



EMPIRICAL STUDY

Monitoring of Reading Comprehension Across the First, Second, and Third Language: Domain-General or Language-Specific?

Razan Silawi ^a Yasmin Shalhoub-Awwad,^{a,b}
and Anat Prior ^{a,b}

^aDepartment of Learning Disabilities, University of Haifa and ^bEdmond J Safra Brain Research Center for the Study of Learning Disabilities, University of Haifa

This study examined the monitoring abilities of trilinguals in reading comprehension, addressing the following questions: Is comprehension monitoring related to reading comprehension across first, second, and third languages? Is comprehension monitoring shared across the languages of trilingual adults (domain-general) or rather linked to language proficiency (language-specific)? Eighty undergraduates, trilingual in Arabic, Hebrew, and English, read three texts in each language, answered multiple-choice questions, and rated their confidence in their responses. From this we derived the absolute and relative accuracy of monitoring. The results showed links between accurate monitoring and successful comprehension in all the languages of participants, but these were weaker in English, the least proficient language. Further, the results lend some support to the involvement of both domain-general and language-specific processes in

This study was supported by grant number 1094/14 from the Israeli Science Foundation to AP, by a research grant from the Arabic Language Academy to AP and YSA, and by the Edmond J. Safra Brain Research Center for the Study of Learning Disabilities. RS was supported by the Otto Werner Excellence Scholarship from The Jewish-Arab Center at the University of Haifa. The authors wish to thank Dr Nachshon Korem for assistance in data analyses, and five anonymous reviewers for comments on a previous version of the manuscript.

Correspondence concerning this article should be addressed to Anat Prior, Edmond J. Safra Brain Research Center for the Study of Learning Disabilities, University of Haifa, 199 Abba Hushi Ave., Mount Carmel, Haifa, Israel, 3498838. E-mail: aprior@edu.haifa.ac.il

[Corrections made on 25 May 2020, after first online publication: The author byline and affiliations have been corrected in this version.]

comprehension monitoring. Specifically, 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.

Keywords comprehension monitoring; reading comprehension; calibration; bilingual; multilingual; foreign language; domain-general; second language; third language

Introduction

Students in postsecondary education spend a significant portion of their study efforts reading course materials. Thus, reading comprehension is a crucial factor for achievement and success. Reading comprehension is a complex skill that relies on both lower level lexical skills, such as vocabulary knowledge and word-reading efficiency, and higher level text processing and metacognitive abilities, such as inference and monitoring of comprehension (e.g., the RAND model, developed by the RAND Reading Study Group [2002] defining reading comprehension as a multi-componential process engaging different elements, inter alia, the readers' cognitive abilities: Kintsch, 2002; Perfetti, Marron, & Foltz, 1996; Snow, 2002).

Increasingly, higher education students worldwide are reading materials in a second language (L2) or even a third language (L3). In light of the demonstrated importance of metacognition in facilitating comprehension in the first language (L1; e.g., Dunlosky & Rawson, 2012), the current study addressed the question of whether similar patterns can be identified in reading comprehension in the L2 and L3. Further, we asked whether metacognitive skills in reading comprehension are a core skill, which can be used similarly by individual readers across the different languages they read, or might be better conceived of as language-specific. To this end, we measured trilingual university students' comprehension monitoring in the three languages they read: Arabic, Hebrew, and English.

Metacognition is comprised of two main components: evaluation of cognition or *monitoring*, and regulation of cognition or *control* (Baker & Brown, 1984; Dunlosky & Thiede, 2013; Flavell, 1979). Monitoring is one's ability to actively evaluate ongoing cognitive processes, whereas control refers to the use of appropriate and effective strategies to regulate cognitive processes (Lin & Zabrucky, 1998). In the context of reading comprehension, an example of monitoring is when the reader recognizes comprehension difficulties, and an example of control is when the reader chooses to reread the paragraph or text (Lin & Zabrucky, 1998). Hence, the decision to regulate the learning process is based mainly on ongoing evaluation and monitoring (Ackerman & Goldsmith,

2008). One measure for evaluating monitoring of comprehension, which is the focus of the current study, is calibration of comprehension. Thus, in the current study, we asked readers to directly estimate their comprehension, and we then compared these judgments with their actual comprehension performance.

Background Literature

Metacognition in L1 Reading Comprehension

Reading comprehension, or the ability of a reader to extract meaning from written text, is a complex undertaking. The “simple view of reading” (Hoover & Gough, 1990) suggested that reading comprehension is the product of decoding ability and language comprehension ($R = D \times C$). Indeed, in young readers, the lower level skills of decoding and reading ability are highly predictive of reading comprehension (García & Cain, 2014; Kendeou, Van den Broek, White, & Lynch, 2009; Perfetti & Hart, 2001; Vellutino, Tunmer, Jaccard, & Chen, 2007). However, the contribution of lower level skills to the comprehension of adults and skilled readers is reduced (e.g., Landi, 2010; Tilstra, McMaster, Van den Broek, Kendeou, & Rapp, 2009).

More complex, componential models describe reading comprehension as recruiting additional skills, including cognitive and metacognitive abilities (as in the RAND model: Johnston & Kirby, 2006; Snow, 2002). The construction-integration model (Kintsch, 1988) and the landscape model (Van den Broek, Young, Tzeng, & Linderholm, 1999; Yeari & Van den Broek, 2011) both emphasize that readers engage in constructive processes to fully comprehend texts. Such effortful meaning making is often triggered by metacognitive monitoring processes, when a problem arises in the reader’s unfolding mental model of the text.

The involvement of metacognitive processes in reading comprehension has been supported in numerous studies. Metacognitive processes have been found to distinguish skilled from less skilled comprehenders, and can enhance students’ comprehension (Dunlosky & Rawson, 2012; Kasperski & Katzir, 2013; Langer, Bartolome, Vasquez, & Lucas, 1990; Zabrocky, Agler, & Moore, 2009). Comprehension strategies in general, and monitoring in particular, are associated with an active process in which the reader attempts to construct a coherent representation (Perfetti et al., 1996).

There is clear evidence of differences in comprehension monitoring between poor and good comprehenders (Dunlosky & Lipko, 2007; Ehrlich, Remond, & Tardieu, 1999; Glover, 1989; Lin, Moore, & Zabrocky, 2001; Maki, Shields, Wheeler, & Zacchilli, 2005; Zabrocky et al., 2009). Garner (1980), for instance, found that good adolescent comprehenders noticed an

inconsistency in a passage whereas poor comprehenders did not. In other words, poor comprehenders could not monitor their comprehension of the passages they read. Nevertheless, Garner speculated that poor monitoring could be either a cause or a result of poor comprehension.

Along similar lines, Maki, Jonas, and Kallod (1994) reported that better and faster comprehenders assessed their comprehension performance more accurately than did poorer comprehenders. Specifically, readers who are less accurate in responding to comprehension questions tend to provide higher confidence ratings, leading to an overestimation of their comprehension abilities. Such patterns of performance have been documented in both children (e.g., Ehrlich et al., 1999) and adults (Glover, 1989; Klassen, 2007; Maki et al., 2005). Several studies have specifically investigated college students and demonstrated that they are relatively poor at accurately judging their comprehension (Baker, 1989; Dunlosky & Lipko, 2007; Lin et al., 2001; Thiede, Griffin, Wiley, & Redford, 2009).

Metacognition in L2 Reading Comprehension

Most of the literature on reading comprehension in a L2 has investigated the underlying skills that predict comprehension, focusing mainly on linguistic and/or cognitive abilities such as word reading, vocabulary, and working memory (e.g., Lesaux, Koda, Siegel, & Shanahan, 2006). Thus, although models of native language reading comprehension (e.g., Van den Broek et al., 1999) should apply to non-native languages as well, few studies have directly investigated the importance of metacognition and comprehension monitoring in reading comprehension in a L2 or L3.

These studies seem generally to support a positive contribution of metacognition to reading comprehension in the L2 (mostly English). For example, two studies have reported that metacognitive knowledge, measured by a questionnaire, made a significant contribution to reading comprehension for L1-Dutch adolescents reading in English as a L2 (Van Gelderen et al., 2004) or a L3 (Van Gelderen et al., 2003). In addition, Trapman, van Gelderen, van Schooten, and Hulstijn (2017) reported that metacognitive knowledge was a significant predictor of reading comprehension in young low-achieving adolescents reading Dutch as a L1 or as a L2. Studies of university students have also reported a significant positive relationship between reading comprehension and self-reported metacognitive strategy use in both the L1 and L2 (Sheorey & Mokhtari, 2001; Taki, 2016).

Thus, these studies suggest that metacognition can contribute to reading comprehension in both the L1 and L2, but the specific conditions for this and

the nature of its contribution are still under debate. Importantly, hardly anything is known about the contribution of metacognition to L3 reading comprehension. Furthermore, most previous research focused mainly on offline self-reports of strategy knowledge and examined use of general metacognitive skills and did not specifically address the issue of comprehension monitoring. Thus, the first goal of the current study was to directly examine comprehension monitoring across the L1, L2, and L3 of trilingual adult university students, enabling us to investigate whether previously identified links between monitoring and comprehension are also attested across the languages of this understudied population when using online measures of monitoring.

Comprehension Monitoring: Shared or Language-Specific?

The studies reviewed so far do not speak directly to the question of possible transfer and sharing of comprehension monitoring skills and strategies across the languages of multilinguals, because in most cases different participants performed in the L1 and in the L2.

However, crosslinguistic transfer and sharing of metalinguistic skills in the two languages of bilinguals is attested in various domains. Thus, a meta-analysis showed that phonological awareness is correlated across the L1 and L2 (Melby-Lervåg & Lervåg, 2011; see also Saiegh-Haddad & Geva, 2008; Verhoeven, 2007). Other studies have found evidence for crosslinguistic transfer of morphological awareness (e.g., Deacon, Wade-Woolley, & Kirby, 2007; Pasquarella, Chen, Lam, Luo, & Ramirez, 2011; Ramirez, Chen, Geva, & Kiefer, 2010). Durgunoğlu (2002) summarized a large body of research showing cross-language associations in a variety of additional metalinguistic skills, including functional awareness, decontextualized language use, and meaning-making strategies (see also Chung, Chen, & Geva, 2019). To the extent that metacognitive abilities, and specifically comprehension monitoring, can be conceptualized as similar to such metalinguistic skills, it would stand to reason that multilinguals would be able to utilize such abilities across the languages they use.

According to this view, metacognitive skills are seen as a general ability, even though they may initially develop through literacy acquisition and practice mostly in the L1. Thus, we would expect individuals to benefit from applying metacognitive skills to reading comprehension to a similar degree in the different languages they read. Several studies support this possibility. For example, Jiménez, García, and Pearson (1996) showed evidence of transfer of some metacognitive strategies (evaluating, monitoring, rereading, and questioning) from one language to another in Spanish–English bilingual sixth- and

seventh-grade students. Similarly, Langer et al. (1990) examined fifth-grade Spanish–English bilinguals and found that some students used good meaning-making strategies in both English and Spanish, regardless of their different proficiency in the two languages. In addition, students who showed poor use of meaning-making strategies in their less proficient language showed limited use in their proficient language as well. A study with Chinese learners of English as a foreign language also reported a moderate positive relationship between L1 and foreign language comprehension monitoring (Han, 2013). Lin and Yu (2015), used a think-aloud methodology, and found that the use of metacognitive strategies was shared between the two languages of the participants (English L2 and Chinese L1), thus supporting the domain general view of shared high-level cognitive processes.

However, there are theoretical and empirical reasons to hypothesize that metacognitive abilities might not apply to the same degree to all the languages of multilinguals. Specifically, comprehension monitoring is considered a higher order skill within reading comprehension, and according to models of reading comprehension (Grabe & Stoller, 2013; Perfetti, Landi, & Oakhill, 2005), higher order skills can only be brought to bear on comprehension when lower level lexical skills (decoding and vocabulary knowledge) have reached a threshold of efficiency. Accordingly, the ability of readers to make use of metacognitive skills and monitoring in reading comprehension might be reduced in their less proficient languages (see also Cummins, 1976).

This theoretical explanation is supported by several empirical studies. Tsai, Ernst, and Talley (2010), in their study of Chinese undergraduates learning English as a foreign language, found evidence for cross-language transfer of reading strategies among highly proficient readers, who appeared to use similar reading strategies in the L1 and L2. However, less skilled readers used different strategies in L1 and L2 reading, suggesting that the use of these skills might be dependent on language proficiency. Along similar lines, Han and Stevenson (2008) reported that Chinese university students learning English as a foreign language performed significantly better in comprehension monitoring in L1 reading than in L2 reading.

Finally, one study also showed differences in metacognitive performance across the languages of bilinguals, but in the opposite direction, that is, more accurate judgments of their own comprehension in the L2 than in the L1. Sarac and Tarhan (2009) examined comprehension monitoring in the L2 (English) of Turkish undergraduates and found that students were more accurate in evaluating their performance on a comprehension test in the L2 compared to one in the L1. The researchers explained this counterintuitive finding by claiming

that a L2 is usually learned with conscious effort, and that readers are therefore more aware of the processes in their L2 than in their L1 and are able to evaluate their performance more accurately.

In the current study, we therefore measured trilingual participants' comprehension monitoring across the languages they used, allowing us to directly test whether such monitoring is shared or language-dependent. In addition, the study also included a nonlinguistic task, and we measured participants' monitoring of performance on this task. This allowed us to examine whether individual differences in monitoring were correlated across the nonlinguistic task and the reading comprehension tasks. Such a finding would support the claim that the metacognitive skill of performance monitoring generalizes across language and nonlinguistic performance, suggesting that it is of a wide domain-general nature.

Measuring Comprehension Monitoring

Monitoring of reading comprehension has been investigated in different manners across studies. Some have investigated it indirectly by presenting inconsistent information in the text (e.g., Block, 1992; Kroll & Ford, 1992; Zabrocky & Commander, 1993), and others directly by asking individuals to evaluate their comprehension after reading a text (Kasperski & Katzir, 2013; Lin et al., 2001; Lin & Zabrocky, 1998; Maki et al., 1994; Sarac & Tarhan, 2009).

The term *calibration* is widely used in the field of self-regulated learning (Ackerman & Goldsmith, 2008, 2011; Thiede, Anderson, & Theriault, 2003). Pieschl (2009) defined calibration as the accuracy of learners' perceptions of their own performance, that is, the ability to accurately evaluate comprehension (see also Alexander, 2013; Glenberg, Sanocki, Epstein, & Morris, 1987). In order to calculate calibration accuracy, researchers ask participants to provide confidence judgments, in which they subjectively rate their confidence regarding their comprehension (Ackerman & Goldsmith, 2011; Kasperski & Katzir, 2013; Lin & Yu, 2015; Pressley & Ghatala, 1988; Zabrocky et al., 2009). Calibration accuracy is then calculated in terms of the discrepancy between actual performance and confidence judgments, which is termed *calibration bias*.

There is much diversity in the literature concerning calibration as a measure of monitoring in the domain of reading comprehension, which can be summed up in three orthogonal dimensions:

- the variable being monitored: text comprehension, test performance, text features (i.e., text difficulty), or others;

- timing of confidence rating: prediction or postdiction (Lin et al., 2001; Lin & Zabrocky, 1998; Pieschl, 2009; also called posttest: see Maki, 1998); and
- the statistical measure used to quantify monitoring.

To illustrate such variability, Zabrocky et al. (2009), for example, focused on text comprehension and test performance as the monitored variables. Thus, they asked students to evaluate their comprehension of the text after reading it, and then to estimate their performance on a comprehension test. In contrast, Lin et al. (2001) asked participants to provide their subjective judgment of text ease after reading the text, and then compared these judgments with comprehension test performance.

Confidence judgment timing is also important. Thus, in prediction judgments participants are required to estimate expected test performance prior to exposure to the test, whereas in postdiction judgments participants are required to evaluate their performance on a comprehension test after taking it (Ackerman & Goldsmith, 2011; Lin et al., 2001; Sarac & Tarhan, 2009; Thiede et al., 2003). In general, postdictions are more accurate than predictions (Pieschl, 2009), due to additional information that is available to the participants about the nature of the test.

Finally, different statistical measures have been used across studies to quantify monitoring. Dunlosky and Thiede (2013) distinguished between measures of absolute accuracy and measures of relative accuracy. Calibration bias is an absolute measure of monitoring accuracy. Theoretically, it addresses the discrepancy between the perception of comprehension or performance and the actual accuracy on the test. Positive calibration bias reflects overconfidence and negative calibration bias reflects underconfidence (Ackerman & Goldsmith, 2011). In the current study, we were less interested in the distinction between overconfidence and underconfidence, and chose to focus on the magnitude of the discrepancy between confidence and performance. Therefore, we used the absolute value of the difference scores. *Resolution* is a relative measure of monitoring accuracy and captures the degree to which confidence judgments reflect differences in performance across test items. The gamma coefficient is an association measure that assesses the accuracy of discrimination independently of individuals' response bias (Ackerman & Goldsmith, 2008, 2011; Masson & Rotello, 2009; Sarac & Tarhan, 2009; Zabrocky et al., 2009). Readers show high resolution if they rate their confidence as higher when their answers are indeed correct but rate their confidence as lower when their answers are incorrect (Baker & Brown, 1984).

Despite statistical differences between the absolute and relative measures of monitoring, some studies have found them to be correlated with one another within individuals (Hadwin & Webster, 2013). However, others emphasize that calibration bias and resolution tap into distinct aspects of monitoring and so do not necessarily correlate (Griffin, Wiley, & Salas, 2013; Wiley et al., 2016).

Given this variability in the literature, in the present study we chose to include measures of both absolute and relative accuracy (as advocated by Hadwin & Webster, 2013). We selected the following parameters:

- Confidence ratings were collected for both text comprehension and test performance.
- Confidence ratings were collected according to both prediction and postdiction paradigms.
- Calibration bias and resolution were the numerical measures derived from the confidence ratings.

The Current Study

As the studies reviewed above demonstrate, the role of monitoring in reading comprehension has received a fair amount of research interest. However, there are still several open issues. We addressed two of these in the current study, investigating university students in Israel across the three languages they used: Arabic, Hebrew, and English. Our specific research questions were as follows:

1. How is monitoring related to reading comprehension across the L1, L2, and L3? Specifically, to what extent is low comprehension associated with greater calibration bias and lower resolution in the nonnative languages of adults, as has been demonstrated for L1 reading comprehension?
2. Is the monitoring accuracy of reading comprehension shared across the languages of trilinguals and across a nonlinguistic domain, or rather, is comprehension monitoring linked to language proficiency, such that it is more accurate in languages in which there is higher proficiency?

Method

Participants

The study included 80 undergraduates from the University of Haifa: 74 females and six males (mean age 21.5, range 19–33). Forty of them were in their first year at the university, and 40 were in their third year.¹ All participants were native Arabic speakers born in Israel and were living in Arabic-speaking communities. For all participants, schooling in elementary and high school had been conducted in their native and dominant language, Arabic. Hebrew, which is

the majority language in Israel, had been studied as a second language in school beginning in the third grade; it is also often encountered in the environment and the media. English had been studied as an additional, foreign language, from the fourth grade. At the time of testing, participants were enrolled students at the University of Haifa, where the language of instruction is Hebrew and where course readings are in either Hebrew or English. None of the participants had experienced living in a country other than Israel, and none of them had a history of a learning disability, an attention deficit disorder, or a hearing or vision impairment. 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 ethics Review Board (IRB).

Materials

Language Proficiency

We used one subjective and one objective measure of proficiency.

Language experience and proficiency questionnaire. For the subjective measure, we used a Hebrew translation (Prior & Beznos, 2009) of the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian, Blumenfeld, & Kaushanskaya, 2007²). The questionnaire included questions regarding participants' history and context of acquiring the languages they knew and their present language use. Participants rated on a 10-point scale their proficiency in speaking, listening, reading, and writing each of the languages (Arabic, Hebrew, and English), and we averaged these ratings to arrive at a single self-rated proficiency score for each language.

Multilingual naming test. For the objective measure of proficiency, we used the Multilingual Naming Test (MINT; Gollan, Weissberger, Runqvist, Montoya, & Cera, 2012), a picture-naming task developed for English and Hebrew (as well as Spanish and Mandarin). For the current study, the test was also translated into Arabic and piloted. Participants completed the MINT in Arabic, Hebrew, and English (split-half reliability in the current sample was .549 for Arabic,³ .930 for Hebrew, and .934 for English). Sixty-seven pictures of different objects were presented to the participants, in order of increasing difficulty (see the list of items in Appendix S1 in the Supporting Information online or at <https://osf.io/emaz3/>). Participants had to name each one of the objects orally. If a participant could not name a specific picture, it was counted as an error and the tester proceeded to the next item. The final score on this task was the number of correctly named pictures in each language.

Reading Comprehension Task

Participants read three expository texts in each language from a computer screen, with each text being followed by five four-option multiple-choice questions (Cronbach's alpha was .459 for Arabic, .668 for Hebrew, and .599 for English). Texts and questions were adapted from preparation guides for a university entry exam and were on a variety of topics (listed below). We identified 20 texts (seven in Arabic, six in Hebrew, and seven in English) as an initial text pool, and each text was read by 28 pilot undergraduates (who did not participate in the main study). Based on the pilot data, we selected three texts in each language, aiming for a medium level of accuracy, roughly equated across the three languages. Mean accuracy scores on the texts were 59% for Arabic, 64% for Hebrew, and 53% for English. We used a one-way ANOVA to test for differences between scores on the texts, followed by Tukey HSD posthoc tests which were all *ns*, $p > .136$ (for standard deviations, test statistics, and exact p values, see Appendix S2 in the Supporting Information online). Difficulty levels were equated across languages because monitoring accuracy and calibration have been linked to absolute performance levels (Garner, 1980; Maki et al., 1994; Pressley & Ghatala, 1988). Thus, given the language profile of our participants, the texts in Arabic (L1) were objectively the most complex, in terms of length, the inclusion of less-frequent vocabulary words, and syntactic complexity (as evaluated by the research team), followed by the texts in Hebrew (L2); finally, the texts in English (L3) were objectively the least complex. The average length of the chosen texts was 486 words in Arabic, 408 words in Hebrew, and 296 words in English.

The topics of the L1 texts were sign language, human behavior, and the plaintiff and defendant parties of the judiciary; those of the L2 texts were genetic engineering, episodic memory, and game theory in politics; and those of the L3 texts were extraversion–introversion, chemical compounds, and a tale of two revolutions (all texts are available in Appendix S3 in the Supporting Information online and also at <https://osf.io/emaz3/>). Participants' prior knowledge of these topics was not directly examined. However, the text topics relate to general knowledge, and topics in all three languages were deemed by the research team to be of similar familiarity. Each text was followed by five four-option multiple-choice questions that participants were required to answer, testing different levels of comprehension such as inference, summarizing, and recall of details (see questions and correct answers at <https://osf.io/emaz3/>). Within each language, the three texts were presented in a random order to each participant. Further, following each text, the order of the questions and the answers was also

randomly determined for each participant. Question order was randomized to control for order effects in the confidence judgments.

Comprehension Monitoring Task

We obtained both prediction and postdiction confidence judgments from the participants.

Prediction confidence judgments. After reading each text, but before seeing any of the questions, participants were asked, “How accurate do you think you will be when responding to comprehension questions regarding the text you have just read?” They rated their confidence using a continuous analog scale between 25% and 100% (Ackerman & Goldsmith, 2011).

Postdiction confidence judgments. Following each multiple-choice question, participants were asked, “How confident are you that you selected the correct answer?” They again rated their confidence on a continuous scale between 25% and 100%.

Monitoring of a Nonlinguistic Task

Raven’s standard progressive matrices test (Raven, 1938) is commonly used to measure nonverbal intelligence (the odd–even split-half reliability is .96). Fifteen matrices were chosen from the original test and presented to participants item by item. Each item included a matrix of geometric figures with one piece missing, and participants were asked to choose the correct missing piece to complete the pattern from a set of eight answer choices. For illustration purposes, the participants were presented first with an example (see Appendix S4 in the Supporting Information online). Questions and answers were completed nonverbally. The items were presented in an increasing level of difficulty, on a computer screen. For this task, only postdiction confidence judgments were collected. Thus, participants were asked to rate their confidence regarding each item from the Raven test, by answering the question “How confident are you that you selected the correct answer?” The ratings were recorded on a continuous scale between 12.5% and 100%.

Procedure

Each participant completed three separate sessions, with an interval of 2–7 days between sessions. Only one language was tested in each session, randomly ordered across participants. Each session lasted up to 60 minutes and took place individually in a quiet room at a research laboratory at the university. Participants completed the LEAP-Q questionnaire online before the first session. At

the beginning of the first session, each participant gave informed consent, and completed the Raven test (including confidence judgments).

Next, in each session, three texts in the relevant language were presented on a computer monitor. Participants were first asked to read each text for comprehension at their own pace for an unlimited time, and to estimate their comprehension of the text. The text was then followed by five comprehension questions, each presented on a separate screen. Participants were able to return to the text from the question screens as many times as they wished. After selecting their response, participants rated their confidence in their answer, and proceeded to the next question. Once a response was selected and a confidence rating was entered, they could not be returned to or changed. Participants were given unlimited time for answering the questions and estimating their confidence. All performance was self-paced, and participants controlled the presentation of the subsequent texts and questions. The following data were recorded during the reading comprehension task: accuracy of responses to the questions and participants' confidence judgments (per text and per question).⁴ Presentation and response collection were controlled by E-Prime (PST, www.pstnet.com; Schneider, Eschman, & Zuccolotto, 2002; version 3.0).

In addition, in each session, the MINT test was administered after the reading comprehension task, according to the language for that session.

Results

An initial examination of results demonstrated that seven participants showed comprehension performance at less than chance level in at least one of the languages (average accuracy below 30% across the three texts). These participants were excluded from all analyses, so that the final sample included 73 participants.

As expected from their language background, participants were highly proficient in Arabic, were less so in their L2 Hebrew, and generally showed the lowest proficiency in English, their L3, both in self-rated proficiency and in the MINT score (see Table 1).

Computing Monitoring Measures

We calculated calibration bias score in prediction by subtracting the average performance for each text from the predicted score for the same text. We computed calibration bias at postdiction by calculating the absolute discrepancy between the average performance for each text and the average estimated scores of all questions for that text. Prediction and postdiction judgments were highly correlated in all three languages (for Arabic $r = .904$; for Hebrew $r = .909$;

Table 1 Descriptive statistics for the study variables in the three languages and in the nonverbal task (the Raven test)

Domain	Measure	Language or nonverbal task	Descriptive statistics		
			<i>M</i>	<i>SE</i>	<i>p</i>
Language proficiency	LEAP-Q (self-rated proficiency out of 10)	L1	9.7	0.6	<.001
		L2	8.3	1.3	
		L3	6.8	1.4	
	MINT (correct items out of 67)	L1	58.4	2.8	
		L2	33.8	9.6	
		L3	29.1	8.7	
Reading comprehension	% accuracy (reading comprehension/Raven)	L1	59.7	14.1	<.001
		L2	67.2	17.4	
		L3	62.9	16.3	
		Raven	48.0	21.8	
Metacognitive monitoring	Confidence judgments of postdiction (magnitude out of 100)	L1	80.29	10.9	<.001
		L2	78.14	11.3	
		L3	72.80	12.9	
		Raven	70.78	14.7	
	Absolute calibration bias prediction (0 indicates perfect calibration, higher values indicate greater bias)	L1	22.65	13.8	<.001
		L2	15.92	11.9	
		L3	15.80	11.9	
	Absolute calibration bias post-diction (difference between confidence judgments of post-diction and reading accuracy, 0 indicates perfect calibration, higher values indicate greater bias)	L1	25.59	10.9	<.001
		L2	20.55	9.4	
		L3	20.03	10.1	
		Raven	23.80	14.6	
	Resolution (gamma coefficient, max = 1)	L1	0.37	0.4	<.001
L2		0.33	0.5		
L3		0.51	0.4		
Raven		0.64	0.3		

Note. Values of *p* indicate significant differences in means between all groups on the group level in a repeated measures ANOVA test. LEAP-Q = Language Experience and Proficiency Questionnaire; L1 = first language (here, Arabic); L2 = second language (here, Hebrew); L3 = third language (here, English); MINT = Multilingual Naming Test.

for English $r = .884$, all $ps < .001$), so all further analyses focus only on the postdiction judgments. Because of our interest in monitoring accuracy, we used the absolute difference between confidence and comprehension.

In addition, we calculated a resolution score for each participant across all three texts in each language, using the Goodman–Kruskal gamma correlation of the concordance between the confidence judgment of each comprehension item and its accuracy. The results for the monitoring measures are shown in Table 1.

Monitoring and Comprehension in the L1, L2, and L3

To answer our first research question, we examined whether more accurate calibration and higher resolution were linked with improved comprehension across all three languages. Thus, we analyzed comprehension performance using a linear mixed-effects model (Model 1; Baayen, Davidson, & Bates, 2008) in R (R Core Team, 2018), with the lme4 library (version 1.1-7; Bates, Maechler, Bolker, & Walker, 2015). We created plots using the ggplot2 package (version 2.3.00; Wickham, 2016).

The dependent variable was the percentage of correct responses per text. The model included the following fixed effects:

- language (Arabic, Hebrew, English), which was a categorical variable, with Arabic set as the reference, and was deviation coded;
- absolute values of calibration bias (a continuous variable);
- resolution (continuous);
- language proficiency measured by the MINT⁵ (continuous); and
- text length (continuous).

We included text length as a control variable in the model, because readers tend to show better comprehension of shorter than of longer texts (Commander & Stanwyck, 1997). Continuous variables were log transformed to normalize the distribution.

The model also included the interaction between language and calibration bias, and the interaction between language and proficiency. The model included a maximal random effect structure (Barr, Levy, Scheepers, & Tily, 2013), with random intercepts by participants and by items (thus controlling for unmeasured sources of heterogeneity between texts), and random slopes for language by participant. We assessed significance based on the Satterthwaite approximations of degrees of freedom in the lmerTest function, as described by Luke (2017), to produce acceptable Type 1 error rates (see full model in Table 2).

Table 2 Fixed and random effects for Model 1, predicting comprehension accuracy

Fixed effects	β	<i>SE</i>	<i>t</i>	<i>p</i>
Language	121.561	15.586	7.799	<.001
Proficiency	0.165	0.153	1.079	0.281
Calibration bias	-0.755	0.042	-17.883	<.001
Resolution	-1.925	1.793	-0.677	0.498
Text length	-0.105	0.033	-3.152	0.025
Language \times calibration bias				
Hebrew vs. Arabic	0.065	0.096	0.673	0.501
English vs. Arabic	0.368	0.099	3.726	<.001
Language \times proficiency				
Hebrew vs. Arabic	0.955	0.427	2.233	0.026
English vs. Arabic	1.178	0.430	2.735	0.006
Random effects		Variance σ^2		
Subjects		58.33		
Items		33.76		
Slope: ^a language Hebrew		21.16		
Slope: ^a language English		43.80		
Residual		240.89		

Note. Interactions are presented by the different levels of the categorical variable “language,” which were dummy coded to probe the interaction, such that each level is presented in relation to Arabic (the reference level).

^aRandom slope adjustments for language across subjects.

The effect of language was significant ($\beta = 121.561$, $SE = 15.586$, $t = 7.799$, $p < .001$), demonstrating higher comprehension scores in Hebrew than in English and Arabic (see Table 1). The main effect of proficiency, as indicated by vocabulary knowledge, was not significant ($\beta = 0.165$, $SE = 0.153$, $t = 1.079$, $p = .281$), so that, on the level of individual differences, no association was found between comprehension performance and language proficiency (but this was modulated by a two-way interaction with language; see below). The effect of text length was significant ($\beta = -0.105$, $SE = 0.033$, $t = -3.152$, $p = .025$), indicating that longer texts seemed to be broadly associated with poorer comprehension performance, within each language. Importantly, calibration bias contributed significantly to comprehension ($\beta = -0.755$, $SE = 0.042$, $t = -17.883$, $p < .001$); specifically, smaller bias was associated with higher comprehension scores. On the other hand, the main effect of resolution was not significant ($\beta = -1.925$, $SE = 1.793$, $t = -0.677$, $p = .498$).

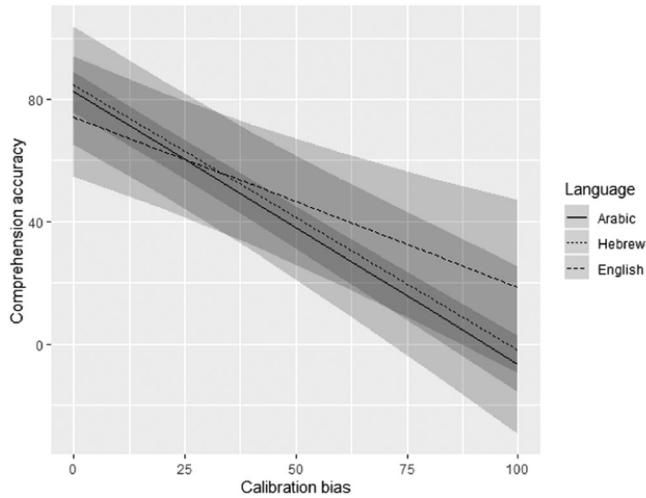


Figure 1 The relation between calibration bias and comprehension accuracy by language (Model 1).

We also found a significant interaction between calibration bias and language ($F = 7.348, p < .001$). Follow-up contrasts demonstrate that the association between comprehension and monitoring was weaker (though still significant) in English than in Arabic and Hebrew, which did not differ from each other (see Table 2 and Figure 1). Finally, we found a significant interaction between language and proficiency ($F = 3.847, p = .023$; see Table 2 and Figure 2). In the follow-up analyses, we computed a regression model for predicting comprehension accuracy by language proficiency for each language separately. The models indicated that in Hebrew and English, language proficiency predicted comprehension accuracy significantly ($p < .001$ for both), whereas in Arabic, language proficiency did not contribute to the model ($p = .132$).

To summarize, we found different levels of comprehension accuracy in the three languages, in which the highest scores were found in Hebrew. Although texts were selected in accordance with participants' proficiency profiles (i.e., texts in Arabic were objectively more difficult than texts in Hebrew, which in turn were objectively more difficult than texts in English), this is still a surprising finding, given that Hebrew was the participants' L2, a point to which we return in the discussion section. Further, although an association was found in Hebrew and English between comprehension and proficiency, no

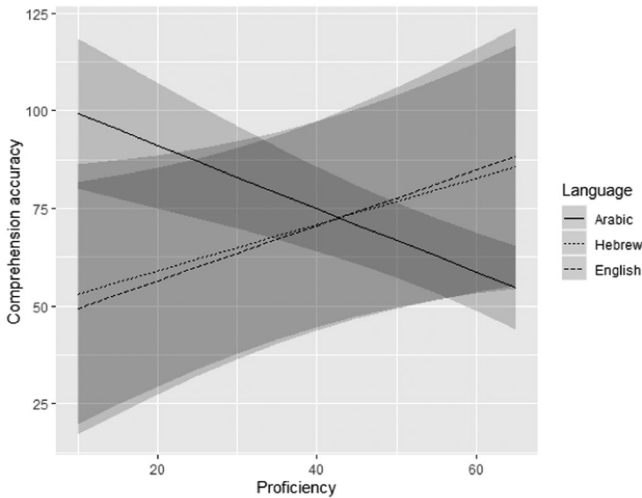


Figure 2 The relation between language proficiency and comprehension accuracy by language (Model 1).

such association was found in Arabic, which we attribute to the low reliability of the proficiency measure (the MINT vocabulary test) in Arabic (see Note 3) and a likely ceiling effect on the Arabic version of the test given that this was the participants' L1 (note that participants scored an average of 58.4 out of 67 with a relatively low standard error suggesting that there may have been low variance). Critically, accurate monitoring was associated with higher comprehension accuracy across all three languages. Finally, the association between monitoring and comprehension was attenuated in the L3, participants' least proficient language.

Is Comprehension Monitoring Shared or Language-Specific?

To investigate the second research question, we first analyzed monitoring accuracy using a linear mixed-effects model (Model 2). The dependent variable was calibration bias. The model included the following fixed effects: language (Arabic, Hebrew, English), which was a categorical variable, with Arabic set as the reference, and was deviation coded; language proficiency measured by the MINT (a continuous variable), which was log transformed to normalize the distribution; and text length (a continuous variable). The model also included the interaction between language and proficiency. Subject and item were included in the model as random effects (see Table 3).

Table 3 Fixed and random effects for Model 2, predicting calibration bias

Fixed effects	β	SE	t	P
Language	4.069	33.939	0.120	0.905
Proficiency	0.382	18.411	0.021	0.983
Text length	0.036	0.026	1.347	0.236
Language \times proficiency				
Hebrew vs. Arabic	-56.421	54.247	-1.040	0.299
English vs. Arabic	-46.300	54.085	-0.856	0.392
Random effects		Variance σ^2		
Subjects		15.18		
Items		20.54		
Residual		241.76		

Note. Interactions are presented by the different levels of the categorical variable "language," which were dummy coded to probe the interaction, such that each level is presented in relation to Arabic (the reference level).

The model yielded no significant effects or interactions. Calibration accuracy was equivalent across the three languages ($\beta = 4.069$, $SE = 33.939$, $t = -0.120$, $p = .905$). Text length also was not a significant predictor of calibration ($\beta = 0.036$, $SE = 0.026$, $t = 1.347$, $p = .236$). Further, we did not find an association between objective proficiency in the language (measured by the MINT) and calibration accuracy in that language ($F = 0.786$, $p = .456$). These findings suggest that comprehension monitoring is not associated with proficiency. Monitoring accuracy was not lower in general in participants' least proficient language as a group, nor was lower proficiency in a given language at the individual level associated with less accurate monitoring.

Next, we wished to address the question of whether a given trilingual individual showed similar tendencies in monitoring accuracy across the three languages and the nonlinguistic task. To this end, we calculated Pearson correlations in postdiction calibration bias and resolution across languages (see Tables 4 and 5). Results demonstrated that calibration bias was positively and significantly correlated between Arabic and Hebrew ($r = .478$, $p < .001$), and between Hebrew and English ($r = .347$, $p = .003$), but not between Arabic and English ($r = .144$, $p = .226$). Alpha levels were corrected using the Holm Bonferroni correction with a corrected level of 0.025 as each set of language data was used in two comparisons. For this measure, we also correlated calibration bias in the reading comprehension task with bias in the nonlinguistic task and found significant correlations for Arabic ($r = .392$, $p = .001$) and Hebrew

Table 4 Correlations across languages and the nonlinguistic task (the Raven test) for calibration bias

	Hebrew		English		Raven	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
Arabic	.478	< .001	.144	.226	.392	.001
Hebrew			.347	.003	.353	.002
English					.225	.071

Table 5 Correlations across languages and the nonlinguistic task (the Raven test) for resolution

	Hebrew		English		Raven	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
Arabic	.021	.861	.173	.152	-.081	.492
Hebrew			.003	.980	.018	.940
English					.099	.461

($r = .353, p = .002$), but not for English ($r = .225, p = .071$). These results demonstrate medium-strength correlations among bias in Arabic, Hebrew, and the nonlinguistic task, which indicates that participants who tended to show accurate monitoring in L1 Arabic also tended to show more accurate monitoring in L2 Hebrew and in the nonlinguistic task. However, monitoring accuracy of L3 English reading comprehension was mostly unassociated with the other tasks, with the exception of a medium-sized correlation with monitoring in the L2.

In contrast, there were no significant correlations for resolution across the different languages of the participants, or with the nonlinguistic task (all $ps > .15$; see Table 5).

Discussion

The present study investigated comprehension monitoring in trilingual undergraduate speakers of Arabic, Hebrew, and English, and focused on two main questions. First, we asked whether monitoring (measured by calibration) is related to reading comprehension in the L1, L2, and L3. Our results demonstrate a consistent relationship between monitoring and comprehension across all three languages, extending previous findings of the importance of metacognition in reading comprehension in a L1 to nonnative languages as well. Second, we

asked whether comprehension monitoring is shared across languages and can be used similarly by individual readers across the different languages they read or whether it is linked to language proficiency, such that readers will exhibit more accurate monitoring in more proficient languages. The results lend some support to both possibilities: Monitoring seemed to be utilized similarly by individuals to support comprehension across the L1 and L2, but was less well generalized to the L3.

Monitoring and Reading Comprehension Across the L1, L2, and L3

Across all three languages examined in the current study, better reading comprehension was related to low calibration bias, that is, more accurate monitoring. Good monitoring may have been the reason why individuals were more accurate on the comprehension test or, vice versa, accurate performance and devoting more resources to the text may have helped individuals to be more aware of their cognitive processes. Thus, the present finding of an association does not allow us to determine the direction of influence, and it is most likely that relations between comprehension and monitoring are reciprocal (Boulware-Gooden, Carreker, Thornhill, & Joshi, 2007; Maki, Foley, Kajer, Thompson, & Willert, 1990). Notably, the current pattern of association between reading comprehension and monitoring replicates findings from several previous studies examining readers' performance in their native language or L1 (Kasperski & Katzir, 2013; Langer et al., 1990; Zabucky et al., 2009), and extends them to the L2 and L3. This finding supports the generality of componential models of reading comprehension such as the RAND model and the landscape model (Van den Broek et al., 1999), which describe the importance of metacognition in supporting reading comprehension, and show that they apply to L2 and L3 reading as well.

Moreover, the current study used online measures of monitoring, unlike most previous L2 studies, which focused on strategy knowledge, mainly through questionnaires, think-aloud tasks, or interviews (Lin & Yu, 2015; Sheorey & Mokhtari, 2001; Trapman et al., 2017; Tsai et al., 2010; Van Gelderen et al., 2004). The current findings, therefore, extend these previously identified patterns to encompass processing measures of monitoring as well. This finding further generalizes the importance of monitoring in reading comprehension not only in the L1, but in the different languages that individuals use, and suggests the relevance of developing learning and intervention programs concerning monitoring in reading comprehension.

Interestingly, the link between monitoring and comprehension was somewhat weaker in the L3, participants' least proficient language, than in the L1 and

L2. This finding can most readily be explained by the claim that monitoring, as a higher order skill, might be easily available to readers who have mastered the lower level skills required for reading comprehension, such as decoding and vocabulary (Grabe & Stoller, 2013; Perfetti et al., 2005). Previous studies have reported that the transfer of reading comprehension strategies and monitoring might be limited by language proficiency (Han & Stevenson, 2008; Schoonen, Hulstijn, & Bossers, 1998; Tsai et al., 2010), a notion that also aligns well the ideas of linguistic interdependence and proficiency thresholds (Cummins, 1976). Importantly, because the current study investigated trilinguals, it allows us to conclude that multilinguals can use higher order skills in a nonnative language (as evident in their L2 performance), but that they might be utilized more easily beyond a specific level of proficiency.

Surprisingly, the analysis showed lower comprehension performance in Arabic, the L1, than in Hebrew and English. Importantly, text difficulty was not matched across languages, but rather in a pretest we selected texts that led to similar levels of comprehension performance across the three languages. Equating levels of comprehension across language was important, because previous research has demonstrated that monitoring accuracy and calibration can vary based on task difficulty (e.g., Maki et al., 1994). To achieve this, we intended the Arabic texts to be objectively the most difficult, the Hebrew texts to be less difficult, and the English texts to be the easiest (in line with participants' language profile). We relied on pretest data to ascertain text comprehension difficulty because there are currently no objective methods of assessing text complexity that can be used across all three languages (such as the Flesch–Kincaid readability tests <https://stars.library.ucf.edu/istlibrary/56> or Lexile scores <https://lexile.com/>). We hypothesize that although the pretest data were collected from participants belonging to the same population as those participating in the main study, unexpected differences in prior knowledge of the text topics might have led to the discrepancy in results between these two groups of participants. Crucially, however, our research question did not focus on comprehension levels per se, but rather on the relation between monitoring and comprehension.

Is Monitoring Domain-General?

The second research question addressed in the current study was whether monitoring of comprehension is best conceptualized as a general skill that can be shared across languages within an individual, similar to other metalinguistic abilities (Durgunoğlu, 2002; Melby-Lervåg & Lervåg, 2011), or whether it might be better understood in the framework of language proficiency

(Cummins, 1976; Perfetti et al., 2005). The current results mostly support the notion of domain-general monitoring abilities, in that we did not find statistically significant differences between participants' monitoring accuracy in the three languages they used. Further, in the current study calibration bias was not associated with proficiency in the language. These findings therefore could suggest that monitoring in reading comprehension was not strongly influenced by an individual's proficiency in the language tested (though see below for a more nuanced interpretation).

However, it is important to qualify this statement, because the participants tested in the current study were university undergraduates who had achieved at least moderate proficiency in both of their nonnative languages and were using them on a daily basis at the time of testing. Thus, it is possible that there is a threshold proficiency that allows the transfer of monitoring to nonnative languages (e.g., Schoonen et al., 1998) and that the current participants had already surpassed that threshold in all three languages. Future studies recruiting participants with a wider range of proficiency in their foreign languages could directly test this possibility.

To further understand the extent to which monitoring accuracy is domain-general, we also conducted correlation analyses. The results mostly demonstrated significant positive correlations in individual participants' calibration bias: Participants who were well calibrated in the L1 also tended to be better calibrated in the L2, and L2 calibration was linked with L3 calibration. This pattern of results is predicted by the domain-general approach (Chung et al., 2019; Geva & Ryan, 1993), which claims that metacognitive skills, once acquired, can be applied by readers in their various languages. Some previous research on metacognition in reading comprehension across the languages of bilinguals has also reported cross-language similarities, albeit using questionnaire and self-report instruments (Han, 2013; Jiménez et al., 1996; Langer et al., 1990).

This domain-general interpretation was further supported by the finding that calibration bias in L1 and L2 reading comprehension was also significantly and positively correlated with calibration bias in the nonlinguistic task (the Raven test). This leads to a conceptualization of monitoring ability, as expressed through calibration bias, as a domain-general skill that can be applied by individuals to linguistic and nonlinguistic domains in a similar manner (for a similar conceptualization, see Stankov, Kleitman, & Jackson, 2015).

However, these conclusions must be somewhat qualified by two other facets of the current results. First, monitoring accuracy in English, the L3, was not associated with monitoring accuracy in either Arabic or the nonlinguistic task (although it was moderately associated with monitoring accuracy in Hebrew,

the L2). One explanation might be that trilinguals approach reading in their L3, which is a foreign language, somewhat differently from reading in their two more proficient languages, which they use for everyday communication. Thus, perhaps due to the subjective and objective typological distance and differences in the domains of use between the languages, individuals might apply different types and amounts of reading strategies and monitoring behaviors, leading to lower correlations of L3 monitoring with the other tasks. Alternatively, participants' lower proficiency in their L3 (relative to their L1 and L2) might have limited the ability of some participants to fully transfer their comprehension monitoring from their L1. This explanation aligns well with our other finding of weaker links between monitoring and comprehension in the L3 relative to in the L1 and in the L2. Here as well, future studies investigating participants with a wider range of foreign language proficiency could further inform our understanding of this issue.

Second, although we did find significant positive correlations in monitoring ability across tasks, these were only moderately sized, accounting for less than half of the variability in performance. This suggests that there are most probably a number of underlying competencies or skills that contribute to accurate monitoring, which might be engaged to different degrees by individuals across tasks and languages. More research is needed to further elucidate the underlying competencies that contribute to accurate calibration. One promising avenue of research could be an investigation of what confidence judgments are based on, along the lines of research by Dinsmore and Parkinson (2013). That study identified various sources for confidence ratings, including reader characteristics such as prior knowledge and task characteristics of the texts and question items, and found that readers vary in the sources they recruit for confidence ratings. Future research could usefully compare the sources that multilingual readers rely on for confidence judgments in the different languages they speak.

Absolute and Relative Measures of Monitoring

The current study included both a measure of absolute accuracy of monitoring (calibration bias) and a measure of relative accuracy of monitoring (resolution). Whereas calibration bias was significantly linked with comprehension in all three languages, resolution accuracy was not linked to comprehension. In addition, calibration bias was significantly correlated across the L1, the L2, and the nonlinguistic task, whereas resolution showed no significant cross-language or cross-task associations.

A possible explanation for this finding is linked to the psychometric characteristics of the resolution measure used in the current study, namely, the gamma

rank correlation coefficient. Gamma does not capture differences in magnitude of variations in confidence judgments versus performance accuracy, because it relies only on rank ordering of ordinal variables (Wiley et al., 2016). In cases where it is possible, Pearson intraindividual correlations have been suggested to better capture relative monitoring accuracy (resolution), but the design of the current study, with only three texts per language, did not allow us to use this measure (Wiley et al., 2016). Moreover, in the current study the computation of gamma was based on only 15 items, which might not have been a large enough sample, as the gamma coefficient becomes less reliable and less sensitive when based on a small number of items (Spellman, Bloomfield, & Bjork, 2008). Thus, we believe that the measure of relative monitoring accuracy in the current study was not of the necessary quality and stability to allow for meaningful conclusions. Future studies should adopt designs that allow for better measurement of this important construct.

Limitations and Future Directions

One limitation of the current study is that we did not directly assess prior knowledge of the different text topics, which might have influenced comprehension performance. This might be one reason for the unexpected lower comprehension performance in Arabic, the L1, than in Hebrew and English. Another possible explanation for the low comprehension performance in Arabic is that we did not assess objective text complexity and compare it across languages, because there are currently no such objective measures that could be used across all three languages. Instead, different approaches were adopted to control for this limitation. First, we determined text complexity based on accuracy of performance in a pilot, and on that basis selected target texts at the same level of difficulty. Second, in the mixed-model analyses, texts were used as a random effect in order to control for the variability in difficulty within and between languages. Future studies, when possible, should use objective measures to assess text complexity, such as the Flesch–Kincaid readability tests (<https://stars.library.ucf.edu/istlibrary/56>) or Lexile scores (<https://lexile.com/>), as well as a precise measure of prior knowledge. Note that text characteristics such as text difficulty and prior knowledge have been identified as having possible effects on calibration accuracy (Lin & Zabrocky, 1998).

An additional limitation was that resolution, a relative evaluation of calibration accuracy, was not of the necessary quality and stability to allow for meaningful conclusions. We believe that this measure could be very informative in providing the accuracy of discrimination between items independently

of individuals' response bias (Ackerman & Goldsmith, 2008, 2011). Future studies should better measure this important construct.

Further, the moderate-sized correlations in monitoring across the languages and the nonlinguistic task suggest that there are several underlying competencies or skills that contribute to accurate monitoring, which might be engaged to different degrees by individuals across tasks and languages. Thus, additional research is needed to further explore the underlying competencies that contribute to accurate calibration. One important step in this direction could be a focus on cue use and identifying the information that participants rely on in their judgments.

Conclusion

To summarize, the current study clearly demonstrates that metacognitive monitoring, as measured by online calibration, is linked to successful comprehension across all the languages of trilingual university students. Thus, intervention programs for promoting metacognitive monitoring in reading comprehension are likely to be important, not only in the L1, but in the additional languages increasingly used by multilinguals in academic settings. Further, our results mostly support the notion of monitoring ability as a domain-general skill that can be applied by individuals to linguistic and nonlinguistic domains in a similar manner. Thus, at the group level we found equivalent monitoring accuracy across the three languages, and calibration bias was mostly correlated across the languages and the nonlinguistic task. However, at the individual level, our findings suggest that the specific recruitment of comprehension monitoring might vary across languages at different proficiencies, being recruited less in a language in which there is lower proficiency.

These findings could suggest that an intervention program designed to bolster metacognitive skills in a L1 could be successfully generalized to other languages and domains. Future studies should investigate the underlying skills that contribute to better monitoring, as well as developing and evaluating intervention programs that promote and assess accurate calibration and link it to self-regulation skills in learning.

Final revised version accepted 11 January 2020

Notes

- 1 The study was initially designed to assess development of reading comprehension and monitoring in undergraduates through their academic career, by comparing first- and third-year students (Van der Stel & Veenman, 2010). However, when we

included the study year variable in the linear mixed-effects analysis (Model 1), it did not contribute significantly to the model. Therefore, all reported analyses aggregated all participants into a single group, to increase power.

- 2 The original study included factor analyses, and identified separable components for L1 and L2 proficiency. Internal consistency for these components was reported (Cronbach's alpha was for L1 = .92; for L2 = .88). External validity was measured by correlations between self-report and objective measures ($\sim .45$ for L1; $\sim .65$ for L2). In the current study, we relied only on self-reported proficiency for L1, L2, and L3 (four items per language), and we are therefore unable to provide similar information.
- 3 The MINT findings should be interpreted with some caution, however, because the Arabic adaptation of this task was used for the first time in the current study and showed relatively low internal consistency/reliability (.549). The test was used because no other validated tests exist for the same purpose.
- 4 The computerized presentation program also collected reading times for texts and questions, as well as number of returns from question screens to the text, but these data are not presented and analyzed in the current article.
- 5 Correlations between the MINT, the objective proficiency measure, and self-reports from the LEAP-Q were generally high ($.42 < R < .44$). When we ran an equivalent model with self-reported proficiency, the same pattern of results was observed.

Open Research Badges



This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at <https://osf.io/EMAZ3/>. All proprietary materials have been precisely identified in the manuscript.

References

- Ackerman, R., & Goldsmith, M. (2008). Learning directly from screen? Oh-no, I must print it! Metacognitive analysis of digitally presented text learning. In Y. Eshet-Alkalai, A. Caspi, & N. Geri (Eds.), *Proceedings of the Chais Conference on Instructional Technologies Research 2008: Learning in the Technological Era* (pp. 1–7). Raanana, Israel: Open University of Israel.
- Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied*, 17, 18–32. <https://doi.org/10.1037/a0022086>

- Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24*, 1–3. <https://doi.org/10.1016/j.learninstruc.2012.10.003>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1*, 3–38. <https://doi.org/10.1007/bf01326548>
- Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 353–394). New York, NY: Longman.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014. Retrieved from <http://cran.r-project.org/package=lme4>
- Block, E. L. (1992). See how they read: Comprehension monitoring of L1 and L2 readers. *TESOL Quarterly, 26*, 319–343. <https://doi.org/10.2307/3587008>
- Boulware-Gooden, R., Carreker, S., Thornhill, A., & Joshi, R. M. (2007). Instruction of metacognitive strategies enhances reading comprehension and vocabulary achievement of third-grade students. *The Reading Teacher, 61*, 70–77. <https://doi.org/10.1598/rt.61.1.7>
- 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>
- Commander, N. E., & Stanwyck, D. J. (1997). Illusion of knowing in adult readers: Effects of reading skill and passage length. *Contemporary Educational Psychology, 22*, 39–52. <https://doi.org/10.1006/ceps.1997.0925>
- Cummins, J. (1976). The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. *Working Papers on Bilingualism, 9*, 1–43.
- Deacon, S. H., Wade-Woolley, L., & Kirby, J. (2007). Crossover: The role of morphological awareness in French immersion children's reading. *Developmental Psychology, 43*, 732–746. <https://doi.org/10.1037/0012-1649.43.3.732>
- Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. <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*, 228–232. <https://doi.org/10.1111/j.1467-8721.2007.00509.x>
- Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*, 271–280. <https://doi.org/10.1016/j.learninstruc.2011.08.003>
- Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, *24*, 58–61. <https://doi.org/10.1016/j.learninstruc.2012.05.002>
- Durgunoğlu, A. Y. (2002). Cross-linguistic transfer in literacy development and implications for language learners. *Annals of Dyslexia*, *52*, 189–204. <https://doi.org/10.1007/s11881-002-0012-y>
- 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*, *11*, 29–63. <https://doi.org/10.1023/a:1007996502372>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*, 906–911. <https://doi.org/10.1037/0003-066x.34.10.906>
- García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, *84*, 74–111. <https://doi.org/10.3102/0034654313499616>
- Garner, R. (1980). Monitoring of understanding: An investigation of good and poor readers' awareness of induced miscomprehension of text. *Journal of Literacy Research*, *12*, 55–63. <https://doi.org/10.1080/10862968009547352>
- Geva, E., & Ryan, E. B. (1993). Linguistic and cognitive correlates of academic skills in first and second languages. *Language Learning*, *43*, 5–42. <https://doi.org/10.1111/j.1467-1770.1993.tb00171.x>
- Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, *116*, 119–136. <https://doi.org/10.1037/0096-3445.116.2.119>
- Glover, J. A. (1989). Reading ability and the calibrator of comprehension. *Educational Research Quarterly*, *13*(3), 7–11.
- Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *15*, 594–615. <https://doi.org/10.1017/s1366728911000332>
- Grabe, W. P., & Stoller, F. L. (2013). *Teaching and researching: Reading*. New York, NY: Routledge. <https://doi.org/10.4324/9781315726274>

- Griffin, T. D., Wiley, J., & Salas, C. R. (2013). Supporting effective self-regulated learning: The critical role of monitoring. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 19–34). New York, NY: Springer.
- Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgments of confidence. *Learning and Instruction, 24*, 37–47. <https://doi.org/10.1016/j.learninstruc.2012.10.001>
- Han, F. (2013). The relationship between L1 and FL comprehension monitoring with Chinese EFL readers. *The International Journal of Literacies, 19*, 13–24. <https://doi.org/10.18848/2327-0136/cgp/v19i01/48841>
- Han, F., & Stevenson, M. (2008). Comprehension monitoring in first and foreign language reading. *University of Sydney Papers in TESOL, 3*, 73–110.
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*, 127–160. <https://doi.org/10.1007/bf00401799>
- 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*, 90–112. <https://doi.org/10.1598/rrq.31.1.5>
- Johnston, T. C., & Kirby, J. R. (2006). The contribution of naming speed to the simple view of reading. *Reading and Writing, 19*, 339–361. <https://doi.org/10.1007/s11145-005-4644-2>
- 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*, 59–84. <https://doi.org/10.1080/02702711.2011.580042>
- Kendeou, P., Van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology, 101*, 765. <https://doi.org/10.1037/a0015956>
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163–182. [https://doi.org/10.1016/s0166-4115\(08\)61551-4](https://doi.org/10.1016/s0166-4115(08)61551-4)
- Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.), *Converging evidence in language and communication research* (Vol. 3, pp. 157–170). Thematics: Interdisciplinary studies. Amsterdam, Netherlands: John Benjamins. <https://doi.org/10.1075/celcr.3.14kin>
- Klassen, R. M. (2007). Using predictions to learn about the self-efficacy of early adolescents with and without learning disabilities. *Contemporary Educational Psychology, 32*, 173–187. <https://doi.org/10.1016/j.cedpsych.2006.10.001>
- Kroll, M. D., & Ford, M. L. (1992). The illusion of knowing, error detection, and motivational orientations. *Contemporary Educational Psychology, 17*, 371–378. [https://doi.org/10.1016/0361-476x\(92\)90075-a](https://doi.org/10.1016/0361-476x(92)90075-a)

- Landi, N. (2010). An examination of the relationship between reading comprehension, higher-level and lower-level reading sub-skills in adults. *Reading and Writing, 23*, 701–717. <https://doi.org/10.1007/s11145-009-9180-z>
- Langer, J. A., Bartolome, L., Vasquez, O., & Lucas, T. (1990). Meaning construction in school literacy tasks: A study of bilingual students. *American Educational Research Journal, 27*, 427–471. <https://doi.org/10.3102/00028312027003427>
- Lesaux, N., Geva, E., Koda, K., Siegel, L., & Shanahan, A. (2006). Development of literacy in second-language learners. In D. August & T. Shanahan (Eds.), *Developing Reading and Writing in Second Language Learners: Lessons from the report of the National Literacy panel on Language minority children and youth* (pp. 27–60). New York, NY: Routledge.
- 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*, 286–306. <https://doi.org/10.1111/1467-9817.12012>
- Lin, L.-M., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology, 22*, 111–128. <https://doi.org/10.1080/02702710119125>
- Lin, L.-M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*, 345–391. <https://doi.org/10.1006/ceps.1998.0972>
- Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*, 1494–1502. <https://doi.org/10.3758/s13428-016-0809-y>
- Maki, R. H. (1998). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition, 26*, 959–964. <https://doi.org/10.3758/BF03201176>
- 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*, 609–616. <https://doi.org/10.1037/0278-7393.16.4.609>
- Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review, 1*, 126–129. <https://doi.org/10.3758/BF03200769>
- 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*, 723–731. <https://doi.org/10.1037/0022-0663.97.4.723>
- Marian, V., Blumenfeld, H., & Kaushanskaya, R. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language & Hearing Research, 50*, 940–967. [https://doi.org/10.1044/1092-4388\(2007/067\)](https://doi.org/10.1044/1092-4388(2007/067))

- Masson, M. E., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 509–527. <https://doi.org/10.1037/a0014876>
- Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, 34, 114–135. <https://doi.org/10.1111/j.1467-9817.2010.01477.x>
- Pasquarella, A., Chen, X., Lam, K., Luo, Y. C., & Ramirez, G. (2011). Cross-language transfer of morphological awareness in Chinese–English bilinguals. *Journal of Research in Reading*, 34, 23–42. <https://doi.org/10.1111/j.1467-9817.2010.01484.x>
- Perfetti, C. A., & Hart, L. (2001). The lexical basis of comprehension skill. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 67–86). Washington, DC: American Psychological Association. <https://doi.org/10.1037/10459-004>
- Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Malden, MA: Blackwell. <https://doi.org/10.1002/9780470757642.ch13>
- Perfetti, C. A., Marron, M. A., & Foltz, P. W. (1996). Sources of comprehension failure: Theoretical perspectives and case studies. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 137–165). Mahwah, NJ: Lawrence Erlbaum.
- Pieschl, S. (2009). Metacognitive calibration: An extended conceptualization and potential applications. *Metacognition and Learning*, 4, 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, 454–464. <https://doi.org/10.2307/747643>
- Prior, A., & Beznos, M. (2009). Hebrew version of the Language Experience and Proficiency Questionnaire. Unpublished instrument. Retrieved from <https://www.iris-database.org/iris/app/home/detail?id=york%3a822288&ref=search>
- R Core Team. (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Ramirez, G., Chen, X., Geva, E., & Kiefer, H. (2010). Morphological awareness in Spanish-speaking English language learners: Within and cross-language effects on word reading. *Reading and Writing*, 23, 337–358. <https://doi.org/10.1007/s11145-009-9203-9>
- RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Washington, DC: RAND Education.

- Raven, J. C. (1938). *Progressive Matrices: Sets A, B, C, D, and E*. London: HK Lewis. published by HK Lewis.
- Saiegh-Haddad, E., & Geva, E. (2008). Morphological awareness, phonological awareness, and reading in English–Arabic bilingual children. *Reading and Writing, 21*, 481. <https://doi.org/10.1007/s11145-007-9074-x>
- Sarac, S., & Tarhan, B. (2009). Calibration of comprehension and performance in L2 reading. *International Electronic Journal of Elementary Education, 2*, 167–179.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools.
- Schoonen, R., Hulstijn, J., & Bossers, B. (1998). Metacognitive and language-specific knowledge in native and foreign language reading comprehension: An empirical study among Dutch students in grades 6, 8 and 10. *Language Learning, 48*, 71–106. <https://doi.org/10.1111/1467-9922.00033>
- Shorey, R., & Mokhtari, K. (2001). Differences in the metacognitive awareness of reading strategies among native and non-native readers. *System, 29*, 431–449. [https://doi.org/10.1016/s0346-251x\(01\)00039-2](https://doi.org/10.1016/s0346-251x(01)00039-2)
- Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. RAND Corporation.
- Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2008). Measuring memory and metamemory: Theoretical and statistical problems with assessing learning (in general) and using gamma (in particular) to do so. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 95–114). New York, NY: Psychology Press.
- Stankov, L., Kleitman, S., & Jackson, S. A. (2015). Measures of the trait of confidence. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 158–189). London, UK: Academic Press. <https://doi.org/10.1016/B978-0-12-386915-9.00007-3>
- Taki, S. (2016). Metacognitive online reading strategy use: Readers' perceptions in L1 and L2. *Journal of Research in Reading, 39*, 409–427. <https://doi.org/10.1111/1467-9817.12048>
- Thiede, K. W., Anderson, M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73. <https://doi.org/10.1037/0022-0663.95.1.66>
- Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85–106). New York, NY: Routledge/Taylor & Francis Group. <https://doi.org/10.4324/9780203876428>
- Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of Research in Reading, 32*, 383–401. <https://doi.org/10.1111/j.1467-9817.2009.01401.x>

- 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*, 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 den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The Landscape Model of reading. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 71–98). Mahwah, NJ: Lawrence Erlbaum.
- Van der Stel, M., & Veenman, M. V. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences*, *20*, 220–224. <https://doi.org/10.1016/j.lindif.2009.11.005>
- Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Snellings, P., Simis, A., & Stevenson, M. (2003). Roles of linguistic knowledge, metacognitive knowledge and processing speed in L3, L2 and L1 reading comprehension: A structural equation modeling approach. *International Journal of Bilingualism*, *7*, 7–25. <https://doi.org/10.1177/13670069030070010201>
- 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*, 19. <https://doi.org/10.1037/0022-0663.96.1.19>
- Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading*, *11*, 3–32. https://doi.org/10.1207/s1532799xssr1101_2
- Verhoeven, L. (2007). Early bilingualism, language transfer, and phonological awareness. *Applied Psycholinguistics*, *28*, 425–439. <https://doi.org/10.1017/S0142716407070233>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag. Retrieved from <https://ggplot2.tidyverse.org>
- Wiley, J., Griffin, T. D., Jaeger, A. J., Jarosz, A. F., Cushen, P. J., & Thiede, K. W. (2016). Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied*, *22*, 393–405. <https://doi.org/10.1037/xap0000096>
- Yeari, M., & Van den Broek, P. (2011). A cognitive account of discourse understanding and discourse interpretation: The Landscape Model of reading. *Discourse Studies*, *13*, 635–643. <https://doi.org/10.1177/1461445611412748>

- Zabrocky, K. M., Agler, L. M. L., & Moore, D. (2009). Metacognition in Taiwan: Students' calibration of comprehension and performance. *International Journal of Psychology, 44*, 305–312. <https://doi.org/10.1080/00207590802315409>
- Zabrocky, K., & Commander, N. E. (1993). Rereading to understand: The role of text coherence and reader proficiency. *Contemporary Educational Psychology, 18*, 442–454. <https://doi.org/10.1006/ceps.1993.1033>

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Items of the Multilingual Naming Test (MINT) in Arabic, Hebrew, and English.

Appendix S2. Mean Accuracies and Contrasts for Comprehension Texts Based on a Pilot Study.

Appendix S3. Reading Comprehension Texts in Arabic, Hebrew, and English.

Appendix S4. Raven's standard progressive matrices test example.

Appendix: Accessible Summary (also publicly available at <https://oasis-database.org>)

Monitoring of Reading Comprehension Across L1, L2, and L3: Domain-General or Language-Specific? Maybe Both!

What This Research Was About and Why It Is Important

Being able to understand what we read is critical for academic and professional success. Increasingly, students in higher education globally are reading not only in their native language but in other languages as well, notably English. An important skill that facilitates comprehension is our ability to monitor our understanding—did I understand what I just read? For reading in the native language, there is ample evidence that accurate comprehension monitoring (called meta-cognitive monitoring) contributes to successful comprehension. The current study asked whether this is also true for reading comprehension in second (L2) and third L3 (languages). Further, we asked whether metacognitive skills in reading comprehension are a core skill, which can be used similarly by individual readers across their different languages, or might be used in language-specific ways. Our results showed that accurate comprehension monitoring is linked to successful comprehension across all languages of trilingual adults. We also found that comprehension monitoring was to some extent a core skill, because it was used similarly by individuals across L1 and L2. However,

monitoring was less well generalized to L3, readers' least proficient language, suggesting that readers might need to achieve some threshold of proficiency before they can fully implement their core monitoring skills. These findings could suggest that instruction programs that promote metacognitive monitoring in reading comprehension are important, not only for the native language but for additional languages used by multilinguals. Also, the findings could suggest that instruction promoting metacognitive skills in L1, might also benefit other languages, and may thus be doubly, or even triply, effective!

What the Researchers Did

- Eighty trilingual undergraduates in Israel participated in the study. They spoke L1 Arabic, L2 Hebrew, L3 English.
- Each participant read three texts in each language, answered multiple-choice questions, and rated their confidence in their responses following each question.
- These confidence ratings were used to assess comprehension monitoring. Monitoring accuracy was calculated as the discrepancy between actual performance and confidence judgments, which is termed calibration bias.
- We also measured participants' proficiency in each language, using a vocabulary test and self-report.
- We also measured calibration bias in a nonlinguistic task, to check whether it was a general skill (or language specific).

What the Researcher Found

- Better reading comprehension was related to low calibration bias (more accurate monitoring) in all three languages.
- Participants were equally accurate in monitoring their comprehension in the three languages.
- Calibration bias was not associated with proficiency in the language, in other words, comprehension monitoring was not strongly linked to an individual's proficiency in the language tested.
- Participants who were well calibrated in L1, also tended to be better calibrated in L2, and L2 calibration was moderately linked with L3 calibration. However, monitoring in L1 and L3 were not related.

Things to Consider

- The pattern of associations between reading comprehension and monitoring supported findings from previous studies examining readers' performance in

their L1 and extended them to L2 and L3, highlighting the role of metacognition in reading comprehension across languages.

- Our results mostly supported the notion of monitoring ability as a core skill that can be applied by individuals to linguistic and non-linguistic domains in a similar manner. However, at the individual level, our findings suggest that the specific recruitment of comprehension monitoring might vary across languages of different proficiency.
- Future studies should investigate the underlying skills that contribute to accurate better monitoring, and develop intervention programs that assess accurate calibration and link it to self-regulation in learning. Another promising avenue of research could be an investigation of what confidence judgments are based on when reading in the native language compared to when reading in other languages.

Materials and data: Materials and data are publicly available at <https://osf.io/EMAZ3/>

How to cite this summary: Silawi, R., Shalhoub-Awwad, Y., & Prior, A. (2020). Monitoring of reading comprehension across L1, L2, and L3: Domain-general or language-specific? Maybe both. *OASIS Summary* of Silawi, Shalhoub, & Prior (2020) in *Language Learning*. <https://oasis-database.org>

This summary has a CC BY-NC-SA license.