.24	0.07–0.41	0.001>
Z - FL Proficiency	−0.11	−0.34 – 0.13	0.361	−0.11	−0.34 – 0.12	0.347
Z - L1 bias * Z - FL Proficiency				0.23	−0.00 – 0.46	0.051

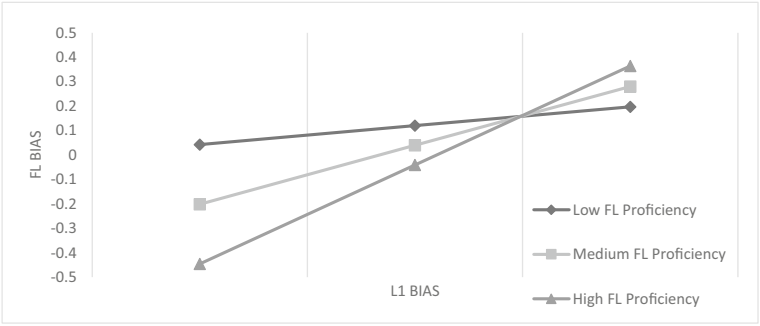

Fig. 2 Interaction effect of L1 bias and FL Proficiency on FL bias ($N = 138$)

Discussion

Self-monitoring is highly important for meaningful reading comprehension and therefore has been extensively studied over the years. Comprehension monitoring is especially important for reading comprehension in English as a foreign language, as practiced in higher education in our current global world. Hence, the main goal of the present study was to understand the underlying sources of confidence ratings, bias and resolution as indicators of comprehension monitoring in English as a foreign language. More specifically, we aimed to clarify whether these processes are trait-oriented or skill-oriented.

Our analysis demonstrated a mixed pattern of results, suggesting that comprehension monitoring involves both domain general and domain specific skills. On the one hand, confidence and absolute monitoring were associated across L1 and FL, as well as a non-verbal task. On the other hand, the strength of the association between absolute monitoring accuracy across the languages was modulated by FL proficiency level, and relative accuracy was not associated across languages and domains.

The fact that confidence in both reading comprehension tasks was correlated with the non-verbal task as well strengthens the trait-oriented assumption suggesting shared self-evaluation processes and supports the claim of a "confidence trait" (Kasperski and Katzir 2013; Moore et al. 2005; Pallier et al. 2002; Stankov et al. 2015). Further, correlation between confidence ratings in the L1 and the FL was evident at all FL proficiency levels, and the strength of this association was not moderated by FL proficiency. These results appeared despite the existing differences between comprehension accuracy in the L1 and the FL, and thus support the argument that confidence ratings are trait oriented and are influenced by the general subjective

Table 6 Main and interaction effects of L1 resolution and FL proficiency on FL resolution ($N = 138$)

	Main model: Z- FL resolution			Interaction model: Z- FL resolution		
Predictors	*β*	*95%CI*	*p*	*β*	*95%CI*	*p*
(Intercept)	0.00	−0.17 – 0.17	1.000	0.01	−0.16 – 0.18	0.931
Z - L1 resolution	0.02	−0.15 – 0.19	0.776	0.03	−0.14 – 0.19	0.770
Z - FL Proficiency	−0.05	−0.29 – 0.18	0.662	−0.05	−0.29 – 0.18	0.659
Z - L1 resolution * Z - FL Proficiency				−0.19	−0.47 – 0.09	0.175

feelings of competence, such as self-efficacy (Bandura 1982), and not exclusively by objective performance in the task.

However, inspecting both measures of comprehension monitoring accuracy reveals some variation across measures. Absolute accuracy (bias) was associated across languages and domains, whereas relative accuracy (gamma) showed no associations at all. Similar results were reported in earlier studies exploring both predictive monitoring accuracy (Glenberg and Epstein 1987; Kelemen et al. 2000; Leonesio and Nelson 1990), postdictive monitoring accuracy (Pressley and Ghatala's 1988; Silawi et al. 2020) and postdictive accuracy across a learning process (Mengelkamp and Bannert 2010). Specifically, Kelemen et al. (2000) did not detect any consistency in relative monitoring accuracy across different task-domains, different types of judgments and different time points (in contrast to the associations found in confidence and absolute accuracy). Thus, they concluded conclusively against a "general metacognitive ability".

These different patterns between relative and absolute measures of monitoring can be explained by the way each measure is calculated. Bias scores are derived from the gap between confidence ratings and performance accuracy, and thus are dependent on the variance of both confidence and the performance in the criterion test (Nelson 1996). For this reason, it is possible that the commonality of absolute accuracy simply reflects the consistency in the estimation of knowledge (confidence) or the knowledge/ability itself (performance accuracy), or both (Dunlosky et al. 2015; Mengelkamp and Bannert 2010). Thus, in the context of the current study, it can be argued that the commonality observed in bias scores merely reflects the stability of comprehension accuracy and confidence ratings rather than the relationship between monitoring accuracy across tasks.

In contrast, the measure of relative monitoring accuracy (resolution) is a more independent measure which is not reliant on the overall levels of performance accuracy or confidence ratings. Thus, it can be regarded as a more precise measure of metacognitive monitoring (Nelson 1996) which is based on participants' specific situational experiences (Koriat 2007) with a task and hence tends to be unstable (Leonesio and Nelson 1990). Additionally, our findings are aligned with the claim that relative and absolute accuracies capture different aspects of monitoring (e.g., Maki 1998).

In the metalinguistic domain, these findings align with recently reported findings of cross-linguistic associations in a population of trilingual university students (Silawi et al. 2020). In this study, using the same 'calibration of comprehension' paradigm, absolute monitoring was consistently associated across the three languages, whereas relative monitoring was not (also calculated with gamma coefficient). Additional support derives from a metalinguistic study which used a think aloud paradigm in first and second language reading comprehension indicating cross-linguistic associations (Jiménez et al.,1996), as well as from cross linguistic interventional studies of reading comprehension strategies (Abu-Rabia et al. 2013; Schwartz et al. 2013). This cumulative evidence lends support for the common underlying cognitive processes hypothesis (Geva and Ryan 1993; Chung et al. 2019), which suggests that metacognitive skills are shared across languages so that once gained in one language, they can be applied to other languages as well.

Bearing this evidence for generality in mind, the current results also provide some evidence for the skill-oriented approach. First, we found significant differences in both absolute and relative monitoring accuracies between the verbal and non-verbal domains. The bias scores in L1 and FL reading comprehension tasks were significantly smaller than bias in the non-verbal task. Similarly, resolution was also more accurate in the verbal tasks than in the non-verbal

task: In both reading comprehension tasks, mean gamma scores indicated quite high resolution (L1, $M = .45$, $sd = .46$; FL, $M = .44$, $sd = .41$), suggesting good distinction between correct and incorrect responses (Glenberg and Epstein 1987). In comparison, resolution in the non-verbal task was extremely low, close to zero ($M = .01$, $sd = .52$), signifying no discrimination between correct and incorrect items. These variations might be explained by task difficulty since it is clearly apparent that the non-verbal task (which contained 15 items from the most challenging sections of the Raven test), was the hardest task for most students. Yet, the fact that distinctions between verbal and non-verbal tasks were observed in both monitoring accuracy measures, and especially in resolution (which is a rather independent measure), reinforces the differentiation across domains. Mengelkamp and Bannert (2010) have suggested that relative monitoring accuracies are domain-specific, thus might generalize between tests within the same domain, but not between different domains. Although our results did not demonstrate generalization of resolution within the verbal domain, our descriptive data is somewhat supportive of this claim.

Second, the analysis of whether FL proficiency moderated the cross-language association in absolute monitoring accuracy, revealed an interesting pattern of results. Thus, although bias scores at the group level were similar across the two languages, at the individual level they were correlated across languages only for participants who had medium to high proficiency levels in FL. This pattern suggests that a certain proficiency threshold in the FL might be necessary in order to successfully share accurate comprehension monitoring across the languages. These findings are consistent with Maki et al.'s (2005) results showing that verbal ability accounted for confidence ratings and predictive absolute monitoring accuracy. It also corresponds with various metalinguistic studies reporting that comprehension monitoring (tested in different paradigms such as "think aloud", error detection and questionnaires) was modulated by a foreign language proficiency threshold (Block 1992; Han 2012; Han and Stevenson 2008; Khonamri and Mahmoudi Kojidi 2011; Lin and Yu 2015). Finally, this finding also aligns with the report that in trilinguals, monitoring seems to be utilized similarly by individuals to support comprehension across the first and second languages, but is less well generalized to the third language (the least proficient language; Silawi et al. 2020). The linguistic interdependence theory (Cummins 1979, 2012), which postulates that FL reading proficiency level might influence comprehension monitoring processes, is in line with these findings.

In sum, our results seem to support both trait and skill as the sources of confidence ratings and comprehension monitoring accuracies in L1 and FL reading comprehension. Post-diction confidence ratings, as subjective indicators of monitoring (Ackerman and Leiser 2014), seem to be mainly domain-general, resting upon theory-based cues such as self-efficacy perceptions (Bandura 1982), which go beyond the properties of the specific testing tool (Pallier et al. 2002). Absolute monitoring accuracy, as derived from the mean difference between the components of performance accuracy and confidence, seems to reflect both theory/self-belief cues (of each student of herself as a learner, a reader etc.) and experience/situational cues (Koriat 2007) such as the actual experience of effort or difficulty in reading and answering test items. This finding is aligned with Kasperski and Katzir (2013) which indicated that the over- and under-confidence in children were dependent on reading ability as well as on personality traits and reading self-perceptions. This finding is also consistent with Dinsmore and Parkinson's (2013) study which investigated students' explanations of their confidence ratings in reading comprehension tasks, suggesting that confidence ratings are constructed on the combination of both personal factors such as prior knowledge (i.e. trait-based) and task factors such as text characteristics and difficulty (i.e. skill-based).

With regard to the relative monitoring accuracy, it seems that it acts as a more domain-specific measure by its own nature (Mengelkamp and Bannert 2010; Nelson 1996) and is thus more influenced by experience-based cues such as skill and a the specific characterization of a given test (Koriat 2007).

There are several limitations to the current study. First, previous research has noted that monitoring accuracy is related to task difficulty (Dunlosky et al. 2015; Maki and Berry 1984; Pressley and Ghatala 1988). In the current study, there were differences in difficulty between the three tasks (despite piloting text comprehension in each language), which are probably inherent when testing comprehension in a mother tongue versus a foreign language (Ercikan and Por 2020). Future studies should take this confound of underlying comprehension accuracy into consideration. Second, the fact that texts within each language varied between participants, in difficulty and content, due to the constraints of the training design, probably added extra "noise" to the data and might have influenced results. Third, as part of our data analysis was based on correlations, other variables could have affected individual differences such as intelligence, experience, prior knowledge and motivation. All our participants were undergraduate students, thus it was assumed that they share some background knowledge and day to day learning experiences. Future studies may try to control some of these variables by testing comprehension monitoring outside the lab, in actual university classroom settings, thus applying results to "real life" learning situations within higher education systems.

To conclude, the results of the current research add to what is known about the nature of comprehension monitoring across languages, theoretically and practically. Theoretically, the results suggest that comprehension monitoring is both trait and skill oriented. It seems that monitoring is shared across languages, as long as a sufficient proficiency threshold in the language has been reached. Thus, practically, when building new instructional methods for effective FL reading comprehension in higher education, both dimensions of domain-general and language-specific skills should be considered. In addition to direct instruction of FL learning and meaning construction, specific training of self-monitoring techniques while reading should be implemented.

Acknowledgements This research was supported by grant 1094/14 from the Israeli Science Foundation to AP and TK and by the Edmond J. Safra Brain Research Center for the study of Learning Disabilities. The authors wish to thank Dr. Nachshon Korem for programming assistance, Razan Silawi and Gali Yosephi for diligent assistance in data collection and coding, and Sandra Zuckerman for statistics consulting.

Compliance with ethical standards

Conflict of interest The authors hereby declare that they have no conflicts of interest.

Informed consent The study was approved by the IRB of the University of Haifa, and all participants gave full informed consent, and were compensated for their participation in the study.

References

Abu-Rabia, S., & Siegel, L. S. (2003). Reading skills in three orthographies: The case of trilingual Arabic-Hebrew-English-speaking Arab children. *Reading and Writing: An Interdisciplinary Journal, 16*(7), 611–634. https://doi.org/10.1023/A:1025838029204.

Abu-Rabia, S., Shakkour, W., & Siegel, L. (2013). Cognitive retroactive transfer (CRT) of language skills among bilingual Arabic-English readers. *Bilingual Research Journal, 36*(1), 61–81. https://doi.org/10.1080/15235882.2013.775975.

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied, 17*(1), 18–32. https://doi.org/10.1037/a0022086.

Ackerman, R., & Leiser, D. (2014). The effect of concrete supplements on metacognitive regulation during learning and open-book test taking. *British Journal of Educational Psychology, 84*(2), 329–348.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*(2), 122–147. https://doi.org/10.1037/0003-066X.37.2.122.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194.

Bernhardt, E. B., & Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interderpendence hypotheses. *Applied Linguistics, 16*, 15–34.

Beznos, M. & Prior, A. (2009). Hebrew version of the Language Experience and Proficiency Questionnaire, https://www.iris-database.org/iris/app/home/detail?id=york:822288.

Block, E. L. (1992). See how they read: Comprehension monitoring of L1 and L2 readers. *TESOL Quarterly, 26*(2), 319. https://doi.org/10.2307/3587008.

Brown, J. I., Fishco, V. V., & Hanna, G. S. (1993). *Nelson–Denny Reading test*. Rolling Meadows: Riverside Publishing.

Buratti, S., Allwood, C. M., & Kleitman, S. (2013). First- and second-order metacognitive judgments of semantic memory reports: The influence of personality traits and cognitive styles. *Metacognition and Learning, 8*(1), 79–102. https://doi.org/10.1007/s11409-013-9096-5.

Chiappe, P., Siegel, L. S., & Gottardo, A. (2002). Reading-related skills of kindergartners from diverse linguistic backgrounds. *Applied PsychoLinguistics, 23*, 95–116.

Chung, S. C., Chen, X., & Geva, E. (2019). Deconstructing and reconstructing cross-language transfer in bilingual reading development: An interactive framework. *Journal of Neurolinguistics, 50*, 149–161. https://doi.org/10.1016/j.jneuroling.2018.01.003.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*, 222–25l.

Cummins, J. (2012). The intersection of cognitive and sociocultural factors in the development of reading comprehension among immigrant students. *Reading and Writing, 25*(8), 1973–1990. https://doi.org/10.1007/s11145-010-9290-7.

Dabarera, C., Renandya, W. A., & Zhang, L. J. (2014). The impact of metacognitive scaffolding and monitoring on reading comprehension. *System, 42*, 462–473. https://doi.org/10.1016/j.system.2013.12.020.

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence ratings made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001.

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*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003.

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2015). *Methodology for Investigating Human Metamemory* (J. Dunlosky & S. (Uma) K. Tauber, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.14.

Ehrlich, M. F., Remond, M., & Tardieu, H. (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing: An Interdisciplinary Journal, 11*, 29–63.

ErçEtin, G., & Alptekin, C. (2013). The explicit/implicit knowledge distinction and working memory: Implications for second-language reading comprehension. *Applied PsychoLinguistics, 34*(04), 727–753.

Ercikan, K., & Por, H. H. (2020). Comparability in multilingual and multicultural assessment contexts. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). Washington, DC: National Academy of Education.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906.

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing, 25*(8), 1819–1845.

Geva, E., & Ryan, E. B. (1993). Linguistic and cognitive correlates of academic skills in first and second language. *Language Learning, 43*, 5–42.

Gilboa, A. (Unpublished). Hebrew version of the Shipley Vocabulary Scale.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 11*(4), 702–718.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*(1), 84–93. https://doi.org/10.3758/BF03197714.

Goodman, J. T., Streiner, D. L., & Woodward, C. A. (1974). Test-retest reliability of the Shipley-Institute of Living Scale: Practice effects or random variation. *Psychological Reports, 35*(1), 351–354.

Grabe, W. (2014). Key issues in L2 reading development. *Centre for English Language Communication*, 8–18.

Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (2nd ed.). Harlow: Longman/Pearson.

Grabe, W., & Zhang, C. (2013). Reading and writing together: A critical component of English for academic purposes teaching and learning. *TESOL Journal, 4*(1), 9–24. https://doi.org/10.1002/tesj.65.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*(1), 93–103.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*(7), 1001–1013. https://doi.org/10.3758/MC.37.7.1001.

Han, F. (2012). Comprehension monitoring in Reading English as a foreign language. *New Zealand Studies in Applied Linguistics, 18*(1), 36–49.

Han, F., & Stevenson, M. A. R. I. E. (2008). Comprehension monitoring in first and foreign language reading. *University of Sydney Papers in TESOL, 3*, 73–110. https://doi.org/10.3138/cmlr.61.1.77.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Hudson, T. (2007). *Teaching second language reading*. Oxford: Oxford University Press.

Jiménez, R. T., García, G. E., & Pearson, P. D. (1996). The reading strategies of bilingual Latina/o students who are successful English readers: Opportunities and obstacles. *Reading Research Quarterly, 31*(1), 90–112. https://doi.org/10.1598/RRQ.31.1.5.

Kahn-Horwitz, J., Shimron, J., & Sparks, R. L. (2005). Predicting foreign language Reading achievement in elementary school students. *Reading and Writing, 18*(6), 527–558. https://doi.org/10.1007/s11145-005-3179-x.

Kasperski, R., & Katzir, T. (2013). Are confidence ratings test- or trait-driven? Individual differences among high, average, and low comprehenders in fourth grade. *Reading Psychology, 34*(1), 59–84.

Katzir, T., Schiff, R., & Kim, Y.-S. (2012). The effects of orthographic consistency on reading development: A within and between cross-linguistic study of fluency and accuracy among fourth grade English- and Hebrew-speaking children. *Learning and Individual Differences, 22*(6), 673–679.

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28*(1), 92–107. https://doi.org/10.3758/BF03211579.

Khonamri, F., & Mahmoudi Kojidi, E. (2011). Metacognitive awareness and comprehension monitoring in Reading ability of Iranian EFL learners. *PROFILE: Issues in Teachers' Professional Development, 13*(2), 99–111.

Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessment. In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 21–38). Plymouth: Rowman & Littlefield.

Kleider-Tesler, E., Prior, A., & Katzir, T. (2019). The role of calibration of comprehension in adolescence: From theory to online training. *Journal of Cognitive Education and Psychology, 18*(2), 190–211.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*, 161–173. https://doi.org/10.1016/j.lindif.2007.03.004.

Kolić-Vehovec, S., & Bajšanski, I. (2007). Comprehension monitoring and reading comprehension in bilingual students. *Journal of Research in Reading, 30*(2), 198–211. https://doi.org/10.1111/j.1467-9817.2006.00319.x.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289–326). New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511816789.012.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*(2), 147–162. https://doi.org/10.1037/0096-3445.131.2.147.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and Reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30.

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 464–470. https://doi.org/10.1037/0278-7393.16.3.464.

Lervåg, A., & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and foreign language learners: Growth in L1 and FL reading comprehension. *Journal of Child Psychology and Psychiatry, 51*(5), 612–620.

Lesaux, N. K., & Siegel, L. S. (2003). The development of reading in children who speak English as a foreign language. *Developmental Psychology, 39*(6), 1005–1019.

Lesaux, N., Koda, K., Siegel, L., & Shanahan, T. (2006). Development of literacy. Developing literacy in second-language learners. *Report of the National Literacy Panel on Language-Minority Children and Youth*. Mahwah: Lawrence Erlbaum Associates, Publishers.

Lin, L. C., & Yu, W. Y. (2015). A think-aloud study of strategy use by EFL college readers reading Chinese and English texts. *Journal of Research in Reading, 38*(3), 286–306.

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased Processing Enhances Calibration of Comprehension. *Journal of experimental Psychology: Learning, Memory and Cognition, 16*(4), 609–616.

Maki, R. H. (1998). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition, 26*(5), 959–964. https://doi.org/10.3758/BF03201176.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 10*(4), 663–679.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*(4), 723–731.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and Multilinguals. *Journal of Speech Language and Hearing Research, 50*(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067).

Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence ratings: Stability and generality in the learning process and predictive validity for learning outcome. *Memory & Cognition, 38*(4), 441–451. https://doi.org/10.3758/MC.38.4.441.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502.

Moore, D., Lin-Agler, L. M., & Zabrucky, K. M. (2005). A source of Metacomprehension inaccuracy. *Reading Psychology, 26*(3), 251–265. https://doi.org/10.1080/02702710590962578.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13 Retrieved from http://jalt-publications.org/tlt.

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance of an individual item comments on Schraw. *Applied Cognitive Psychology, 10*(3), 257–260.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence ratings. *The Journal of General Psychology, 129*(3), 257–299. https://doi.org/10.1080/00221300209602099.

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

Pressley, M., & Ghatala, E. S. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly, 23*(4), 454. https://doi.org/10.2307/747643.

Prior, A. (2012). Too much of a good thing: Stronger bilingual inhibition leads to larger lag-2 task repetition costs. *Cognition, 125*, 1–12.

Prior, A., Zeltsman-Kulick, R. & Katzir, T. (2020). Adolescent word reading in English as a foreign language. *Journal of Research in Reading, 43*(1), 116–139.

Raven, J. C. (1960). *Guide to the standard progressive matrices*. London: H.K. Lewis & Co. Ltd..

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*(6), 1004–1010. https://doi.org/10.3758/BF03209348.

Sarac, S., & Tarhan, B. (2009). Calibration of comprehension and performance in L2 reading. *International Electronic Journal of Elementary Education, 2*(1), 167–179.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-prime User's guide*. Pittsburgh: Psychology Software Tools, Inc..

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*, 33–45. https://doi.org/10.1007/s11409-008-9031-3.

Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory & Cognition, 22*(1), 63–69. https://doi.org/10.3758/BF03202762.

Schwartz, A. I., Mendoza, L., & Meyer, B. (2013). The impact of text structure reading strategy instruction in a foreign language: Benefits across languages. *The Language Learning Journal*, 1–19.

Sheorey, R., & Mokhtari, K. (2001). Differences in the metacognitive awareness of reading strategies among native and non-native readers. *System, 29*(4), 431–449. https://doi.org/10.1016/S0346-251X(01)00039-2.

Shipley, W. C. (1940). A self administering-scale for measuring intellectual impairment and deterioration. *The Journal of Psychology, 9*, 371–377.

Silawi, R., Shalhoub-Awwad, Y. & Prior, A. (2020). Comprehension monitoring in L1, L2 and L3: Domain general or language specific? *Language Learning, 70*(3), 886–922. https://doi.org/10.1111/lang.12410.

Snow, C. (2002). Reading for understanding: Toward an R&D program in reading comprehension. Rand Corporation.

Stankov, L. (1999). Mining on the "no man's land" between intelligence and personality. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 315–337). Washington, DC: American Psychological Association.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*(6), 747–758. https://doi.org/10.1016/j.lindif.2012.05.013.

Stankov, L., Kleitman, S., & Jackson, S. A. (2015). Measures of the trait of confidence. In *Measures of Personality and Social Psychological Constructs* (pp. 158–189). https://doi.org/10.1016/B978-0-12-386915-9.00007-3.

Stevenson, M., Schoonen, R., & de Glopper, K. (2007). Inhibition or compensation? A multidimensional comparison of Reading processes in Dutch and English: Inhibition or compensation? *Language Learning, 57*, 115–154. https://doi.org/10.1111/j.1467-9922.2007.00414.x.

Taki, S. (2016). Metacognitive online reading strategy use: Readers' perceptions in L1 and L2. *Journal of Research in Reading, 39*(4), 409–427. https://doi.org/10.1111/1467-9817.12048.

Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*(2), 129–160.

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. https://doi.org/10.1037/0022-0663.95.1.66.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology, 91*, 579–594.

Trapman, M., van Gelderen, A., van Schooten, E., & Hulstijn, J. (2017). Reading comprehension level and development in native and language minority adolescent low achievers: Roles of linguistic and metacognitive knowledge and fluency. *Reading & Writing Quarterly, 33*(3), 239–257. https://doi.org/10.1080/10573569.2016.1183541.

Tsai, Y.-R., Ernst, C., & Talley, P. C. (2010). L1 and L2 strategy use in reading comprehension of Chinese EFL readers. *Reading Psychology, 31*, 1–29. https://doi.org/10.1080/02702710802412081.

Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology, 96*(1), 19–30. https://doi.org/10.1037/0022-0663.96.1.19.

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

Wenden, A. L. (1999). An introduction to metacognitive knowledge and beliefs in language learning: Beyond the basics. *System, 27*(4), 435–441. https://doi.org/10.1016/S0346-251X(99)00043-3.

Zachary, A. (1986). *Shipley Institute of Living Scale. Revised manual*. Los Angeles: Western Psychological Services.

Zachary, R. A. (1991) *The manual of the Shipley Institute of Living Scale.* Los Angeles, CA: Western Psychological Services.

Zhang, D., & Zhang, L. J. (2019). Metacognition and self-regulated learning (SRL) in second/foreign language teaching. In X. Gao (Ed.), *Second Handbook of English Language Teaching (pp. 1–15)*. Springer International Publishing. https://doi.org/10.1007/978-3-319-58542-0_47-1.

Author notes:

1. We confirm that for the experiment described here we have reported all measures, conditions and data exclusion decisions. In addition to the measures reported in the current manuscript, we also administered two WMC tasks (operation-span and Letter-Number sequencing) which will be analyzed in a future manuscript. Reading times and response times were also recorded in the reading comprehension task, in order to ascertain that participants were indeed engaged in reading the texts but were not further analyzed. In addition, the participants in the current experiment were also followed longitudinally in a training and feedback paradigm for 4 additional sessions, and a final follow up session. These longitudinal data will also be reported in a separate manuscript. Sample size was based on estimated power analyses of training and feedback effects, and on feasibility of conducting a longitudinal intervention study.

Affiliations

Lilach Temelman-Yogev[1] · Tami Katzir[1] · Anat Prior[1]

Lilach Temelman-Yogev
lilachtem@gmail.com

Tami Katzir
tkatzir@edu.haifa.ac.il

[1] Edmond J Safra Brain Research Center for the Study of Learning and Learning Disabilities, Department of Learning Disabilities, University of Haifa, Haifa, Israel