What Works Clearinghouse

ies INSTITUTE OF EDUCATION SCIENCES

June 2012

WWC Review of the Report "Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets"[1,2]

The findings from this review do not reflect the full body of research evidence on the effectiveness of *linguistic modification* of math test items.

What is this study about?

The study is a randomized controlled trial in which seventh- and eighth-grade students were randomly assigned to complete a set of 25 math questions delivered with either standard language or language that had undergone *linguistic modification* by the research team.

The purpose of the study was to assess the effects of using *linguistic modification* as a way of removing language barriers for English language learners and non-English language learners struggling with reading.

Nearly 3,000 students from 13 middle schools in five school districts in California were randomly assigned to complete traditional math assessments or math assessments that had undergone *linguistic modification*. Researchers then examined the results for three subgroups of students: Spanish-speaking English language learners (EL), non-English language learners who were not proficient in English language arts (NEP), and non-English language learners who were proficient in English language arts (EP). Comparisons were made between students who took the modified test and those who took the non-modified test.

WWC Rating

The research described in this report meets WWC evidence standards without reservations

Strengths: This study is a well-implemented randomized controlled trial.

Features of *Linguistic Modification*

The complexity of language used in test items may interfere with students' abilities to demonstrate understanding of content, especially when students are struggling with English. *Linguistic modification* is a test accommodation strategy aimed at removing language barriers. This strategy requires the creation of carefully constructed test items that are accessible to all students, regardless of language background, while still maintaining the integrity of the content being tested. This study focuses on the *linguistic modification* of math content that is typically presented on standardized math achievement tests.

What did the study find?

The study found a positive effect on math scores for students struggling with English who completed the *linguistic modification* item set relative to similar students who did not. The estimated six percentage-point gain on math achievement is statistically significant. The study found neither statistically significant nor substantively important differences for EP students who took the modified test, relative to those who did not.

Appendix A: Study details

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C.-W. (2010). *Accommodations for English language learner students: The effect of linguistic modification of math test item sets* (NCEE 2009-4079). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Setting

The study was conducted in 13 middle schools located in five school districts in California serving high percentages of students whose native language is Spanish. Nine schools were urban, three were rural, and one was located in a small city.

Study sample[3]

The original randomized sample included 6,342 students (3,171 each in the *linguistic modification* and traditional item set groups) in grades 7 and 8. Sample attrition occurred because of lack of completed booklets; reasons for lack of completed booklets included:

- an overestimation of the number of eligible students,
- student absenteeism, and
- 16 teachers declining to participate in the study.

After the remaining 5,380 students completed and returned the math assessment (2,687 in the *linguistic modification* group and 2,693 in the traditional item set group), researchers categorized the students into three subgroups based on prior state test score data:

- Spanish-speaking English language learners (EL),
- non-English language learners who were not proficient in English language arts (NEP), and
- non-English language learners who were proficient in English language arts (EP).

Students in the EL condition were excluded from the analysis if they did not speak Spanish as a first language (a loss of 160 students in the *linguistic modification* group and 146 students in the traditional item set group) or if they lacked scores from a qualifying California English Language Development Test (50 students in the *linguistic modification* group and 67 students in the traditional item set group). Non-EL students were excluded from the analysis if they lacked California Standards test information and thus could not be categorized into the NEP or EP subgroups (170 students each in the *linguistic modification* group and traditional item set group). The final analysis sample included 4,617 students, which included 2,307 in the *linguistic modification* item set group (608 EL, 804 NEP, 895 EP) and 2,310 in the traditional item set group (606 EL, 821 NEP, 883 EP). For the purposes of the WWC rating of effectiveness, the analysis sample focuses on the 1,214 EL and 1,625 NEP students: 1,412 in the *linguistic modification group* and 1,427 in the traditional item set group.

Intervention group[4]

Students in the *linguistic modification* item set group were administered a set of items that had been revised: the language load was removed to make the items more accessible and easier to comprehend. To create the modified item set, the research team convened a working group of experts in mathematics, linguistics, measurement, curriculum and instruction, and English language learning. The team followed an eight-step process related to item selection, development, and administration to ensure that the traditional and *linguistic modification* item sets measured the same constructs.

Comparison group

Students in the comparison condition received the traditional, unmodified math items.

Outcomes and measurement

Because of the nature of this study, the intervention and comparison groups received different versions of a 25-item math assessment that measured the same item constructs. For a more detailed description of these outcome measures, see Appendix B.

Support for implementation

No training was necessary.

Reason for review

This study was identified for review by the WWC because it is an Institute of Education Sciences (IES)-funded study conducted by 2006-11 Regional Education Laboratory West administered by WestEd.

Appendix B: Outcome measures for each domain

Mathematics achievement	
Study-developed math assessment	The *linguistic modification* and traditional math item sets included 25 matched pairs of items focusing on two content strands often covered in standardized math assessments: (1) measurement, and (2) number sense/operations. The study research team conducted a multi-layered procedure for designing the assessments that included convening a working group consisting of the study team and experts in mathematics, linguistics, measurement, curriculum and instruction, and EL student populations. Specifically, the team followed an eight-step procedure related to item selection, development, and the administration process: 1. initially collecting 256 items from the grade 8 National Assessment of Education Progress and the grade 7 California state test, 2. retaining 115 items with sufficient language to modify linguistically (those focused on measurement and number sense/operations), 3. retaining 81 items that covered a diverse pool of items aligned to state standards, 4. retaining 51 items to which specific *linguistic modification* strategies could be applied (those that included words or grammatical structures typically unfamiliar to EL students), 5. selecting 42 matched pairs of linguistically modified and traditional items endorsed by content specialists, 6. conducting a cognitive interview process with nine students to determine the range of complexity and range of content assessed that reduced the number to 63 items, 7. selecting 30 matched pairs of items based on the student cognitive interviews, and 8. conducting a pilot test that narrowed the set to 25 matched pairs of items. The experts reviewed the final contents and verified that the constructs intended to be assessed had not been changed substantively in the *linguistic modification* item set.[5] Internal consistency reliability for the original item set ranged from 0.61–0.79 across the three student subgroups (EL, NEP, and EP). Internal consistency for the linguistically modified item set ranged from 0.68–0.78 across the three subgroups.

Appendix C: Study findings for the mathematics achievement domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Mathematics achievement								
Study-developed math assessment (raw score)	Grades 7 and 8 (EL and NEP students)	2,839 students	10.07 (4.06)	9.49 (3.83)	0.58	0.15	+6	0.00
Domain average for mathematics achievement						**0.15**	**+6**	**Statistically significant**

Table Notes: Positive results for mean difference, effect size, and improvement index favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average student's outcome that can be expected if the student is given the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is given the intervention. The statistical significance of the study's domain average was determined by the WWC; the study is characterized as having a statistically significant positive effect because a univariate statistical test is reported for the single outcome measure and the effect is positive and statistically significant. EL = Spanish-speaking English language learners; NEP = non-English language learners who were not proficient in English language arts.

Study Notes: No corrections for clustering or multiple comparisons were needed. The reported effect sizes and *p*-values were computed by the WWC, using the pooled standard deviations and *p*-values for the subgroups included above (EL and NEP students). This study was not designed to assess the effectiveness of a program but rather to assess the effect of modifying a mathematics assessment to remove language barriers for students struggling with English. As such, the "intervention group" included students who were assigned to the linguistically modified item set, and the "comparison group" included students who were assigned to the traditional item set.

Appendix D: Subgroup findings for the mathematics achievement domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Mathematics achievement								
Study-developed math assessment (raw score)	Grades 7 and 8 (EL students)	1,214 students	9.16 (3.91)	8.40 (3.52)	0.76	0.20	+8	0.00
Study-developed math assessment (raw score)	Grades 7 and 8 (NEP students)	1,625 students	10.69 (4.05)	10.23 (3.85)	0.46	0.12	+5	0.01
Study-developed math assessment (raw score)	Grades 7 and 8 (EP students)	1,778 students	15.63 (4.58)	15.59 (4.66)	0.04	0.01	0	0.86

Table Notes: Positive results for mean difference, effect size, and improvement index favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average student's outcome that can be expected if the student is given the intervention. The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is given the intervention. EL = Spanish-speaking English language learners; NEP = non-English language learners who were not proficient in English language arts; EP = non-English language learners who were proficient in English language arts.

Study Notes: No corrections for clustering or multiple comparisons were needed. The reported effect sizes and *p*-values were computed by the WWC. The WWC calculations pooled the standard deviations of the *linguistic modification* item set and the traditional item set for each subsample individually. This study was not designed to assess the effectiveness of a program but rather to assess the effect of modifying a mathematics assessment to remove language barriers for students struggling with English. As such, the "intervention group" included students who were assigned to the linguistically modified item set, and the "comparison group" included students who were assigned to the traditional item set.

Endnotes

[1] Single study reviews examine evidence published in a study (supplemented, if necessary, by information obtained directly from the author[s]) to assess whether its design meets WWC evidence standards. The review reports the WWC's assessment of whether the study meets WWC evidence standards and summarizes the study findings following WWC conventions for reporting evidence on effectiveness. The WWC rating applies only to the summarized results, and not necessarily to all results presented in the study. This study was reviewed using the Single Study Review Protocol, version 2.0.

[2] Absence of conflict of interest: The Regional Educational Labs were provided technical assistance by Mathematica Policy Research, which also operates the WWC. Because Mathematica operates the WWC, this study was reviewed by staff from subcontractor organizations.

[3] The randomized sample included 6,342 students: 3,171 students who completed the *linguistic modification* item set and 3,171 students who completed the original item set. The final research sample included 4,617 students: 2,307 who completed the *linguistic modification* item set and 2,310 who completed the original item set. Overall attrition is 27%. There is no information available on attrition rates for the three student subgroups, but the final sample sizes for the two conditions are similar within each subgroup, which suggests that differential attrition rates within each subgroup are low.

[4] This study was not designed to assess the effectiveness of a program, but rather to assess the effect of modifying a mathematics assessment to remove language barriers for students struggling with English. As such, the "intervention group" included students who were assigned to the linguistically modified item set, and the "comparison group" included students who were assigned to the traditional item set.

[5] The outcome results were reported in four ways: the raw score, 1-PL model, 2-PL model, and 3-PL model. The #-PL models are item response theory (IRT) models, allowing each item to have its own difficulty level, using a variety of assumptions. The authors included these three IRT models so that strategies for estimating students' underlying abilities through a test would be consistent with states' practices. The raw score is used for WWC rating purposes.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2012, June). *WWC review of the report: Accommodations for English language learner students: The effect of linguistic modification of math test item sets*. Retrieved from http://whatworks.ed.gov.

Glossary of Terms

Attrition Attrition occurs when an outcome variable is not available for all participants initially assigned to the intervention and comparison groups. The WWC considers the total attrition rate and the difference in attrition rates across groups within a study.

Clustering adjustment If intervention assignment is made at a cluster level and the analysis is conducted at the student level, the WWC will adjust the statistical significance to account for this mismatch, if necessary.

Confounding factor A confounding factor is a component of a study that is completely aligned with one of the study conditions, making it impossible to separate how much of the observed effect was due to the intervention and how much was due to the factor.

Design The design of a study is the method by which intervention and comparison groups were assigned.

Domain A domain is a group of closely related outcomes.

Effect size The effect size is a measure of the magnitude of an effect. The WWC uses a standardized measure to facilitate comparisons across studies and outcomes.

Eligibility A study is eligible for review and inclusion in this report if it falls within the scope of the review protocol and uses either an experimental or matched comparison group design.

Equivalence A demonstration that the analysis sample groups are similar on observed characteristics defined in the review area protocol.

Improvement index Along a percentile distribution of students, the improvement index represents the gain or loss of the average student due to the intervention. As the average student starts at the 50th percentile, the measure ranges from –50 to +50.

Multiple comparison adjustment When a study includes multiple outcomes or comparison groups, the WWC will adjust the statistical significance to account for the multiple comparisons, if necessary.

Quasi-experimental design (QED) A quasi-experimental design (QED) is a research design in which subjects are assigned to intervention and comparison groups through a process that is not random.

Randomized controlled trial (RCT) A randomized controlled trial (RCT) is an experiment in which investigators randomly assign eligible participants into intervention and comparison groups.

Single-case design (SCD) A research approach in which an outcome variable is measured repeatedly within and across different conditions that are defined by the presence or absence of an intervention.

Standard deviation The standard deviation of a measure shows how much variation exists across observations in the sample. A low standard deviation indicates that the observations in the sample tend to be very close to the mean; a high standard deviation indicates that the observations in the sample tend to be spread out over a large range of values.

Statistical significance Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than 5% ($p < 0.05$).

Substantively important A substantively important finding is one that has an effect size of 0.25 or greater, regardless of statistical significance.

Please see the WWC Procedures and Standards Handbook (version 2.1) for additional details.