IJEIR
International Journal of Educational Innovation and Research ISSN: 2810-0808
Volume 1, Number 2, 2022, pp. 191-199
DOI: https://doi.org/10.31949/ijeir.v1i2.2617

The role of language in word-problem solving: a meta-analysis

Martín Flores[1]*, Mabel Urrutia[2]
[1]Universidad Central de Chile
[2]Universidad de Concepción, Chile
*Coresponding Author: fq.martin@gmail.com

ABSTRACT
The relationship between language proficiency and word-problem solving has been extensively studied in the last three decades with one main finding: language proficiency is directly proportional to word-problem solving scores. Researchers have focused on language accommodations to standardized tests to level the playing field for nonnative speakers. Interestingly, several meta-analyses on language accommodation have noted that there are no significant effects on scores. At the same time, most research has reported significance and not effect sizes, which does not allow to establish comparisons between studies. Therefore, the purpose of this meta-analyses was to study the effect size of studies conducted in the US and the rest of the world to provide a new population effect size and identify possible moderator variables that have not been considered before. The main findings were that the differences in scores between native speakers and nonnative speakers (g=.360) were less pronounced than previously reported (g=.604). The participants' age and the language of instruction and testing moderated a small percentage of the effect size, which leads to the need to include specific information about the participants to provide a clearer picture of this relationship and eventually provide language accommodations that have significant effects on scores.

ARTICLE HISTORY
Received 2022-05-03
Accepted 2022-07-13

KEYWORDS
Word-Problem Solving
Language Proficiency
Bilingualism
Word Problems

INTRODUCTION

Although researchers have posited the notion that language and mathematical knowledge are two distinct independent cognitive abilities (Guthormsen et al., 2016), the relationship between language proficiency and mathematical achievement has been a topic of study for more than 30 years. Educators, researchers, and policy makers have paid special attention to this interaction because of the need to fairly assess large groups of students (Ockey, 2007). The large corpus of studies conducted so far points towards one conclusion in particular: the lower the language proficiency, the lower the mathematical achievement, especially in word-problem solving (Abedi & Lord, 2001; Kieffer et al., 2009; Kieffer et al., 2012; Pennock-Roman & Rivera, 2011; Rios et al., 2020). When testing native speakers of English and English learners on mathematical achievement, the scores are quite similar in computational problems, such as basic addition and subtraction (Bernardo, 2002; Martin & Fuchs, 2019; Powell et al., 2020). The differences arise when these two groups of students are tested on word problem solving. According to Driver and Powell (2017, p.41),

> A word problem is a mathematics calculation embedded within sentences (…) To solve word problems, students use text, typically presented in English, to

identify missing information, make a plan to solve the problem, and perform one or more calculations to get the solution.

Ideally, the use of words should not affect the construct of mathematical knowledge that is being measured. Nevertheless, research has repeatedly shown that language of testing has a detrimental effect on nonnative speakers, especially if they are not proficient in the language (Swanson et al., 2019). Interestingly, most studies report that they do not assess language proficiency independently with a standardized language test. They report the level of English of the participants based on the information provided by head teachers or the school, mostly stating that this piece of information in particular is missing or not easily available (Rios et al., 2020).

On the relationship between language and word problem solving, the Cognitive Load Theory (Campbell et al., 2007) has posited the idea that linguistically and mathematically complex word problems should yield the poorest performance due to working memory overload. Barbu and Beal (2010) have tested this notion by analyzing how a group of English learners performed when word problems differed in language and computational complexity. Unsurprisingly, easy computations, such as simple addition and multiplication, with easy to understand texts were correctly solved 90% of the total times. Whereas difficult computations, such as multi-digit multiplication and division along with texts difficult to understand, were solved correctly approximately 50% of the total times. In general, participants achieved the lowest scores when the text was complex and the operations difficult, proving the role of cognitive load to successfully solve word problems.

This relationship does not only affect foreign speakers of a language, but also native speakers. Rodríguez and Domínguez (2016) focused on identifying the difficulties Spaniard students face when solving word problems and proposing a teaching strategy to soften the effects of these difficulties on performance. After observing a class of 3rd graders, they identified four main difficulties: oral expression and reading comprehension, identification of the key elements of the problem, question identification, and proper articulation of the answer. After applying a reading comprehension intervention and contrasting the results of the experimental group with a control group, the number of difficulties the students faced was reduced. Therefore, this intervention in particular might offer a possible means to improve word problem solving.

Considering that in the US schools are measured by the results their students achieve on standardized tests (Ockey, 2007), bridging the gap between native and nonnative speakers of English has been the focus of a considerable body of studies (Ríos et al., in press). Interventions focused on strategies to identify relevant information and solve word problems (Swanson et al., 2019), improve reading comprehension (Driver and Powell, 2017) and accommodations on tests (Abedi and Lord, 2001) have shown promising results. Nevertheless, meta-analyses on accommodations have shown little (Kieffer et al., 2009) to no (Rios et al., 2020) significant effect on improving scores.

The effect of language proficiency and word problems has been mostly studied in the US, where currently 1 in 10 students speak English as a second language (Powell et al., 2020). It is not coincidental that most studies are conducted there because there are several policies that state schools need to follow in order to ensure that English learners (ELs) are learning the content even when their L1 differs from English (Kieffer et al., 2009). Kieffer et al. (2009, p. 1170), after conducting a meta-analysis on the efficiency of language accommodations, reported that "there is indeed a substantial link between students' English language proficiency and their performance on tests of math, science, and social studies." Namely, ELs achieve significantly lower scores than their peers. Moreover, Kieffer et al. (2009, p. 1170) note that "several correlational studies have found that assessments and individual test items that have more linguistic complexity yield larger performance gaps between ELLs and non-ELLs." Therefore, language proficiency might be unfairly assessing the knowledge ELs may possess in school due to construct-irrelevant variance (Banks et al., 2016).

Although the relationship between language proficiency and word problem solving has been studied extensively, early research on this topic focused mostly on two dimensions of the problem that need to be revisited. First, the focus was on reporting significant differences between ELs and native speakers' scores and not on effect size, something that might lead to misleading conclusions (Field & Gillet, 2010). Second, the studies were mostly conducted in the US, with English as the language of instruction and testing, which leaves aside research conducted in the rest of the world with other instruction languages, which might create an incomplete picture of the problem.

Since the relationship between word problem solving and language has shown solid results in the last decades, research has focused on the different accommodations that can be done to tests to make them fair for ELs, without affecting the construct they are measuring. To date, two narrative reviews (Abedi & Lord., 2004; Sireci, et al., 2003), two systematic reviews (Acosta et al., 2019; Baker et al., 2016) and several meta-analyses (Kieffer et al., 2009; Kieffer et al., 2012; Pennock-Roman & Rivera, 2011; Rios et al., 2020) have focused on summarizing and analyzing the effects of language accommodation in standardized tests, particularly on science, math and English language arts. These reviews and analyses have considered the evidence of research mostly conducted in the US on the relation of language proficiency and word problem solving in English learners (ELs) as a fact, and have focused on ways to narrow the gap. That goal in itself is important considering that "By 2030, ELs are projected to represent 40% of all school-age students" (Powell et al., 2020, p.122).

Kieffer et al. (2009) comment on the results of a meta-analysis regarding the difference in scores between ELs and native English speakers in the US. In relation to math, they point out that the mean effect size in non-national assessments is g = .604, whereas in national assessments the mean size was d = .831 for 4^{th} graders and d = 1.006 for 8^{th} graders. They propose the notion that other variables might explain this difference, such as social class and the quality of the schools. Nevertheless, they do report an intermediate effect size for the difference in score in favor of the native speakers under experimental conditions.

In relation to the meta-analyses on language accommodation, surprisingly, researchers have found limited effectiveness to improve scores on ELs (Kieffer et al., 2009) or no effectiveness "statistically different from zero" (Rios et al., 2020), which clearly indicates that the measures taken to level the field for ELs have not worked. Even more, previous research has also suggested that standardized math tests are fair when assessing arithmetical skills because the minute biases found may have been due to chance or methodological flaws with the sample (Ockey, 2007). Furthermore, recent research has found contradictory results related to the role of reading comprehension when successfully solving word problems (Pavón & Cabezuelo, 2019; Trakulphadetkrai et al., 2020) with subjects receiving instruction in Spain and the UK, respectively. These conclusions suggest that the foundational findings might have missed possible key elements.

Finally, Rios et al. (2020) have recently pointed out that focusing on particular populations may improve the knowledge that we have regarding the effectiveness of language accommodations for ELs because there is need "to recognize EL heterogeneity and begin to study specific EL subpopulations (…) by collecting large sample sizes that account for idiosyncrasies". This highlights the need to have more contextualized knowledge about the effects that language of instruction may have on students that have a different L1.

Therefore, the purpose of this meta-analysis is to survey the differences in scores in word problems between native speakers and non-native speakers, regardless of the language of instruction, and to evaluate their effect size.

The purpose of this meta-analysis was to determine the average effect size of studies that have focused on documenting the difference in scores in word problem solving between native speakers of the language of testing and nonnative speakers, trying to identify moderator variables that should be considered

in interventions and accommodations. Consequently, studies that have used a different language than English to test differences were also included.

METHODS

Search strategy

Two main search strategies for suitable studies were followed. First, two databases were searched: Google Scholar and ERIC. The search terms used were (math OR mathematical OR mathematics OR arithmetic) AND "word problem" AND (bilingualism OR bilingual OR language). The results were limited to studies between 2001 and 2020. Both the search terms and date range are broad to include as many studies as possible, particularly from outside the US. The Google Scholar search yielded 2630 results, whereas the EIRC search yielded 218 results. The second strategy was to examine the references of the studies that met the eligibility criteria (illustrated below).

Eligibility criteria

Studies were included if, first, they 1) quantitatively measured the performances of second language learners in contrast to native speakers or 2) if they compared second language learners' performance in their L1 and L2. Second, the studies were not conducted by the government since the sample size would immediately outweigh the effect size of other studies. If the studies were conducted in the US, they had to be published after 2006 to include studies that were not considered by Kieffer et al. (2009). Hence, the studies included were quasi-experimental in nature.

Studies were excluded if they: 1) did not provide separate scores when they had mixed samples (evaluating word problems for second language learners and SEN students, for example); did not provide number of participants, mean scores and standard deviations to calculate effect size; and 3) were found methodologically flawed.

From the 2848 studies found, and after removing the duplicates and studies that were not peer-reviewed or part of a dissertation, 35 were fully read. After checking for the eligibility criteria, 10 composed the final sample. These studies were written by 10 main authors, published between 2001 and 2020 in peer-reviewed journals. The final sample was composed of 26 effect sizes for 2,528 native speakers and 1,786 second language learners. Most studies reported focused on testing either a language different from English or testing learners on their native language and English (69.2% of effect sizes). In addition, 69.03% of the participants were part of the English as language of instruction and testing studies. Languages other than English included Filipino (3 studies), German and Turkish (1 study) and Spanish (1 study). The study that focused on German and Turkish was the only study that did not measure English as well.

Analysis procedure

In order to compute the population effect size, EpiGear's MetaXL package for Excel was used. Since the samples of the individual studies were small, Hedges's g was computed (Kieffer et al., 2009).

The procedures followed to conduct the meta-analysis were the ones proposed by Field and Gillet (2010) and are described as follows. After tabulating the samples and effect sizes for each study, a random-effects method was run on MetaXL. This method was chosen because it allows to "generalize beyond the studies included in the meta-analysis" (Field & Gillet, 2010, p. 673), which serves a higher purpose when trying to level the field for non-native speakers. After calculating the average effect size for all studies, moderator variable analysis was conducted to identify the influence of moderator variables on the effect sizes. This was done using Stata SE version 13. In this case, two moderator variables were considered: age and language of testing. Age was composed of two categories: ages from 6 to 12 and ages from 13 to 18. Language used for testing was categorized between English and other. If a study tested the participants in English and their native language, this was also considered part of the other language category. Level of

proficiency was not considered as a possible moderator since almost no studies provided this information. Finally, publication bias was addressed by means of a funnel plot and heterogeneity test. First, the purpose of the funnel plot was to identify if the effect sizes were evenly distributed across the median, which would indicate that no publication bias was found. At the same time, a heterogeneity test was conducted to identify "the proportion of the variation in effect size estimates due to heterogeneity as opposed to chance" (Rios et al., 2020).

RESULTS AND DISCUSSION

The following stem-and-leaf plot summarizes the frequencies of the effect sizes of the different studies included after being calculated.

Table 1. Stem-and-leaf plot of all effect sizes (gs)

Stem	Leaf
-.4	9
-.3	1
-.1	4
.0	7
.1	3, 5, 6
.2	4
.3	0, 3, 4, 6
.4	1
.5	3, 7
.6	0, 1, 3, 8
.7	5, 6, 7
.8	4
.9	
1.	01, 5
2.	7

As can be seen, the smallest effect size was g = -.49 and the largest was g = 2.7. Most effect sizes were in the .3 to .7 range.

Average effect size and heterogeneity test

The computed average effect size was g = .360 (95% CI: .176, .543). I^2 was 85.02%, which suggests considerable heterogeneity, cementing the need for a moderator analysis to identify if other variables might explain the effect sizes.

Moderator Analysis

The moderator analysis indicates that age significantly moderated 3.5% of the effect sizes (p = .002). In relation to language of instruction and testing, the moderator analysis indicates that language of instruction and testing significantly moderated 4.7% of the effect sizes (p < .001).

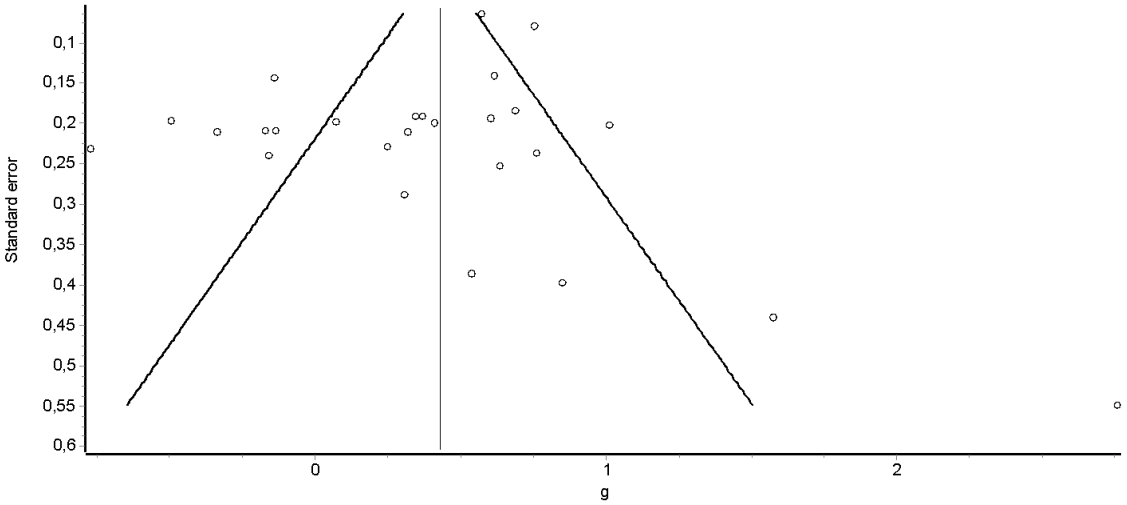
In conjunction, when the participants are in the 6 to 12 age range and the language of instruction and testing is English, 5.7% of the effect sizes are moderated (p=.001). If the language of instruction and testing is other, only 2.7% of the effect sizes are significantly moderated (p=.05).

When the participants are in the 13 to 18 age range, language of instruction and testing moderates a lower percentage of effect sizes (0.5% for English and approximately 0% for other), with no value reaching significance (p=.66 and p=.991, respectively).

Publication bias

The following funnel plot illustrates the distribution of the effect sizes along the median.

Figure 1. Funnel plot



This funnel plot illustrates that more than half effect sizes were distributed symmetrically. The 12 effect sizes outside the triangular region can be related to the considerable heterogeneity reported before.

Discussion

The reported population effect size of this meta-analysis was g = .360, an intermediate effect size. According to Coe (2002), this effect size would mean that 62% of the language learners would be below the average score of the native speakers. This result contrasts markedly when compared to Kieffer et al.'s (2009) reported population effect size of g = .604. This difference in population effect sizes might be explained by the studies that were included. Mainly, their study did not include languages of testing other than English and focused on studies conducted in the US. Nonetheless, when language of instruction and mother tongue coincide, scores would be higher, although not as pronounced as reported before.

This meta-analysis illustrates that most of the studies included have measured the relationship between word problem solving and language in language of instruction and testing other than English (17 out of 26 effect sizes). Nevertheless, only one study reported effect sizes where English was not the language of instruction or testing. This clearly shows how most research has focused on English. This might suggest that language of instruction and testing could be a major moderator of the population effect size, which would be congruent to the conclusions reached by previous research. However, the moderator analysis showed that language of instruction and testing moderates a small proportion of the population effect size. Moreover, the same can be said about the age range of the participants. Both results might be explained by the characteristics of the sample. First, English, as the only means of instruction and testing, was part of approximately a third of the effect sizes reported. Furthermore, the sample size for the 13-18 age range was a quarter in comparison to all the participants, focusing on two studies only. These differences might have influenced the results of this meta-analysis, possibly underrepresenting the role of language of instruction and testing, and age range. This should be considered in further research as a mediating effect.

In relation to publication bias and heterogeneity, there was considerable heterogeneity in this meta-analysis (I^2 = 85.02%), which is expressed as well in the number of studies that are outside the expected ranges in the funnel plot (almost 50%). As mentioned in the previous paragraph, age range and language of instruction and testing only account for a small percentage of the effect sizes (5.7%). Heterogeneity can be explained by the different research designs of the studies (Higgins & Thompson, 2002) and by the different information reported by the studies that did not allow to provide more moderators (Rios et al., 2020).

CONCLUSION

The results in this meta-analysis corroborate the disaggregated findings of previous research: participants achieve higher scores when their language of instruction, testing and mother tongue are the same. When complemented with the low efficiency of accommodation strategies to improve the scores of nonnative speakers (Kieffer et al., 2009; Ríos et al., 2020), these results point out the need to continue researching, not only accommodation to level the playing field, but also nonnative speakers' scores on word problems on languages different from English to describe the extent of the differences in scores.

Naturally, due to the nature of the inclusion criteria, this meta-analysis could not encompass a balanced description of age range and language of instruction and testing. Having a more balanced sample size might provide a better picture of the interaction between language on word problems. At the same time, this meta-analysis illustrates that age and language of instruction and testing only account for less than 6% of the population effect size. Therefore, it is important to explore other variables that might explain why native speakers achieve higher scores in math. Kieffer et al. (2009) posited that notion that sociodemographic factors could explain the difference in effect sizes, particularly in controlled studies and national tests. As such, future research should also report these variables to be included in future meta-analyses.

Unfortunately, promising studies could not be included because key information regarding samples, means and standard deviations were missing. Attention must be paid to these data, even when results are not significant, to broaden the perspectives related to this phenomenon.

Finally, future research should pay attention to language proficiency as a variable that might explain not only the difference in scores but also why accommodations are not as efficient as needed. Therefore, researchers should measure language proficiency independently by means of a standardized test to clearly identify the L2 proficiency of nonnative speakers. Additionally, adult learners should be included, considering that some universities might require to test mathematical knowledge in a foreign language.

REFERENCES

Abedi, J. & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education, 14*(3), 219-234. http://dx.doi.org/10.1207/S15324818AME1403_

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1-28.

Acosta, B. D., Rivera, C., & Willner, L. S. (2008). *Best practices in state assessment policies for accommodating English language learners: A delphi study*. Arlington: Center for Equity and Excellence in Education, the George Washington University

Baker, D., Basaraba, D. & Polanco, P. (2016). Connecting the Present to the Past: Furthering the Research on Bilingual Education and Bilingualism. *Review of Research in Education, 40*(1), 821-883. https://doi.org/10.3102/0091732X16660691

Banks, K., Jeddeeni, A. & Walker, C. (2016). Assessing the Effect of Language Demand in Bundles of Math Word Problems. *International Journal of Testing, 16*(4) 269-287. https://doi.org/10.1080/15305058.2015.1113972

Barendregt, J. & Doi, S. (2016). MetaXL User Guide (Version 5.3). EpiGear International Pty Ltd.

Bautista, D., Mitchelmore, M. & Mulligan, J. (2009) Factors influencing Filipino children's solutions to addition and subtraction word problems. *Educational Psychology, 29*(6), 729-745.

Bernardo, A. (2002). Language and Mathematical Problem Solving Among Bilinguals, *The Journal of Psychology, 136*(3), 283-297. http://dx.doi.org/10.1080/00223980209604156

Bernardo, A. (2005) Language and Modeling Word Problems in Mathematics Among Bilinguals, *The Journal of Psychology: Interdisciplinary and Applied, 139*(5), 413-425. http://dx.doi.org/10.3200/JRLP.139.5.413-425

Campbell, A., Adams, V., & Davis, G. (2007). Cognitive demands and second language learners: A framework for analyzing mathematics instructional contexts. *Mathematical Thinking and Learning, 9*(1), 3-30.

Coe, R. (2002). *It's the Effect Size, Stupid. What Effect Size Is and Why It Is Important* [Paper]. The British Educational Research Association Annual Conference, Exeter, U.K. http://www.leeds.ac.uk/educol/documents/00002182.htm

Driver, M. & Powell, S. (2017). Culturally and Linguistically Responsive Schema Intervention: Improving Word Problem Solving for English Language Learners With Mathematics Difficulty. *Learning Disability Quarterly, 40*(1), 41-53.

Field, A. & Gillet, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665-694. https://doi.org/10.1348/000711010X502733

Guthormsen, A., Fisher, K., Bassok, M., Osterhout, L., DeWolf, M., & Holyoak, K. (2016). Conceptual Integration of Arithmetic Operations With Real-World Knowledge: Evidence From Event-Related Potentials. *Cognitive Science*, *40*(3), 723-757. http://dx.doi.org/10.1111/cogs.12238

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558

Kempert, S., Saalbach, H. & Hardy, I. (2011). Cognitive Benefits and Costs of Bilingualism in Elementary School Students: The Case of Mathematical Word Problems. *Journal of Educational Psychology, 103*(3), 547-561.

Kieffer, M., Lesaux, N., Rivera, M. & Francis, D. (2009). Accommodations for English Language Learners Taking Large-Scale Assessments: A Meta-Analysis on Effectiveness and Validity. *Review of Educational Research, 79*(3), 1168-1201. https://doi.org/10.3102/0034654309332490

Kieffer, M., Rivera, M., Francis, D.J. (2012). *Research-based recommendations for the use of accommodations in large-scale assessments: 2012 update*. Practical Guidelines for the Education of English Language Learners. Book 4. New York: Center on Instruction.

Lager, C. (2006). Types of Mathematics-Language Reading Interactions that Unnecessarily Hinder Algebra Learning and Assessment. *Reading Psychology, 27*(2-3), 165-204.

Martin, B. & Fuchs, L. (2019). The Mathematical Performance of At-Risk First Graders as a Function of Limited English Proficiency Status. *Learning Disability Quarterly, 42*(4) 244-251. https://doi.org/10.1177/0731948719827489

Ockey, G. (2007). Investigating the Validity of Math Word Problems for English Language Learners with DIF. *Language Assessment Quarterly, 4*(2), 149-164

Pavón, V. & Cabezuelo, R. (2019). Analysing mathematical word problem solving with secondary education CLIL students: A pilot study. *Latin American Journal Of Content & Language Integrated Learning, 12*(1), 18-45. https://doi.org/10.5294/laclil.2019.12.1.2

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*(3), 10-28.

Powell, S., Berry, K. & Tran, L. (2020). Performance Differences on a Measure of Mathematics Vocabulary for English Learners and Non-English Learners with and without Mathematics Difficulty. *Reading and Writing Quarterly, 36*(2) 124-141. https://doi.org/10.1080/10573569.2019.1677538

Rios, J., Ihlenfeldt, S. & Chavez, C. (2020). Are Accommodations for English Learners on State Accountability Assessments Evidence-Based? A Multistudy Systematic Review and Meta-Analysis. *Educational Measurement Issues and Practices.* https://doi.org/10.1111/emip.12337

Sireci, S., Li, S., & Scarpati, S. (2003). The effect of test accommodation on test performance: A review of the literature (Research Report 495). Amherst: University of Massachusetts School of Education, Center for Educational Assessment.

StataCorp. 2013. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.

Swanson, H., Kong, J., Moran, A. & Orosco, M. (2019). Paraphrasing Interventions and Problem-Solving Accuracy: Do Generative Procedures Help English Language Learners with Math Difficulties? *Learning Disabilities Research and Practice, 34*(2), 68-84. https://doi.org/10.1111/ldrp.12194

Trakulphadetkrai, N., Courtney, C., Clenton, J., Treffers-Daller, J., & Tsakalaki, A. (2020). The contribution of general language ability, reading comprehension and working memory to mathematics achievement among children with English as additional language (EAL): an exploratory study. *International Journal of Bilingual Education and Bilingualism, 23*(4), 473-487. https://doi.org/10.1080/13670050.2017.1373742