Early Childhood Research Quarterly 57 (2021) 89–101

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

journal homepage: www.elsevier.com/locate/ecresq

Are translated mathematics items a valid accommodation for dual language learners? Evidence from ECLS-K

John Marc Goodrich*, Natalie A. Koziol, HyeonJin Yoon

University of Nebraska-Lincoln, Lincoln, Nebraska

ARTICLE INFO

Article history:
Received 13 May 2020
Revised 26 May 2021
Accepted 1 June 2021
Available online 8 July 2021

Keywords:
Dual language learners
Differential item functioning
Mathematics
Regression discontinuity design

ABSTRACT

When measuring academic skills among students whose primary language is not English, standardized assessments are often provided in languages other than English. The degree to which alternate-language test translations yield unbiased, equitable assessment must be evaluated; however, traditional methods of investigating measurement equivalence are susceptible to confounding group differences. The primary purposes of this study were to investigate differential item functioning (DIF) and item bias across Spanish and English forms of an assessment of early mathematics skills. Secondary purposes were to investigate the presence of selection bias and demonstrate a novel approach for investigating DIF that uses a regression discontinuity design framework to control for selection bias. Data were drawn from 1,750 Spanish-speaking Kindergarteners participating in the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999, who were administered either the Spanish or English version of the mathematics assessment based on their performance on an English language screening measure. Evidence of selection bias—differences between groups in SES, age, approaches to learning, self-control, social interaction, country of birth, childcare, household composition and number in the home, books in the home, and parent involvement—highlighted limitations of a traditional approach for investigating DIF that only controlled for ability. When controlling for selection bias, only 11% of items displayed DIF, and subsequent examination of item content did not suggest item bias. Results provide evidence that the Spanish translation of the ECLS-K mathematics assessment is an equitable and unbiased assessment accommodation for young dual language learners.

Introduction

Despite evidence of significant progress in the education of dual language learners (DLLs), significant achievement gaps remain in mathematics between monolingual children and DLLs (Kieffer & Thompson, 2018). Prior research indicates that achievement gaps begin early in life and grow throughout the elementary years. Specifically, Kieffer (2008) reported that DLLs who enter kindergarten with typical levels of English proficiency have developmental trajectories of academic skills that do not diverge from those of monolingual children. In contrast, DLLs who enter kindergarten with limited English proficiency have significantly slower development of academic skills than do monolingual children, resulting in large achievement gaps by the time children reach late elementary school. Additionally, evidence indicates that early mathematics skills are strongly associated with high school mathematics achievement, even after controlling for reading skills, cognitive ability, and socioeconomic status (e.g., Watts, Duncan, Siegler & Davis-Kean, 2014). Consequently, it is critical that researchers identify DLLs who are at risk for mathematics learning difficulties early in life so that supports can be provided to those children to prevent them from falling behind their monolingual peers.

One issue that is central to accurate identification of children at risk for mathematics learning disabilities is the use of achievement tests that are supported by strong reliability and validity evidence. Extant evidence indicates that assessing general cognitive ability does not improve the ability to distinguish presence of or risk for learning disabilities, and current best practice is to rely on direct assessment of academic achievement (e.g., norm- or criterion-referenced mathematics assessments), alongside evaluations of student responsiveness to instruction, to identify potential learning disabilities (Fletcher & Miciak, 2017; Miciak, Taylor, Denton & Fletcher, 2015). However, assessment of mathematics achievement among DLLs is complicated by children's varied English proficiency (Abedi, 2002; Solano-Flores & Li, 2008), creating issues of equity and potential bias when assessing DLLs with tests normed for use

* Correspondence author.
E-mail addresses: jgoodrich4@unl.edu, marcgoodrich5@gmail.com (J.M. Goodrich).

https://doi.org/10.1016/j.ecresq.2021.06.001
0885-2006/

with monolingual populations. Some DLLs may have typical levels of English proficiency, allowing them to demonstrate knowledge and skills in mathematics through English-language assessments. Other DLLs have limited English proficiency, and poor performance on English mathematics achievement tests may reflect limitations in language ability rather than deficits in mathematics skills. One common assessment accommodation for DLLs with limited English proficiency is to administer assessments adapted to children's home language. However, even when assessments are available in children's home language (e.g., Spanish), the equivalence of these adapted tests may be under-evaluated and under-reported. Validity evidence to support alternate-language adaptations of English mathematics achievement tests is thus lacking.

To address this limitation, the current study evaluated construct validity evidence of the Spanish-adapted kindergarten mathematics assessment used in the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K). Specifically, we used differential item functioning (DIF) analysis to evaluate whether, among Spanish-speaking DLLs, the Spanish-adapted items measured the same construct and were on the same scale (i.e., had equivalent item difficulty) as the items on the English version of the mathematics assessment. Subsequently, we conducted item content scrutiny to determine whether items flagged for DIF indicated item bias. Notably, traditional methods for investigating DIF only control for ability level; if the 2 groups differ on characteristics other than ability level and assessment form/English language proficiency, inferences regarding DIF (and item bias) will be confounded by selection bias. Therefore, we also investigated the presence of selection bias and compared DIF and item bias results from a traditional method for investigating DIF with results from a novel approach that controls for selection bias by drawing on a regression discontinuity design framework. Such an approach may be necessary to ensure that assessments used with DLLs do not lead to bias and over- or under-identification of DLLs as having language or learning disabilities.

Assessment of mathematics achievement for dual language learners

Level of English proficiency represents a critical barrier to discriminating DLLs with and without learning difficulties in an unbiased manner. Limited English proficiency hinders children's ability to understand and respond accurately to mathematics questions presented in English, even when children possess adequate mathematics skills. Because of difficulties related to assessment of DLLs, some evidence indicates that DLLs are often under-identified as having learning difficulties (Morgan et al., 2015), especially in the early elementary years when teachers are hesitant to refer children with limited English proficiency for special education evaluations (Samson & Lesaux, 2009). Conversely, DLLs are over-identified as having learning difficulties in the late elementary years and are identified as needing special education later than are monolingual children (Samson & Lesaux, 2009; Sullivan, 2011). Disproportionate representation of DLLs in special education has important implications for student achievement, as under-identification could result in many DLLs with learning difficulties not receiving the supports they need to succeed in the early school years. Conversely, over-identification leads to misallocation of resources toward children without disabilities and could potentially lead to negative stigma associated with a learning disability label. Therefore, addressing English proficiency in the assessment/identification process is important for equitable, fair assessment of DLLs and subsequent valid identification of learning disabilities.

Federal regulation requires that education agencies provide accommodation for DLLs when administering assessments of academic achievement. According to the Individuals with Disabilities Education Act (2004), evaluation of learning disabilities for DLLs should be "provided and administered in the child's native language…in the form most likely to yield accurate information on what the child knows and can do academically." Similarly, the Every Student Succeeds Act (2015) specifies that states must make efforts to provide assessments in children's home language when there is a "significant" prevalence of a given language. Additional recommendations for DLLs have included conducting assessments in both the home language and English, using informal, curriculum-based assessments, dynamic assessment procedures, or portfolio assessments (e.g., Spinelli, 2008).

Assessment accommodations for DLLs must be supported by both validity and effectiveness evidence (Abedi, Zhang, Rowe & Lee, 2020). According to Abedi et al., accommodations are effective when they make assessments more accessible to the test takers. Robinson (2010) evaluated the effectiveness of the Spanish-language adaptation of the ECLS-K kindergarten mathematics assessment, using a regression discontinuity approach to control for selection bias, and reported that English language learners who completed the mathematics achievement test in Spanish scored significantly higher than English language learners who completed the mathematics assessment in English at 2 out of 3 assessment points (spring of kindergarten and first grade, but not fall of kindergarten). Furthermore, effect sizes reported across Spanish and English mathematics assessments were large (i.e., $ds > 0.85$). This indicates that, as an accommodation, use of the Spanish-language form of the assessment was effective, as children were better able to demonstrate their mathematics knowledge when given the Spanish assessment. However, effectiveness evidence is only meaningful to the extent that the accommodation is supported by validity evidence. Validity of accommodations is established when it is determined that the accommodation does not alter the construct being measured and does not provide an unfair advantage on the assessment (Abedi et al., 2020). It is possible, for example, that group differences observed by Robinson exist not because the English assessment underestimates children's true ability, but because items on the Spanish-adapted assessment function differently (are less difficult) than the corresponding English items, controlling for underlying mathematics ability. If this were the case, Spanish-speakers assigned to the Spanish form of the assessment would have an unfair advantage over Spanish speakers assigned to the English form. Thus, more research is needed to determine the validity of assessment accommodations prior to evaluating the effectiveness of those accommodations.

Differential item functioning

Translating mathematics assessments from one language to another and using the translated tests to measure achievement among DLLs requires the assumption that the adapted-language items operate equivalently to the English-language items. However, this assumption may be violated in practice. For example, a language-based word problem may have information that does not translate directly to another language or is not culturally relevant to certain groups of children. Holding achievement level constant, items that operate differently across forms are said to demonstrate DIF (Embretson & Reise, 2000).

Test items can demonstrate 1 of 2 types of DIF, uniform and non-uniform DIF. Items that demonstrate uniform DIF are more difficult on 1 form of the assessment than the other, holding ability level constant. In other words, for a mathematics item that displays uniform DIF, 2 test takers who complete different forms of the assessment have different probabilities of responding to the item correctly, despite having the same underlying mathematics ability. Non-uniform DIF occurs across assessment forms when a test item is more strongly associated with the underlying con-

struct and better able to differentiate among different ability levels for 1 form of the assessment than the other. For example, an item that is poorly translated may not distinguish well between test takers with different ability levels. In this case, DIF is said to be non-uniform because the difference in difficulty of the item across assessment forms varies across the range of ability (e.g., for test takers with low levels of mathematics ability the item is more difficult on the Spanish assessment but for test takers with high levels of mathematics ability the item is more difficult on the English assessment). There are multiple approaches for evaluating DIF, and for this study we implement a probit regression approach (see Method).

Items and tests that demonstrate substantial amounts of DIF may be biased in favor of 1 group or assessment form, but the presence of DIF alone does not necessarily indicate item or test bias (e.g., Penfield & Lam, 2000). To make the assertion of bias, item content must be examined to determine whether there is a theoretical basis for the existence of bias in the presence of DIF. If items demonstrating DIF are assessing constructs other than the construct intended, it may be concluded that the item is biased. For example, a word problem on a mathematics assessment may measure children's language or reading skills, in addition to their mathematics skills (e.g., Goodrich & Namkung, 2019). Therefore, word problems may have the potential to demonstrate bias across groups with differing levels of language or reading ability, such as when comparing performance across monolingual children and DLLs, as differences in performance across items may not represent differences in mathematics ability.

Prior research examining DIF for DLLs has primarily compared item functioning across monolingual English-speaking children and DLLs in which both groups completed the assessment in English (Farrington, Lonigan, Phillips, Farver & McDowell, 2015; Goodrich, Lonigan & Alfonso, 2019; Kamata, Chaimongkol, Genc & Bilir, 2005; Lakin, Elliott & Liu, 2012; Mahoney, 2008; Martiniello, 2009; Ockey, 2007; Snetzler & Qualls, 2000). Several of these studies have not found substantial evidence of DIF across DLLs and monolingual children on achievement tests (Lakin et al., 2012; Mahoney, 2008; Martiniello, 2009; Ockey, 2007). Other studies have reported that numerous test items demonstrate DIF (e.g., Farrington et al., 2015; Goodrich et al., 2019; Kamata et al., 2005; Snetzler & Qualls, 2000); however, these studies have generally reported that DIF did not systematically favor 1 group over the other across items, providing little indication of test bias.

Despite a lack of evidence of item and test bias across monolingual children and DLLs for achievement tests administered in English, it is possible that DIF and bias exist across adapted-language and original English-language versions of tests used to evaluate academic achievement. For example, if the English language proficiency cut score used to determine language of assessment is too low, then items on the English version may be more difficult than items on the adapted-language form when evaluating DIF among DLLs. Of relevance to this study, in comparing the English-language version of the mathematics assessment to the Spanish-adapted language version, test developers for the ECLS-K study concluded that the items operated similarly across forms (U.S. Department of Education [DoEd], 2002). However, this comparison was based on a heterogeneous sample in which the group who received the English-language version, which included monolingual English speakers, Spanish-speaking DLLs, and DLLs with other home languages who passed the English language proficiency screener, likely differed in many ways from the Spanish-speaking DLLs who did not pass the screener and thus received the Spanish-adapted assessment. Given this threat of selection bias, a more controlled approach would be to evaluate DIF across assessment forms specifically among the Spanish-speaking children in the sample. Embedding a quasi-experimental design (regression discontinuity design) in the evaluation of DIF would offer even greater control.

Controlling for selection bias in evaluation of differential item functioning

When administering alternate-language forms of mathematics assessments for DLLs, typical practice is to screen children for English language proficiency and administer the alternate-language version to children who score below a certain cutpoint and the English-language version to children above that cutpoint. The assumption underlying this practice is that the 2 groups of DLLs differ in their level of English proficiency but are otherwise similar (i.e., there is no selection bias). However, it is possible that DLLs with lower levels of English proficiency differ significantly from DLLs with typical levels of English proficiency on various dimensions (e.g., socioeconomic status, reading skills). Using a traditional method for investigating DIF that controls for ability level on the underlying construct but does not account for differences on other relevant variables may be inappropriate, as assignment to test form is potentially confounded by selection bias. In this situation, it is not possible to know whether DIF (and associated item/test bias) is due to form-related differences or other group differences.

To control for selection bias in the evaluation of DIF across assessment forms for DLLs, we propose a novel approach that capitalizes on naturally occurring regression discontinuity designs (RDD; Imbens & Lemieux, 2008) present in this type of assessment context. RDD is considered to be the most rigorous alternative to randomized controlled trial design to facilitate causal inference (Bloom, 2012). In RDD, treatment assignment is determined by a cutpoint on an assignment measure that typically assesses participants' need for treatment. For example, in the present study, an English language proficiency screener is used to assign Spanish-speaking DLLs to either an English or Spanish-adapted version of a mathematics assessment. If the "treatment" (language form) has an impact on the outcome (in this study, mathematics item response), then a discontinuity in the regression relationship between the assignment variable (English language proficiency score) and the mean outcome (item response) occurs at the cutpoint of the assignment variable (at the score that defines English proficiency). Figure S1 in the supplementary material, for example, illustrates a discontinuity in the relationship between English proficiency score and SES at the cutpoint for defining English proficiency, suggesting systematic differences in SES between DLLs assigned to the English vs Spanish-adapted form.

The major advantage of RDD stems from its ability to estimate an unbiased causal estimate *at the cutpoint*, under the condition that all RDD assumptions are met (see Procedure and Analytic Approach). This is because the participants in the treatment and control groups near the cutpoint are comparable, except in terms of treatment assignment status. In our proposed approach for investigating DIF (and item/test bias), we control for selection bias by embedding an RDD within an existing framework for evaluating DIF. Specifically, by incorporating RDD terms into a traditional model for investigating DIF, we conditionalize estimates of DIF at the English language proficiency cutpoint of the language screener used to assign children to the English vs Spanish version of the mathematics assessment in the ECLS-K study. Thus, characteristics of children who are clustered just above and below the cut-point of the language screener should be comparable. Controlling for selection bias theoretically allows for causal conclusions to be drawn regarding the source of observed DIF (i.e., the language form).

Current study

The primary purposes of this study were to investigate DIF and item bias across Spanish and English forms of an assessment of early mathematics skills. Secondary purposes were to investigate the presence of selection bias and demonstrate a novel approach for investigating DIF that uses a regression discontinuity design framework to control for selection bias. We addressed the following research questions:

1. Do any mathematics items demonstrate DIF across the Spanish- and English-language administrations using:
   a. a traditional approach for evaluating DIF that controls for achievement level but ignores other sources of selection bias?
   b. an RDD approach for evaluating DIF that controls for achievement level and other sources of selection bias?
2. Among items that demonstrate DIF, are there any patterns that suggest item/test bias:
   a. for items flagged as showing DIF using the traditional approach?
   b. for items flagged as showing DIF using the RDD approach?

When using a traditional approach, we expected there to be substantial DIF across assessment forms, primarily because children who completed different versions of the assessment likely differed on many other dimensions besides achievement level and English language proficiency. When using an RDD approach, we expected to observe less evidence of DIF because the potential threat of selection bias would be better controlled. We did not formulate specific a priori hypotheses regarding patterns of potential item/test bias, as we were not able to statistically evaluate bias. Rather, we performed a post hoc, exploratory, subjective evaluation of item content to search for potential patterns across items that may explain observed DIF.

Method

Data source

Secondary data were drawn from the ECLS-K restricted-use dataset (ECLS-K; DoEd, 2000). The authors received a certification of exemption from their university's IRB to carry out the analysis. The ECLS-K, sponsored by the National Center for Education Statistics, followed a U.S. nationally representative sample of 21,260[1] children from kindergarten in 1998–99 through 8th grade in 2006–07 with the purpose of describing and identifying correlates of children's development. A stratified, multistage probability sampling design was used to select participants. Data collection spanned multiple sources and methods, including direct child assessments, child questionnaires, parent interviews, and teacher and school administrator questionnaires.

Participants

The current study used kindergarten data from 1750 children who were identified as having a home language of Spanish and had base year panel weights and mathematics and oral English language proficiency scores (see *Measures and Variables*). Home language was determined by the ECLS-K team by checking school records, and in the absence of this information, by consulting the child's teacher to determine whether: 1) The child spoke a language other than English; 2) The child's family members spoke a language other than English; and/or 3) The child had been observed conversing in a language other than English (DoEd, 2000). If 1 or more criteria were met, teachers were asked to specify the non-English language. Children with an IEP-listed accommodation of Braille, enlarged print, or sign language, or with an IEP that indicated they could not participate in standardized assessments, were excluded from direct cognitive testing and thus excluded from the current study. Less than 1% of children with a home language of Spanish and base year panel weights had missing mathematics scores and English proficiency scores.

Weighted descriptives for the current study's target subpopulation of kindergarteners (N = 381,773) vs the total population of kindergarteners (N = 3865,945) are as follows: 50.1% vs 51.3% male; 7.2% vs 14.7% with a parent-reported disability; 87.8% vs 97.3% born in the U.S.; M = 67.69 (SD = 4.35) vs M = 68.50 (SD = 4.45) months of age at the fall kindergarten assessment; M = −0.63 (SD = 0.62) vs M = −0.03 (SD = 0.79) household SES composite; and M = −1.55 (SD = 0.42) vs M = −1.19 (SD = 0.48) fall kindergarten mathematics IRT theta score.

Measures and variables[2]

Mathematics achievement. The ECLS-K design team, drawing heavily on the "*Mathematics Framework for the 1996 National Assessment of Educational Progress* (National Assessment Governing Board, 1996)" (p. 2–7; DoEd, 2002), developed a direct assessment of kindergarten mathematics achievement that covered 5 content strands: 1) Number sense, properties, and operations (≈73% of items included in the kindergarten scale scores); 2) Measurement (≈4%); 3) Geometry and spatial sense (≈4%); 4) Data analysis, statistics and probability (≈8%); and 5) Patterns, algebra, and functions (≈12%). Items were drawn from the Peabody Individual Achievement Test-Revised (PIAT-r; Markwardt, 1989); Primary Test of Cognitive Skills (PTOC; Huttenlocher & Levine, 1990); Test of Early Mathematics Ability (TEMA-2; Ginsburg & Baroody, 1990), and Woodcock Johnson Tests of Achievement-R (WJ-R; Woodcock & Bonner, 1989), as well as newly developed by curriculum specialists and teachers. A Spanish version was developed using forward- and back-translation procedures, with expert feedback provided by mathematicians whose native language was Spanish (DoEd, 2002). Of the 1750 children in the current study with a home language of Spanish, 1020 (58%) completed the Spanish version of the assessment. Both multiple-choice and open-ended response formats were used. All items were designed to be answered orally or by pointing to the answer.

The kindergarten mathematics assessment was untimed and individually administered using an adaptive design in which all children received a common core of 16 items (plus 2 practice items) that was then used to route them to a low difficulty form (18 items), medium (23 items), or high form (31 items). One of the open-ended routing items was subsequently divided into 2 distinct items at the time of scoring, for a total of 17 non-practice routing items. Among children in the target subsample, 95% were routed to the low form, 4% the medium form, and <1% the high form, compared to 77% of children in the total sample who were routed to the low form, 17% the medium form, and 6% the high form (p. 73; DoEd, 2002). The full assessment included 64 scored items, of which 51 were used in the construction of the kindergarten scale scores and the remaining were linking items used only for vertical scaling purposes.

ECLS-K psychometricians calibrated the mathematics items via item response theory (IRT) using a pooled dataset containing re-

[1] Unweighted sample sizes are rounded to the nearest 10 per Institute of Education Sciences restricted-use data requirements.

[2] Unless otherwise noted variables were drawn from fall of kindergarten (Wave 1).

sponses to all mathematics items across all grade levels. Estimated reliability of the Wave 1 IRT theta scores, using data from all children, was 0.92 (DoEd, 2002). Mathematics scores were strongly correlated with reading ($r = 0.77$) and general knowledge ($r = 0.64$) scores at Wave 1.

The current study evaluated DIF for 28 items that were included on the routing or low form and were used in the construction of the kindergarten scale scores (the 2 practice items and the 7 items on the routing form that were used only for linking purposes were not evaluated). Descriptive statistics for these items are presented in Table S1 of the supplemental materials. Based on data from the target subsample, the proportion correct ranged from 0.05 to 0.96 with $M = 0.42$ (SD = 0.24), and the point-biserial correlation between the item and overall mathematics theta score ranged from 0.16 to 0.71 with $M = 0.39$ (SD = 0.15). Sample size varied slightly across analyses as a function of the items' form assignment (routing vs low vs low and medium and/or high form) and due to item-level missingness that occurred when children refused to answer or said "I don't know." This latter reason for missingness was rarely observed (maximum = 3.4% and median = 0.1% across the 28 items). Too few children in the target subsample were routed to the medium and high forms to facilitate DIF analyses of items that only appeared on those forms.

Oral English language proficiency. Children's oral English proficiency was used to route them to either the English or Spanish version of the mathematics assessment (DoEd, 2002). Conceptualized by the ECLS-K design team in consultation with an expert panel, proficiency was measured via performance on the Oral Language Development Scale (OLDS), an untimed and individually administered direct assessment. The OLDS comprises 3 subtests of the PreLAS 2000 (Duncan & DeAvila, 1998): 1) "Simon Says," a 10-item measure of receptive language with a maximum score of 10; 2) "Art Show," a 10-item measure of expressive language (picture vocabulary) with a maximum score of 10; and 3) "Let's Tell Stories," a measure of receptive and expressive language based on the retelling of 2 story prompts with a maximum score of 40. English proficiency was defined as a total OLDS raw score $\geq$37, which "was based on results of a national norming sample for PreLAS, extrapolated to the 3 selected subtests" (p. 2–22; DoEd, 2002). The sample frequency distribution of OLDS scores is shown in Table S2 of supplemental materials. Assessors had to achieve 90% accuracy on a set of training stories to administer the OLDS. Ongoing interrater reliability was evaluated by recoding 10% of each assessor's stories. Estimated split-half reliability of the Wave 1 scores, using data from all children in the sample with a home language other than English, was 0.97.

Child and family characteristics. Child and family characteristics were obtained via parent interview. Child characteristics included gender, age, disability status (with disability vs not), country of birth reported at Wave 2 (U.S. born vs not), childcare (ever in center-based care vs not), and parent-rated impulsiveness/overactiveness, approaches to learning, self-control, social interaction, and sadness/loneliness which were derived from the Social Rating Scale (SRS; adapted from the Social Skills Rating System; Gresham & Elliott, 1990). A child was categorized by the ECLS-K team as having a disability if the parent reported a professional diagnosis related to the child's difficulty with paying attention, learning, activity level, coordination in moving limbs, communication, hearing, and/or seeing, and/or reported that the child had received therapy services for a disability.

Family characteristics included household SES, parents in the home (both biological parents vs other), number of people in the home, number of children's books in the home, and parent involvement. SES was a composite created by the ECLS-K team that combined information about the education level and occupational prestige of the child's mother and father figures (Wave 1) and household income (Wave 2). Parent involvement was computed as the average of 9 questions related to the frequency (ranging from 1 = Not at all to 4 = Every day) with which a family member engaged with the child by reading books, telling stories, singing songs, helping with arts/crafts, involving the child in chores, playing games/puzzles, talking about nature/science, building/playing with construction toys, and playing sports/exercising.

Procedure and analytic approach

Analyses were performed in R Version 3.6.1 (R Core Team, 2019) and M *plus* Version 8 (Muthén & Muthén, 1998–2017). Fall kindergarten child-level weights (C1CW0) were applied to adjust for unequal probabilities of selection and nonresponse. The jackknife replication method was used to obtain design-adjusted standard errors and test statistics, with α set at 0.05.

Preliminary analysis. Multiple tests were performed to evaluate assumptions underlying the use of the traditional and RDD approaches. Covariate balance under the traditional DIF approach was evaluated in M*plus* by performing bivariate group comparisons (English vs Spanish form) on demographic variables and the mathematics IRT theta scores, with standardized mean differences (SMDs) and odds ratios (ORs) as measures of effect size.

Assumptions underlying the RDD approach were tested as follows. First, frequency statistics disaggregated by group were calculated to determine English and Spanish forms were correctly assigned based on the OLDS cutpoint of 37. (Note for all RDD analyses, we specified a cutpoint of 36.5 instead of 37 to minimize extrapolation on both sides of the cutpoint; see Robinson, 2010.) Second, the rdrobust Version 0.99.4 package (Calonico, Cattaneo, Farrell & Titiunik, 2018) was used to test balance on demographic variables and mathematics achievement at the OLDS cutpoint. Third, a density discontinuity test, or McCrary test (McCrary, 2008), available via the rddensity Version 1.0 package in R (Cattaneo, Jansson & Ma, 2019), was performed to test for discontinuity in the distribution of the OLDS scores at the cutpoint which could signify manipulation of the OLDS scores. Fourth, in estimating DIF, multiple bandwidths were considered to evaluate the sensitivity of the results.

RQ1: Investigation of DIF. Details of the analytic approach are available in the supplementary material. A traditional evaluation of DIF was performed in M*plus* by estimating a series of probit regression models with ability, group, and ability by group terms. Models were estimated via mean- and variance-adjusted weighted least squares (WLSMV). An item was flagged as demonstrating non-uniform DIF if the ability by group interaction was significantly different from 0. In the absence of non-uniform DIF, an item was flagged as demonstrating uniform DIF if the ability term was significantly different from 0 based on the reduced model.

An RDD approach for evaluating DIF was performed in M*plus* via local linear probit regression with WLSMV estimation and robust bias-corrected standard errors (Cattaneo, Idrobo & Titiunik, 2020). The model was similar to that of the traditional approach but with the addition of OLDS and group by OLDS terms, and where only cases with an OLDS score falling within an empirically-derived bandwidth were included in the analysis. Bandwidths 1 point smaller and bigger than the derived values were obtained for sensitivity testing.

For both DIF approaches, practical significance was assessed by estimating the signed and unsigned item difference in the sample (SIDS and UIDS, respectively; Meade, 2010). Because sampling weights are used in the current study, we instead refer to the effect sizes in terms of the population (SIDP, UIDP). The SIDP indicates the average difference in the expected item score across forms (English vs Spanish), holding constant mathematics ability, within the population represented by children who took the Spanish version.

Table 1
Comparison of kindergarten child and family characteristics across mathematics assessment language groups.

	Simple group difference (English Form - Spanish Form)					Group difference at the English proficiency cutpoint (English Form - Spanish Form)				
Variable	*n*	Estimate	SE	*p*	Effect Size	n	Estimate	SE	*p*	Effect Size
Female	1750	0.02[a] (0.01[b])	0.07[a]	0.786	OR = 1.03	550	−0.13[b]	0.16[b]	0.414	OR = 0.59
Disability	1540	0.03[a] (0.00[b])	0.13[a]	.833	OR = 1.06	520	0.05[b]	0.06[b]	0.291	OR = 6.40
U.S. born	1510	0.53[a] (0.10[b])	0.09[a]	<0.001	OR = 2.79	540	0.07[b]	0.08[b]	0.207	OR = 2.12
Age in months	1750	1.00	0.26	<0.001	SMD = 0.23	390	0.44	1.10	0.601	SMD = 0.10
Ever in center-based care	1530	0.41[a] (0.16[b])	0.08[a]	<0.001	OR = 1.50	520	−0.01[b]	0.14[b]	0.789	OR = 0.96
Household SES	1640	0.46	0.04	<0.001	SMD = 0.75	300	0.39	0.28	0.041	SMD = 0.63
Both biological parents in home	1540	−0.15[a] (−0.05[b])	0.07[a]	0.041	OR = 0.86	400	−0.02[b]	0.19[b]	0.886	OR = 0.91
Number people in household	1540	−0.57	0.10	<0.001	SMD = −0.35	610	0.06	0.62	0.922	SMD = 0.04
Number children's books in home	1530	22.39	1.87	<0.001	SMD = 0.75	400	6.11	13.11	0.775	SMD = 0.20
Parent involvement[c]	1540	0.14	0.03	<0.001	SMD = 0.27	520	0.21	0.19	0.406	SMD = 0.39
Mathematics IRT score	1750	0.67	0.07	<0.001	SMD = 0.67	390	0.18	0.22	0.179	SMD = 0.18
Impulsive/overactive[d]	1520	−0.06	0.04	0.101	SMD = −0.09	390	0.04	0.27	0.836	SMD = 0.06
Approaches to learning[d]	1530	0.14	0.03	<0.001	SMD = 0.29	560	0.02	0.20	0.878	SMD = 0.04
Self-control[d]	1530	0.07	0.03	0.045	SMD = 0.12	330	−0.02	0.26	0.580	SMD = −0.04
Social interaction[d]	1530	0.31	0.03	<0.001	SMD = 0.50	560	0.25	0.15	0.066	SMD = 0.40
Sad/lonely[d]	1530	−0.03	0.03	0.225	SMD = −0.07	480	−0.12	0.16	.450	SMD = −0.27

OR = odds ratio. SMD = standardized mean difference.
Note. Sample sizes rounded to the nearest 10.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), "Parent-Guardian Interview," fall 1998 and spring 1999, "Student Assessment," fall 1998.
[a] Probit scale.
[b] Probability scale.
[c] Composite based on the frequency with which family members engage with the child by reading books, telling stories, singing songs, helping with arts/crafts, involving in chores, playing games/puzzles, talking about nature/science, building/playing with construction toys, and playing sports/exercising.
[d] Subscale of the Social Rating Scale (SRS) – parent report (adapted from the Social Skills Rating System; Gresham & Elliott, 1990).

In the presence of non-uniform DIF, SIDP may be close to zero due to positive and negative differences canceling each other out. The UIDP represents the average *absolute value* of the difference in the expected item score across forms and provides a measure of expected item score difference assuming DIF uniformly favors one group. Because the items are dichotomous, SIDP ranges from −1 to 1 and UIDP from 0 to 1. In calculating effect sizes for the RDD approach, the average expected score difference was conditionalized at the OLDS English proficiency cutpoint.

RQ2: Investigation of item bias. Following detection of significant DIF, item content was examined to evaluate the potential for DIF to be representative of item bias, as presence of DIF is not a sufficient condition to determine a given item is biased (Penfield & Lam, 2000). According to Penfield and Lam, to confirm that an item is biased one must be able to conclude that the differential performance across groups (or test forms) is due to a variable other than the underlying construct being assessed. However, to our knowledge there are no specific recommendations regarding procedures for examining item content to confirm bias. For this study, we evaluated whether item content for items displaying DIF revealed patterns that could explain presence of DIF. For example, we evaluated whether items included complex language that may influence student performance or whether item content included specific information (e.g., cultural references) irrelevant to the mathematics construct being assessed. If examination of item content revealed that differential performance of children across forms could be due to systematic reasons unrelated to mathematics ability, we concluded that the items were biased.

Results

Preliminary analysis

Traditional approach. Table 1 shows differences in child/family characteristics across mathematics language forms. Simple group comparisons (left side) indicated no difference in gender, disability, impulsiveness/overactiveness, or sadness/loneliness, but children who were administered the English version were significantly more likely to be born in the U.S., be older, attend center-based care, live in a home with higher SES, fewer people, and more children's books, have more involved parents, and have higher mathematics achievement, approaches to learning, self-control, and social interaction, and significantly less likely to have both biological parents in the home, than children who were administered the Spanish version. Four of the 12 statistically significant differences were moderate to large in size ($|SMD| \geq 0.50$; Cohen, 1988).

RDD approach. Frequency distributions of OLDS scores by language group indicated a sharp cutpoint at 36.5, with no children assigned to the Spanish form scoring above this cutpoint, and no children assigned to the English form scoring below this cutpoint. Table 1 (right side) provides the tests of covariate balance at the OLDS cutpoint. There was only 1 statistically significant group difference: children receiving the English version were significantly more likely to live in a household with higher SES. While the magnitude of this effect was smaller at the English proficiency cutpoint compared to the simple group difference (SMD = 0.63 vs 0.75, respectively), it was still moderate in size. In addition, while statistically non-significant, the magnitude of difference in disability distribution was sizeable (OR = 6.40), with children receiving the English version more likely to have a disability. Figure S1 in the supplementary material provides RD plots (a visual depiction of discontinuity) for all 16 variables. Finally, the McCrary test was non-significant (robust $T = 0.07$, $p = .945$), suggesting no systematic unidirectional manipulation of OLDS scores at the cutpoint (see Figure S2 in the supplementary material for the density discontinuity plot). Additional validity evidence of the RDD approach based on sensitivity of the results to different bandwidths is provided in the next section.

Table 2
Estimated DIF parameters and effect sizes under the traditional approach.

Item	Nonuniform Estimate	Uniform 95% CI	Estimate	95% CI	SIDP	UIDP
1	0.18[a]	(0.05, 0.30)	0.08	(−0.05, 0.21)	0.01	0.05
2	−0.23	(−0.55, 0.10)	0.12	(−0.05, 0.29)	0.03	0.03
3	0.42[a]	(0.06, 0.79)	−0.43[a]	(−0.58, −0.28)	−0.08	0.09
4	0.46[a]	(0.13, 0.79)	0.16	(−0.03, 0.35)	0.01	0.03
5	0.87[a]	(0.60, 1.14)	0.09	(−0.11, 0.28)	0.00	0.04
6	0.09	(−0.06, 0.24)	−0.04	(−0.15, 0.07)	−0.02	0.03
7	0.08	(−0.16, 0.32)	−0.11	(−0.30, 0.07)	−0.02	0.02
9	0.04	(−0.14, 0.23)	0.10	(−0.09, 0.28)	0.02	0.02
10	0.10	(−0.09, 0.28)	−0.14	(−0.29, 0.02)	−0.04	0.04
17	−0.03	(−0.60, 0.54)	0.05	(−0.26, 0.36)	0.00	0.00
18	−0.16	(−0.40, 0.08)	0.12	(−0.03, 0.28)	0.05	0.05
19	0.05	(−0.13, 0.24)	0.02	(−0.12, 0.15)	0.00	0.01
20	0.25	(−0.25, 0.76)	0.08	(−0.17, 0.33)	0.01	0.01
21	0.14	(−0.17, 0.46)	−0.06	(−0.22, 0.11)	−0.02	0.02
22	0.08	(−0.18, 0.34)	0.03	(−0.16, 0.22)	0.00	0.01
23	0.15	(−0.08, 0.37)	0.07	(−0.13, 0.27)	0.00	0.02
24	0.19[a]	(0.04, 0.34)	0.08	(−0.10, 0.27)	0.01	0.05
25	0.21[a]	(0.07, 0.34)	0.04	(−0.12, 0.20)	−0.01	0.05
26	0.03	(−0.11, 0.16)	−0.05	(−0.20, 0.09)	−0.02	0.02
27	0.06	(−0.16, 0.28)	−0.03	(−0.21, 0.15)	−0.01	0.01
28	0.03	(−0.12, 0.18)	0.03	(−0.10, 0.15)	0.01	0.01
29	0.24	(−0.06, 0.55)	−0.07	(−0.32, 0.18)	−0.01	0.01
30	0.19	(−0.03, 0.41)	−0.12	(−0.27, 0.03)	−0.04	0.04
31	0.04	(−0.14, 0.22)	−0.10	(−0.26, 0.06)	−0.04	0.04
32	−0.08	(−0.26, 0.10)	0.02	(−0.14, 0.18)	0.01	0.02
33	−0.01	(−0.17, 0.14)	0.00	(−0.17, 0.16)	0.00	0.00
34	0.17	(−0.02, 0.36)	−0.12	(−0.26, 0.03)	−0.05	0.06
64	−0.34[a]	(−0.62, −0.07)	0.45[a]	(0.24, 0.65)	0.14	0.14

Note. Sample sizes are provided in Table S1. The Spanish group is the reference group. Uniform DIF estimates are based on reduced model with non-uniform DIF parameter fixed at 0. SIDP = signed item difference (English - Spanish) in the population for the focal (Spanish) group. UIDP = unsigned item difference (English - Spanish) in the population for the focal (Spanish) group.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), "Student Assessment," fall 1998.
[a] $p < 0.05$.

RQ1: investigation of DIF

Traditional approach. Table 2 provides results for DIF under the traditional approach. Seven items (25%) were flagged, all of which exhibited non-uniform DIF.[3] Fig. 1 illustrates the predicted probability of a correct response by form for these items. Holding mathematics achievement constant, items 1, 3, 4, 5, 24, and 25 were more discriminating (i.e., more strongly related to mathematics achievement) for the group that took the English version. Effect sizes were relatively small, ranging from UIDP = 0.03 (item 4) to 0.09 (item 3). In contrast, item 64 was more discriminating for the group that took the Spanish version (UIDP = 0.14).

RDD approach. Table 3 provides results for DIF under the RDD approach. Although the total sample size exceeded 1600 for all items, the effective sample size (the number of cases used in the RDD analysis) was less than 700 across all RDD analyses, less than 500 for approximately 60% of the analyses, and as small as 260 under the narrowest bandwidth.

Based on the optimal bandwidth, 3 items (11%) were flagged, all of which exhibited uniform DIF. Fig. 2 illustrates the predicted probability of a correct response by form language. Holding achievement constant, items 6, 28, and 29 were harder for the group that took the English form. Effect sizes ranged from UIDP = 0.01 (item 29) to 0.41 (item 6).

Results based on smaller and larger bandwidths (optimal bandwidth ± 1) are also presented in Table 3. Items 6 and 28 were also flagged for uniform DIF with the larger bandwidth but not the smaller bandwidth, whereas item 29 was also flagged for uniform DIF with the smaller bandwidth but not the larger bandwidth. Items 9 and 17 were additionally flagged for non-uniform DIF (less discriminating for English) with the larger bandwidth, and Item 5 was additionally flagged for uniform DIF (harder for English) with the smaller bandwidth. This indicated results were sensitive to the choice of bandwidth in the RDD analysis; however, magnitude of DIF effect sizes did not differ across bandwidths. It is also of note that for 3 of the 7 items flagged for non-uniform DIF under the traditional approach, the estimate of the non-uniform parameter was the same or larger under the RDD approach.

RQ2: investigation of item bias

*Traditional approach***.** Among the 7 items that were flagged for non-uniform DIF, 3 (items 1, 24, and 25) were Patterns, Algebra, and Functions items (comprising 75% of all Patterns, Algebra, and Functions items). These items required children to identify a pattern that matched a stimulus pattern. The language presented to the child was identical across all 3 of these items, and all items were multiple choice (30% of all multiple-choice items) with 4 possible response options. This pattern of results may be related to differences in children across assessment forms (e.g., higher SES for English test takers); however, the overall small magnitude of effect size suggests any degree of bias is minimal.[4] The remaining 4 items exhibiting DIF were all Number Sense, Properties, and

[3] Two of these items also exhibited uniform DIF, but we do not interpret the results because it would be akin to interpreting a main effect in the presence of a significant interaction effect.

[4] Caution is warranted in interpreting results of scores on the Patterns, Algebra, and Functions items independently of the total score. Given that 75% of these items demonstrated DIF, it is possible that a subscale score derived from these items would be biased.

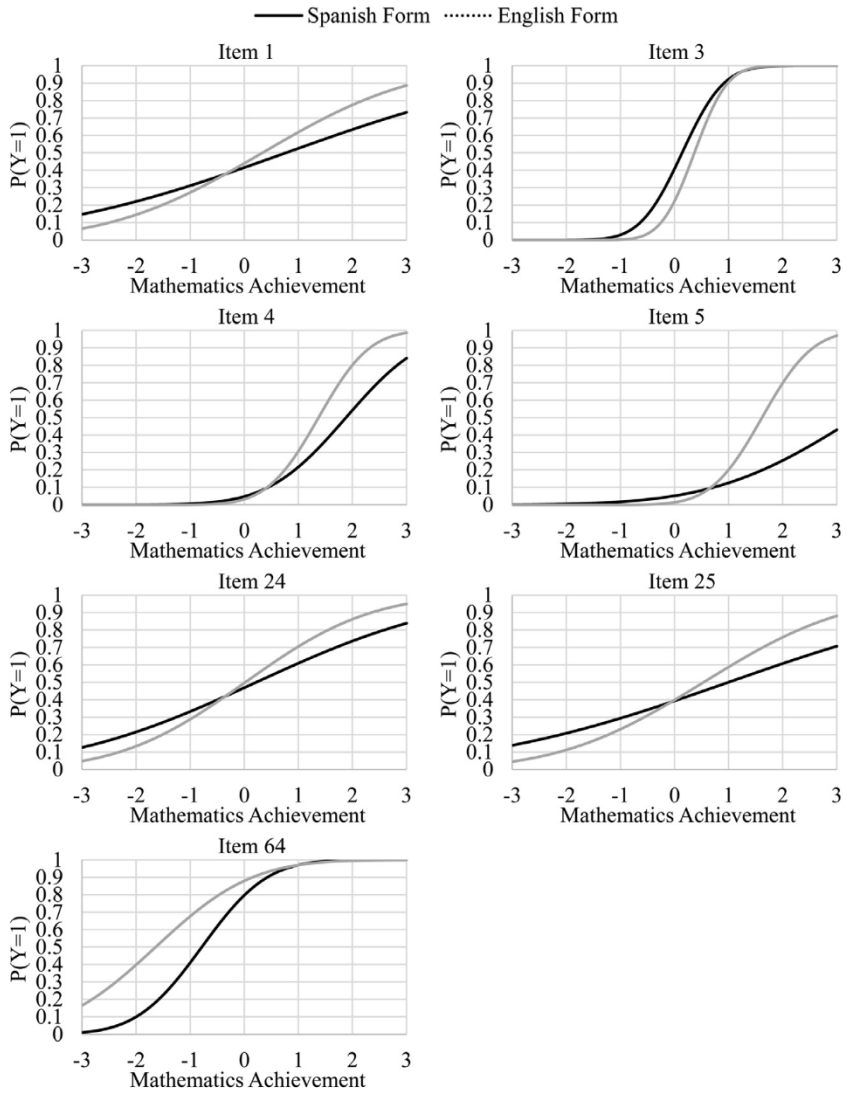

Fig. 1. Predicted probability of a correct response by mathematics assessment language group for items flagged for DIF under the traditional method. Sample sizes are provided in Table S1. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), "Student Assessment," fall 1998.

Operations items (21% of all Number Sense, Properties, and Operations items). All of these items were open ended responses (as was item 64, described below; 22% of all open-ended items). Of the 3 items for which performance was more strongly related to mathematics ability for the group that took the English test, 2 items (items 3 and 4) required children to identify the name of a displayed numeral and had identical stimulus language across items. The remaining item (item 5) required children to identify an object by its order in a sequence. The effect size for DIF was smaller for this item than for some other items, although the estimated coefficient for DIF was larger than it was for other items. The item that was more related to math ability for the group that took the Spanish version (item 64) required children to count objects and credit for the item was given if they counted to a given number. No items were flagged as having DIF for other content areas.

RDD approach. Among the 3 items that demonstrated uniform DIF (such that items were more difficult for the English form) when using the RDD approach, 2 (items 28 and 29) were Number Sense, Properties, and Operations items (11% of all Number Sense, Properties, and Operations items). These items required simple addition and subtraction operations and were open ended. The remaining item (item 6) that displayed DIF was a Measurement item (50% of all Measurement items). This item was a multiple-choice item for which children were shown an object and required to point to the 1 of 4 other objects that was shorter. No items were flagged as having DIF for other content areas.

Discussion

The primary purposes of this study were to investigate DIF and item bias across Spanish and English forms of an assessment of early mathematics skills. Secondary purposes were to investigate the presence of selection bias and demonstrate a novel approach for investigating DIF that uses a regression discontinuity design framework to control for selection bias. Overall, there was minimal evidence of DIF and item/test bias across the Spanish and English forms of the ECLS-K kindergarten mathematics assessment. In contrast, there was significant evidence of selection bias, suggesting the need for a modified approach for evaluating DIF to account for such bias. These findings have important implications for measurement of mathematics skills (and other academic skills) among DLLs and offer a new approach for evaluating the validity of assessment accommodations to ensure equitable, unbiased assessment for all students.

Table 3
Estimated DIF parameters and effect sizes under the RDD approach.

Item	n	Bandwidth	Nonuniform Estimate	Nonuniform 95% CI	Uniform Estimate	Uniform 95% CI	SIDP	UIDP
1	380	6.24	−0.05	(−0.40, 0.31)	−0.28	(−1.02, 1.09)	−0.11	0.11
1	450	7.24	−0.04	(−0.37, 0.30)	−0.27	(−0.93, 0.60)	−0.10	0.10
1	540	8.24	−0.02	(−0.32, 0.28)	−0.23	(−0.88, 0.37)	−0.09	0.11
2	550	7.77	−0.42	(−1.16, 0.32)	0.17	(−0.80, 1.24)	0.03	0.02
2	590	8.77	−0.41	(−1.11, 0.29)	0.17	(−0.70, 1.12)	0.03	0.02
2	640	9.77	−0.42	(−1.08, 0.26)	0.17	(−0.71, 1.12)	0.03	0.02
3	260	4.45	−0.11	(−1.09, 0.92)	0.49	(−1.86, 1.94)	0.09	0.01
3	320	5.45	−0.13	(−1.11, 0.85)	0.47	(−0.85, 1.60)	0.09	0.01
3	390	6.45	−0.06	(−1.01, 0.87)	0.40	(−0.66, 1.72)	0.07	0.01
4	550	7.64	0.95	(−0.41, 2.30)	0.03	(−1.51, 1.21)	0.01	0.02
4	590	8.64	0.88	(−0.32, 2.06)	0.04	(−1.11, 0.89)	0.02	0.02
4	640	9.64	0.82	(−0.17, 1.81)	0.05	(−1.00, 0.87)	0.02	0.02
5	390	5.68	1.27	(−0.21, 2.74)	−1.02[a]	(−3.60, −0.14)	−0.05	0.03
5	460	6.68	1.11	(−0.14, 2.32)	−0.97	(−3.92, 0.88)	−0.05	0.03
5	550	7.68	0.99	(−0.08, 2.04)	−1.04	(−2.58, 0.13)	−0.05	0.03
6	320	5.19	0.01	(−0.32, 0.33)	−0.85	(−1.52, 0.13)	−0.31	0.34
6	390	6.19	0.00	(−0.32, 0.33)	−0.82[a]	(−1.53, −0.05)	−0.30	0.41
6	460	7.19	0.00	(−0.29, 0.29)	−0.79[a]	(−1.53, −0.19)	−0.29	0.38
7	260	3.70	0.19	(−0.74, 1.09)	−0.87	(−2.76, 0.45)	−0.10	0.09
7	320	4.70	0.14	(−0.61, 0.89)	−0.79	(−2.55, 0.27)	−0.09	0.08
7	390	5.70	0.15	(−0.53, 0.84)	−0.71	(−2.42, 0.24)	−0.09	0.07
9	390	5.54	−1.04	(−2.46, 0.21)	−0.29	(−2.52, 0.63)	−0.04	0.03
9	460	6.54	−1.01	(−2.42, 0.10)	−0.24	(−2.36, 0.94)	−0.04	0.03
9	550	7.54	−0.93*	(−1.76, −0.25)	−0.23	(−1.70, 0.76)	−0.04	0.03
10	380	5.55	−0.13	(−0.65, 0.39)	−0.29	(−1.95, 1.16)	−0.04	0.05
10	450	6.55	−0.11	(−0.59, 0.37)	−0.20	(−1.81, 0.88)	−0.03	0.04
10	540	7.55	−0.08	(−0.52, 0.38)	−0.15	(−1.75, 0.87)	−0.02	0.03
17	380	5.90	−1.45	(−5.43, 2.34)	−1.37	(−3.84, 0.24)	−0.06	0.00
17	450	6.90	−1.34	(−3.05, 0.44)	−0.91	(−3.86, 0.53)	−0.04	0.01
17	530	7.90	−1.14[a]	(−2.19, −0.12)	−0.70	(−3.50, 0.79)	−0.03	0.01
18	380	6.06	0.51	(−0.13, 1.14)	0.50	(−0.67, 1.45)	0.10	0.11
18	440	7.06	0.44	(−0.13, 1.02)	0.47	(−0.40, 1.36)	0.09	0.10
18	530	8.06	0.38	(−0.14, 0.90)	0.42	(−0.29, 1.38)	0.08	0.08
19	530	8.00	−0.02	(−0.48, 0.44)	0.03	(−1.74, 1.49)	0.01	0.03
19	570	9.00	−0.01	(−0.43, 0.40)	0.03	(−1.15, 0.95)	0.02	0.03
19	620	10.00	−0.01	(−0.42, 0.39)	0.04	(−1.03, 0.91)	0.02	0.03
20	620	9.59	−0.15	(−1.57, 1.33)	0.55	(−0.50, 1.78)	0.07	0.01
20	670	10.59	−0.11	(−1.31, 1.11)	0.53	(−0.37, 1.67)	0.07	0.01
20	720	11.59	−0.04	(−1.19, 1.10)	0.52	(−0.45, 1.67)	0.07	0.01
21	320	5.41	−0.34	(−1.21, 0.54)	−0.15	(−2.10, 0.84)	−0.02	0.01
21	380	6.41	−0.13	(−0.98, 0.58)	0.00	(−1.79, 0.57)	−0.00	0.00
21	450	7.41	−0.04	(−0.78, 0.60)	−0.01	(−1.26, 0.70)	−0.00	0.00
22	390	6.28	0.29	(−0.62, 1.21)	0.83	(−2.24, 3.89)	0.14	0.06
22	460	7.28	0.35	(−0.48, 1.19)	0.78	(−0.81, 2.64)	0.14	0.05
22	550	8.28	0.41	(−0.38, 1.22)	0.69	(−0.35, 2.39)	0.12	0.04
23	450	6.81	0.06	(−0.35, 0.52)	−0.18	(−2.72, 1.57)	−0.05	0.03
23	530	7.81	0.04	(−0.34, 0.46)	−0.12	(−2.34, 1.40)	−0.03	0.01
23	570	8.81	0.02	(−0.33, 0.40)	−0.09	(−1.92, 1.13)	−0.02	0.00
24	320	5.08	0.05	(−0.31, 0.41)	0.60	(−0.12, 1.82)	0.22	0.10
24	380	6.08	0.08	(−0.25, 0.41)	0.55	(−0.05, 1.67)	0.20	0.08
24	450	7.08	0.09	(−0.21, 0.40)	0.50	(−0.08, 1.56)	0.18	0.09
25	370	5.56	0.22	(−0.16, 0.60)	−0.12	(−0.80, 1.06)	−0.04	0.05
25	440	6.56	0.21	(−0.15, 0.56)	−0.16	(−0.84, 0.94)	−0.06	0.05
25	530	7.56	0.17	(−0.14, 0.47)	−0.16	(−0.89, 0.78)	−0.06	0.05
26	370	5.80	−0.10	(−0.44, 0.26)	0.31	(−1.02, 1.39)	0.12	0.04
26	440	6.80	−0.10	(−0.42, 0.24)	0.31	(−0.98, 1.51)	0.11	0.04
26	530	7.80	−0.10	(−0.41, 0.22)	0.25	(−0.61, 1.35)	0.09	0.04
27	260	4.12	−0.22	(−0.76, 0.32)	0.67	(−0.45, 2.14)	0.16	0.17
27	320	5.12	−0.17	(−0.67, 0.33)	0.64	(−0.54, 2.08)	0.15	0.16
27	390	6.12	−0.13	(−0.59, 0.32)	0.55	(−0.22, 1.89)	0.13	0.13
28	320	5.46	−0.15	(−0.56, 0.25)	−0.98	(−2.41, 0.77)	−0.36	0.36
28	380	6.46	−0.16	(−0.50, 0.19)	−0.88[a]	(−2.02, −0.05)	−0.32	0.33
28	460	7.46	−0.16	(−0.49, 0.16)	−0.84[a]	(−1.84, −0.22)	−0.31	0.31
29	390	5.93	0.14	(−0.93, 0.94)	−0.49[a]	(−3.06, −0.60)	−0.05	0.01
29	460	6.93	0.10	(−0.87, 0.88)	−0.49[a]	(−2.06, −0.05)	−0.04	0.01
29	550	7.93	0.06	(−0.78, 0.74)	−0.57	(−2.15, 0.61)	−0.04	0.01
30	320	5.50	0.16	(−0.43, 0.70)	0.34	(−1.90, 3.41)	0.07	0.02
30	380	6.50	0.11	(−0.38, 0.63)	0.39	(−0.73, 1.70)	0.08	0.04
30	460	7.50	0.08	(−0.37, 0.55)	0.35	(−0.42, 1.36)	0.08	0.06
31	540	8.01	−0.06	(−0.47, 0.35)	0.00	(−0.61, 0.69)	−0.00	0.01
31	580	9.01	−0.05	(−0.42, 0.32)	−0.03	(−0.61, 0.76)	−0.01	0.01
31	630	10.01	−0.05	(−0.41, 0.31)	−0.05	(−0.67, 0.82)	−0.02	0.01
32	320	5.43	0.19	(−0.18, 0.57)	0.27	(−0.40, 1.26)	0.09	0.07
32	390	6.43	0.19	(−0.17, 0.55)	0.16	(−0.57, 1.57)	0.05	0.04

(*continued on next page*)

Table 3 (*continued*)

Item	n	Bandwidth	Nonuniform Estimate	Uniform 95% CI	Estimate	95% CI	SIDP	UIDP
32	460	7.43	0.19	(−0.16, 0.52)	0.11	(−0.64, 1.44)	0.04	0.04
33	590	9.15	0.18	(−0.17, 0.52)	−0.33	(−1.84, 1.76)	−0.12	0.13
33	640	10.15	0.19	(−0.15, 0.52)	−0.34	(−1.40, 1.15)	−0.13	0.13
33	690	11.15	0.19	(−0.13, 0.52)	−0.35	(−1.20, 0.80)	−0.13	0.12
34	440	6.77	0.19	(−0.22, 0.60)	0.22	(−1.05, 1.69)	0.06	0.03
34	530	7.77	0.23	(−0.15, 0.59)	0.26	(−0.76, 1.20)	0.07	0.03
34	570	8.77	0.26	(−0.06, 0.56)	0.28	(−0.74, 1.17)	0.08	0.04
64	550	8.45	−0.35	(−0.97, 0.26)	0.20	(−0.89, 0.96)	0.06	0.03
64	590	9.45	−0.33	(−0.88, 0.21)	0.22	(−0.78, 0.89)	0.06	0.03
64	640	10.45	−0.33	(−0.77, 0.10)	0.27	(−0.75, 0.80)	0.08	0.04

Note. Sample sizes rounded to the nearest 10. The Spanish group is the reference group. Uniform DIF estimates are based on reduced model with non-uniform DIF parameter fixed at 0. Middle row for each item indicates data-driven bandwidth. First and third row for each item (±1 data-driven bandwidth) included for sensitivity testing. SIDP = signed item difference (English - Spanish) in the population for the focal (Spanish) group. UIDP = unsigned item difference (English - Spanish) in the population for the focal (Spanish) group.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), "Student Assessment," fall 1998.
[a] $p < 0.05$.

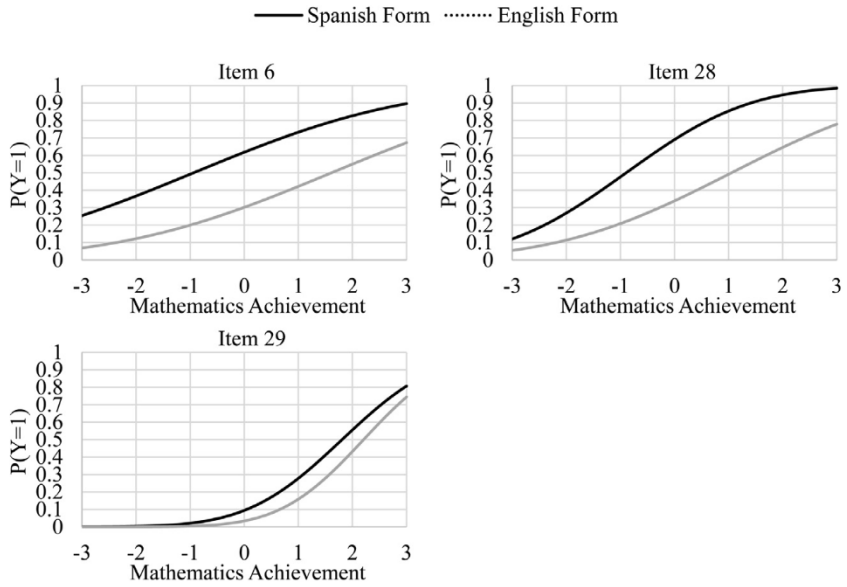

Fig. 2. Predicted probability of a correct response by mathematics assessment language group for items flagged for DIF under the RDD method. Sample sizes are provided in Table 2. SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), "Student Assessment," fall 1998.

Evaluating equivalence of measurement across test forms

Although native-language test accommodations are theoretically rooted in practices designed to promote equity, their validity depends on certain assumptions. In addition to the assumption that children have greater proficiency in their native language than in English, this approach assumes children receiving different language forms are otherwise similar and that the translations or adaptations of items across languages yield items that measure the same construct and are on the same scale. Thus, it is important to evaluate whether items function equivalently across forms while considering potential differences in test takers across forms.

Our results indicated that the groups of DLLs who received the English and Spanish forms differed on several important variables other than English language proficiency. Evidence indicates that many of these variables are associated with language development, e.g., DLLs whose mothers were born in the U.S. have greater English proficiency at kindergarten (Kim, Curby, & Winsler, 2014), children from higher SES homes often have stronger language skills (e.g., Hart & Risley, 1995), and there is a significant association between language skills and early numeracy, particularly for mathematics-related language knowledge (Purpura & Reid, 2016; Toll & van Luit, 2014). Clear differences between groups on variables besides underlying mathematics achievement and the grouping mechanism of interest (i.e., form language) highlight the threat of selection bias in carrying out traditional investigations of DIF and item/test bias. The modified RDD approach we employed resulted in improved covariate balance, but there were still statistically and practically meaningful differences in SES across the English and Spanish forms of the assessment. Taken together, these findings highlight the limitations of both approaches, although the modified RDD approach appears to account for selection bias to a greater extent than does the traditional approach.

Consistent with our hypotheses, fewer items showed evidence of DIF when using the modified RDD approach (3 items) than when using the traditional approach (7 items), and the nature of DIF also differed across approaches. When using the traditional approach, most items with significant DIF exhibited non-uniform DIF such that they were more strongly related to the underlying math construct for the English form of the assessment than they were for the Spanish form. In contrast, when using the RDD approach, all items exhibiting DIF exhibited uniform DIF such that they were

more difficult for children who received the English form. The latter finding suggests that Spanish-speaking DLLs who received the Spanish-language test accommodation may have had an unfair advantage over Spanish-speaking DLLs who were assigned to the English form. However, because only 3 items (out of 28 items examined) demonstrated significant DIF when controlling for selection bias, we do not believe there is substantial bias in the overall mathematics scores derived from the 2 test forms. Therefore, the Spanish-language translation of the routing and low form items of the ECLS-K kindergarten mathematics assessment appears to be a valid assessment accommodation.

Implications for assessment of academic and cognitive skills among DLLs

Traditional approach to DIF. When using the traditional approach, examination of item content did not reveal consistent patterns that would be indicative of substantial item bias across forms. Although 3 of 4 items assessing children's pattern recognition skills exhibited DIF, the item wording was identical for the 1 item that did not exhibit DIF, and pattern stimuli (e.g., shapes, sizes) did not appear to differ in a way that could explain presence or lack of DIF. Prior evidence indicates that parents and preschool teachers engage in pattern-related activities frequently and speak with children about patterns they observe in the world on a regular basis (Rittle-Johnson, Fyfe, Loehr & Miller, 2015). However, the frequency with which children are exposed to conversations about patterns in early home and school experiences may vary by socioeconomic status. Given that children who completed the English form of the assessment were more likely to come from high SES backgrounds, it is possible that differences in the association between performance on items assessing pattern knowledge and underlying mathematics abilities across test forms are reflective of differences in SES rather than mathematics knowledge. However, it should be noted that SES is significantly correlated with early numeracy achievement (e.g., Anders et al., 2012). Given that our analysis controlled for mathematics achievement, differences across forms in SES are at least partially controlled as well, limiting the potential for item bias related to the form of the assessment itself.

As was the case for pattern items, the 4 Number Sense, Properties, and Operations items did not display a consistent pattern of item content to explain DIF. The items primarily involved counting or numeral identification, and several items that did not exhibit DIF also required counting or numeral identification, often with identical stimulus language to those items that exhibited DIF. Among the items displaying DIF, only 1 had questionable item content that could be plausibly tied to DIF. Specifically, item 5 references a water fountain. Notably, the Spanish instructions allow for multiple words to be used related to water fountain, *la fuente de agua* and *bebedero*. Thus, it is possible that variability in the language used for this item across examiners could result in item performance reflecting knowledge of the concepts described in the item rather than mathematics ability, representing potential bias. However, there were 2 other items on the assessment that permitted multiple Spanish words to be used to represent *banana,* and neither of these items exhibited DIF.

RDD approach. The Measurement item that exhibited DIF required participants to point to an object that was shorter than a displayed target object. In contrast, the only other Measurement item required children to identify how long a pictured object was (with a scale provided). The other 2 items that exhibited DIF when using the RDD approach required children to perform addition and subtraction, and did not appear to differ in content from similar items that did not show DIF.

Overall, examination of item content revealed little evidence of bias across either approach. Notably, items exhibiting DIF under the RDD approach were more difficult for children taking the English version of the assessment. Although we do not believe our results show evidence of item/test bias, a consistent pattern of higher difficulty of English items would suggest that the cutpoint used to assign forms may have been too low, disadvantaging children who performed just above the cutpoint on the English language screener.

Within the ECLS-K sample, there was a large correlation between Spanish and English language proficiency, indicating that children with lower English proficiency also had lower Spanish proficiency. Additionally, although Robinson (2010) found evidence of effectiveness of the Spanish-language accommodation in the ECLS-K sample, other recent research suggests that Spanish-language accommodations are not always effective (Abedi et al., 2020). Abedi et al. reported poorer performance on a mathematics assessment for students assigned to receive a Spanish-language form of the test and suggest that factors other than English proficiency (e.g., language of instruction) play a role in DLLs' assessment performance. Thus, decisions regarding alternate-language assessment accommodations need to carefully consider myriad factors, rather than solely determining assessment language based on an English proficiency score.

Considerations for the modified RDD approach for examining DIF

Our analysis highlighted multiple limitations of the RDD approach in this context. First, results were sensitive to choice of bandwidth. Three additional items were identified as exhibiting DIF, and items identified as exhibiting DIF using the data-driven bandwidth sometimes did not exhibit DIF when a different bandwidth was used. This could be problematic for interpretation; however, the magnitude of DIF did not differ substantially across bandwidths.

Second, the English and Spanish groups differed on more than just mathematics assessment language at the OLDS English proficiency cutpoint. It is not possible to rule out SES as a source of differences in item responding. We note, however, that covariate imbalance can occur even in the context of a true randomized experiment.

Third, results of the RDD analysis revealed potential issues related to statistical power when using this approach. Local linear regression uses only a subset of cases that are close to the cutpoint to reduce bias in the estimate of the average treatment effect. Even though the total sample size was large, the effective sample size (the number of cases used in the RDD analysis) was not large. For several items flagged as exhibiting non-uniform DIF when using the traditional approach but not the RDD approach (e.g. items 1 and 24), DIF effect size estimates were larger in the RDD approach. Similarly, discrepancies in significance and effect sizes emerged across items within the RDD approach. For example, item 29 was flagged as exhibiting uniform DIF with a SIDP value of −0.04. Other items with comparable or larger sample sizes were not flagged as exhibiting uniform DIF despite having substantially greater effect sizes (e.g., item 33). Effect size estimates generated from the RDD approach should be interpreted with caution, as these effect size indices were not developed for use with an RDD approach to evaluating DIF.

Limitations

Although the results of this study have the potential to advance knowledge regarding measurement of academic skills among DLLs,

this study has some limitations. First, this dataset is old, as it represents the 1998–1999 cohort of ECLS-K. Item-level data are not available for the newer kindergarten cohort. However, the extent to which this is a noteworthy limitation is mitigated by the fact that items for mathematics assessments have not changed substantially in the last 20 years, and the mathematics test used in the ECLS-K 2011 cohort was based on the mathematics framework used for the 1996 National Assessment of Educational Progress. Relatedly, limitations in information describing the sample limit our ability to examine how heterogeneity within the sample influences the results. For example, it is possible that dialectical differences in Spanish used by children from different countries yield different patterns of item/test bias. However, there is no information about Spanish dialect spoken by children in the ECLS-K dataset, and the only information on country of origin is for children who were not born in the U.S. (approximately 12% of our sample). Therefore, we could not evaluate whether the validity of the Spanish-adapted items varied across dialect or country of origin. Finally, although our novel RDD approach for controlling selection bias showed promise when compared to traditional evaluations of DIF, there were many methodological limitations that warrant additional research. These limitations include: reliance on parametric assumptions to investigate non-uniform DIF; use of empirically-derived bandwidth selectors and robust confidence intervals that were not developed with the purpose of testing interactions or for use with categorical data; and use of effect sizes that have not been studied for use within an RDD framework.

Conclusions

Overall, the results of this study demonstrated little evidence of item or test bias on the ECLS-K kindergarten mathematics assessment, providing evidence that the Spanish-language form is a valid accommodation for DLLs that results in equitable, unbiased assessment. Moreover, we observed that children assigned to different assessment forms differed on characteristics other than mathematics achievement and language form/English proficiency. This introduces selection bias that needs to be controlled when evaluating DIF and subsequently making inferences about item/test bias. Although our approach was not without limitations, it reduced the selection bias present and led to different and potentially more accurate conclusions regarding DIF and item/test bias. Ensuring accurate and fair assessment for all children is a core tenet of the Every Student Succeeds Act (2016). Additional research is needed to develop and validate approaches for ensuring equitable and unbiased assessment.

Authors' contributions

J. Marc Goodrich: Conceptualization, writing – original draft, writing – review and editing, funding acquisition, project administration. Natalie Koziol: Conceptualization, methodology, writing – original draft, writing – review and editing, formal analysis, project administration, funding acquisition. HyeonJin Yoon: Methodology, writing – original draft, writing – review and editing.

Acknowledgments

This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under NSF award NSF-DRL #1,749,275. Opinions reflect those of the author and do not necessarily reflect those of AERA or NSF.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecresq.2021.06.001.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*, 231–257.
Abedi, J., Zhang, Y., Rowe, S. E., & Lee, H. (2020). Examining effectiveness and validity accommodations for English language learners in mathematics: An evidence-based computer accommodation decision system. *Educational Measurement: Issues and Practice, 39*(4), 41–52.
Anders, Y., Rossbach, H.-G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., & von Maurice, J. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly, 27*, 231–244.
Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness, 5*, 43–82.
Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2018). rdrobust: Robust data-driven statistical inference in regression-discontinuity designs. *R package version 0.99.4*.
Cattaneo, M.D., .Jansson, M., & Ma, X. (2019). rddensity. R package version 1.0.
Cattaneo, M., Idrobo, N., & Titiunik, R. (2020). *A Practical Introduction to Regression Discontinuity Designs: Foundations* (Elements in Quantitative and Computational Methods for the Social Sciences). Cambridge. *Cambridge University Press*. https://doi.org/10.1017/9781108684606.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Duncan, S. E., & De Avila, E. (1998). *PreLAS 2000*. Monterey, CA: CTB/McGraw-Hill.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
Every Student Succeeds Act, 20 U.S.C. § 6301 (2015).
Farrington, A. L., Lonigan, C. J., Phillips, B. M., Farver, J. M., & McDowell, K. D. (2015). Evaluation of the utility of the *Revised Get Ready to Read!* for Spanish-speaking English-language learners through differential item functioning analysis. *Assessment for Effective Intervention, 40*, 216–227.
Fletcher, J. M., & Miciak, J. (2017). Comprehensive cognitive assessments are not necessary for the identification and treatment of learning disabilities. *Archives of Clinical Neuroscience, 32*, 2–7.
Ginsburg, H. P., & Baroody, A. J. (1990). *The test of early mathematics (TEMA-2)*. Austin, TX Pro-Ed.
Goodrich, J. M., Lonigan, C. J., & Alfonso, S. V. (2019). Measurement of early literacy skills among monolingual English-speaking and Spanish-speaking language-minority children: A differential item functioning analysis. *Early Childhood Research Quarterly, 47*, 99–110.
Goodrich, J. M., & Namkung, J. M. (2019). Correlates of reading comprehension and word-problem solving skills of Spanish-speaking dual language learners. *Early Childhood Research Quarterly, 48*, 256–266.
Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system*. Circle Pines, MN: AGS.
Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children*. Baltimore, MD: Brookes Publishing.
Huttenlocher, J., & Levine, S. C. (1990). *The primary test of cognitive skills (PTCS)*. New York: CTB/McGraw Hill.
Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*, 615–635.
Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).
Kamata, A., Chaimongkol, S., Genc, E., & Bilir, K. (2005). Random-effect differential item functioning across group unites by the hierarchical generalized linear model April. Montreal, Canada: Paper presented at the annual meeting of the American Educational Research Association.
Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology, 100*, 851–868.
Kieffer, M. J., & Thompson, K. D. (2018). Hidden progress of multilingual students on NAEP. *Educational Researcher, 47*, 391–398.
Kim, Y. K., Curby, T. W., & Winsler, A. (2014). Child, family, and school characteristics related to English proficiency development among low-income, dual language learners. *Developmental Psychology, 50*, 2600–2613.
Lakin, J. M., Elliott, D. C., & Liu, O. L. (2012). Investigating ESL Students' performance on outcomes assessments in higher education. *Educational and Psychological Measurement, 72*, 734–753.
Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the National Assessment of Educational Progress. *International Journal of Testing, 8*, 14–33.
Markwardt, F. C., Jr. (1989). Peabody individual achievement test-revised (PIAT-R). Circle Pines, MN: American Guidance Services, Inc.
Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14*, 160–179.
McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*, 698–714.
Meade, A. W. (2010). A taxonomy of effect size measures of the differential functioning of items and scales. *Journal of Applied Psychology, 95*, 728–743.

Miciak, J., Taylor, W. P., Denton, C. A., & Fletcher, J. M. (2015). The effect of achievement test selection on identification of learning disabilities within a patterns of strengths and weaknesses framework. *School Psychology Quarterly, 30*, 321–334.
Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., et al. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. *Educational Researcher, 44*, 278–292.
Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide. Eighth edition* 1998. Los Angeles, CA: Muthén & Muthén.
Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly, 4*, 149–164.
Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessments: Review and recommendations. *Educational Measurement, 19*, 5–15.
Purpura, D. J., & Reid, E. E. (2016). In *Mathematics and language: Individual and group differences in mathematical language skills in young children: 36* (pp. 259–268). Early Childhood Research Quarterly.
R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Retrieved from: https://www.R-project.org/ .
Rittle-Johnson, B., Fyfe, E. R., Loehr, A. M., & Miller, M. R. (2015). Beyond numeracy in preschool: Adding patterns to the equation. *Early Childhood Research Quarterly, 31*, 101–112.
Robinson, J. P. (2010). The effects of test translation on young English learners' mathematics performance. *Educational Researcher, 39*, 582–590.
Samson, J. F., & Lesaux, N. K. (2009). Language-minority learners in special education: Rates and predictors of identification for services. *Journal of Learning Disabilities, 42*, 148–162.
Snetzler, S., & Qualls, A. L. (2000). Examination of differential item functioning on a standardized achievement battery with limited English proficient students. *Educational and Psychological Measurement, 60*, 564–577.
Solano-Flores, G., & Li, M. (2008). Examining the dependability of academic achievement measures for English language learners. *Assessment for Effective Intervention, 33*, 135–144.
Spinelli, C. G. (2008). Addressing the issue of cultural and linguistic diversity and assessment: Informal evaluation measures for English language learners. *Reading & Writing Quarterly, 24*, 101–118.
Sullivan, A. L. (2011). Disproportionality in special education identification and placement of English language learners. *Exceptional Children, 77*, 317–334.
Toll, S. W. M., & van Luit, J. E. H. (2014). The developmental relationship between language and low early numeracy skills throughout kindergarten. *Exceptional Children, 81*, 64–78.
U.S. Department of Education, National Center for Education Statistics. (2000). ECLS-K restricted-use base year data files and electronic codebook *(NCES 2000-097)*. Washington, DC: Author.
U.S. Department of Education, National Center for Education Statistics. (2002). ECLS-K psychometric report for kindergarten through 1st grade *(NCES 2002-005)*. Washington, DC: Author.
Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*, 352–360.
Woodcock, R. W., & Bonner, M. (1989). *The Woodcock-Johnson tests of achievement-revised*. Itasca, IL: Riverside.