CEMETR-2021-05
MAY 2021

CEME Technical Report

The Center for Educational Measurement and Evaluation

The Examining Potential Differential Item Functioning of Placements on *GOLD*® Developmental Progressions across Gender, Race, and Primary Language Spoken in the Home

Richard G. Lambert

David E. Lynn

A PUBLICATION OF THE CENTER FOR EDUCATIONAL MEASUREMENT AND EVALUATION

CEME
The Center for Educational Measurement and Evaluation
College of Education
UNC CHARLOTTE

Examining Potential Differential Item Functioning of Placements on *GOLD*® Developmental Progressions across Gender, Race, and Primary Language Spoken in the Home

Richard G. Lambert

David E. Lynn

Center for Educational Measurement and Evaluation

University of North Carolina at Charlotte

May, 2021

Abstract

The current study focused on gathering validity evidence for the use of an authentic formative assessment called Teaching Strategies GOLD (*GOLD*®). Differential item functioning analyses were used with a large sample of young children (n = 32,063) to gather evidence that teachers use *GOLD*® developmental progressions in a consistent manner across sub-groups of children based on gender, race, and primary language spoken in the home. Nominal item difficulty classifications (Easy, Average, or Difficult) were the same for each group across nearly all sub-group comparisons and developmental progressions. For all progressions, across all developmental domains and across all group comparisons, the DIF magnitudes were negligible (≤.43). This study provided evidence that teachers are generally using *GOLD*® developmental progressions in a similar manner across the sub-groups of young children compared in this study.

Keywords: Differential item functioning, formative assessment, early childhood

Examining Potential Differential Item Functioning of Placements on *GOLD*® Developmental Progressions across Gender, Race, and Primary Language Spoken in the Home

Standard 3.0 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014, p. 63) makes it clear that "All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct irrelevant variance and to promote the validity of interpretations for intended uses for all examinees in the intended population." The Standards identify the following potential sources of construct-irrelevant variance in assessment scores: a.) assessment content, b.) assessment context, c.) response process, and d.) opportunity to learn. Validity evidence based on the integrity and fairness of the response process is one of the most important components of an argument to support the meaningfulness, usefulness, and appropriateness of the interpretations of assessment scores (Messick, 1995; AERA, APA, & NCME, 2014).

The current study focused on gathering validity evidence for the use of an authentic formative assessment called Teaching Strategies GOLD (*GOLD*®). *GOLD*® was designed and validated to provide teachers with information about the developmental trajectories of young children from birth to third grade. Teachers use this information for instructional planning, and to support the growth, development, and learning of young children. The task of establishing evidence for the validity of an authentic formative assessment such as *GOLD*® presents unique challenges. Unlike direct summative assessments, authentic formative assessment measures for young children have no standardized test administration or response protocol. Rather, the information they provide is wholly dependent upon a complex response process that involves an interaction between teacher and child behaviors. Within this complex cycle, the child component of the response process resides in the evidences (child work samples, anecdotes, etc.) elicited by the teacher, classroom activities, interactions with other children, and the classroom environment itself. The teacher component of

the response process resides in a teacher's ability to recognize, select, record, and analyze representative evidences of child developmental progress. Finally, the teacher uses the evidence gathered and analyzed to support placements on GOLD® developmental progressions.

Heritage (2013) described this process as a continuous cycle of activities embedded within instructional activities in the classroom. This cycle includes the following phases: 1.) understanding what is next for a child and setting learning goals, 2.) defining and understanding criteria that will indicate progress toward the next level of development, 3.) gathering evidences of growth, development and learning, 4.) analyzing and interpreting these evidences, 5.) making placements on developmental progressions, and 6.) adapting instruction to support the unique needs of the individual child (Heritage, 2013). When a child masters a particular task, this cycle repeats itself as the child moves toward the next level on a specific developmental progression. This process is ongoing and plays out simultaneously across multiple developmental domains.

If teachers misunderstand or misapply any of the steps in the assessment cycle, they can introduce construct-irrelevant variance into the assessment scores. For example, this process will only include evidences that are representative of a child's true abilities when a teacher has mastered the complex set of tasks involved in all phases of the assessment cycle. Teachers have to understand fully how each progression corresponds to learning objectives in the applicable curricular model and child learning standards that govern their work. They have to understand how to recognize valid evidences that relate to the behavioral anchors on the progressions, match those evidences accurately with the appropriate levels on the progressions, and determine when they have sufficient evidence to support finalized placements on the progressions (Lambert, 2020).

Given the multi-tiered nature of this complex response process, and given the multiple possibilities for teachers to introduce construct-irrelevant variance that can impact the fairness of assessment scores, authentic formative assessments such as *GOLD*® qualify as what Engelhard

(2002; Engelhard & Wind, 2018) has called rater-mediated assessments. Therefore, *GOLD*® assessment scores are mediated by, or emerge through, the judgement of the teacher (Engelhard & Wind, 2018). Within this process, teachers analyze and interpret evidences of child progress that were elicited by the instructional process, not the developmental progressions. Then the teacher, not the child, responds to the stimuli contained within the developmental progressions by making ratings. However, interpretation of the resulting assessment scores has consequences for the child, not the teacher. Ideally, the scores, when properly interpreted, lead to decisions about instructional goals, strategies, and support provided to the child. Therefore, the scores yielded by measures such as *GOLD*® lead to meaningful, useful, and actionable conclusions for all children in the classroom only when teachers can apply the rating scales in a reliable and fair manner across all subgroups of children.

Furthermore, the standards specifically require evidence that assessment scores can be interpreted in the same way, or have the same essential meaning, across all subgroups of children (AERA, APA, & NCME, 2014). The Standards, when applied to rater-mediated assessments for young children, require rater effects on the response process be minimized such that an assessment system is fair to all subgroups of children. Therefore, variability in assessment scores should correspond to variability in true differences in child ability and be free of the influence of any other teacher or child characteristics.

However, teachers can introduce construct-irrelevant variance into a set of scores derived from an authentic formative assessment by applying the complex response process in different ways across sub-groups of children. This can result in a lack of fairness within the assessment process; and fairness in testing is a fundamental principle of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014). The standards define fairness in terms of the absence of measurement bias. Any introduction of construct-irrelevant variance that results in assessment

scores taking on different meaning for varying sub-groups of children is evidence of a lack of fairness. The Standards identify differential item functioning (DIF) analyses as a viable strategy for contributing to an argument for the fairness of the information provided by assessment scores under specific conditions of use (AERA, APA, & NCME, 2014, p. 51).

Specifically, DIF analyses can address these issues by providing evidence that the response process does not advantage any particular subgroup of children. DIF analyses separate performance differences on specific items from individual differences on the measured latent construct. Focal and reference categories of examinees are identified to investigate DIF. Researchers examine the focal group to gather evidence to demonstrate that the assessment process is fair to the members of that particular subgroup. The reference group is typically the majority group and the presumption is that the assessment process is generally fair for them. Researchers compare the focal and reference groups by examining their item responses after first matching them on their overall ability estimates. Respondents are, therefore, matched based on the estimated quantity of the latent construct under investigation they are presumed to possess (Walker, 2011). In a Rasch modeling context, DIF can be thought of as differences between members of the focal and reference groups in the conditional probability of a particular item response.

Non-Rasch DIF procedures investigate the focal and reference groups for interactions between sample characteristics and item responses, while holding overall ability levels of the test takers constant. The analytical process estimates overall ability for each participant from total performance across the full item set. Therefore, such DIF statistical procedures answer the following research question: Do test takers in the focal and reference groups who have the same overall ability levels respond in similar ways to each item? Evidence of DIF can lead to further investigation of item content, item interpretation by respondents, rater characteristics and behavior, construct validity of the items and measure, and potential test bias. However, it is important to

distinguish item bias from item impact. Item impact in the context of this study refers to true score differences between subgroups of children in their average developmental level on the latent construct in question. Item bias in the context of this study means differences between subgroups of children related to construct-irrelevant sources of variance that can lead to inaccurate interpretations of the information that particular *GOLD*® developmental progressions provide. One possible source of construct-irrelevant variance could be the fairness of the content of the behavioral anchors on a particular rating scale when teachers apply that content to different subgroups of children. Another such factor could be teacher or rater effects, suggesting that teachers apply a particular rating scale in an unfair manner when rating different subgroups of children.

From a test validity perspective, item functioning needs to be investigated for possible interactions with characteristics of subgroups of test takers to ensure that the test is measuring the same construct of interest, and only the construct of interest, in the same way for all subgroups (Badia, X., Prieto, L., and Linacre, J. M., 2002). This broad principal of validity extends to teacher ratings and authentic formative assessments and is an important part of the validity argument to support the use of the *GOLD*® assessment system.

DIF Studies in Early Childhood Education

Previous research examined validity evidence to support the use of *GOLD*® for assessing children with disabilities and those for whom English is not their first language. Assessment information was collected on three-, four-, and five -year-old children at the fall (n = 79,324), winter (n = 132,693), and spring (n = 50,558) checkpoints. DIF analyses indicated that in general, teachers' ratings were similar for children of similar abilities, regardless of their subgroup membership. The majority of items in *GOLD*® displayed little or no DIF with the exception of one item, "uses conventional grammar" (Kim, Lambert, & Burts, 2013). This progression was somewhat higher in difficulty level for children who are not native English speakers.

Associations of teacher ratings with child demographics (e.g., age, gender, disability status, and English language status) and classroom composition characteristics (e.g., class mean age and percentage English language learners (ELLs), children with disabilities, and males) were examined with a sample of 21,592 children ages 12 months through 59 months. Using three-level growth curve modeling, findings indicated that teachers' *GOLD*® ratings were associated in anticipated directions for both child and classroom characteristics. Children with disabilities began the year behind their typically developing peers and grew more slowly throughout the year. Girls demonstrated advantages in some developmental domains over boys. ELLs were rated lower at the beginning of the year but exhibited somewhat faster growth rates than native English-speakers. Variance associated with potential rater effects (i.e., between teacher differences in how teachers used the *GOLD*® to rate the children in their classrooms) ranged from 16% to 25%, which is considerably lower than reported in some other studies (Lambert, Kim, & Burts, 2013).

In two studies that examined social-emotional assessment for young children, DIF procedures were used to identify potential bias in specific items. For example, in a study of (n = 7,179) responses of 3-year-olds on a social emotional questionnaire, differences were found on specific items between boys and girls with similar socio-emotional levels (Vaezghasemi et al., 2020). A study in the United States found potential bias on an emotional behavioral assessment when screening (n = 1,985) first graders who were of Latin American heritage (Lambert, Garcia, January, & Epstein, 2018). In both males and females, DIF was found between English Language Learners (ELL) and non-ELL participants. Researchers in both studies concluded that the DIF identified did not impact the overall scores provided by the instrument.

Three studies were identified that applied DIF procedures to uncover potential language bias in assessments of young children. When examining preschool results for the Preschool Language Scale-4 (PLS-4) assessment of Head Start participants (n=440), specific questions were found to be

more challenging for Latinx students while others were more difficult for European-American students (Qi & Marley, 2009). In another study of (n=210) prekindergarten students, latent class analysis was used to compare results on the Peabody Picture Vocabulary Test-Third Edition (PPVT-III) and the Expressive Vocabulary Test (EVT) between lower ability and higher ability children. The results uncovered potential bias, related to one particular response strategy (Webb, Cohen, & Schwanenflugel, 2008). In a study of 4-year-old dual language learners (DLL) (n = 133,732) using *GOLD©*, researchers found that most items functioned similarly when comparing three distinct latent subgroups of DLL children (Kim, Lambert, Durhams, & Burt, 2018). Just two of the *GOLD©* developmental progressions displayed intermediate or large DIF when assessing language and literacy developmental progress.

Kim, Lambert, & Burts (2014) used DIF analysis to examine data from the *GOLD©* assessment system in a study focused on children ages 3 to 5 years (n = 362,575) with regards to DLL and children with disabilities. A majority of the assessment items displayed little to no DIF. Only a few items exhibited DIF for children with disabilities and DLL. The study concluded that while DIF analyses generally provided validity evidence in support of the use of *GOLD©* with young children, the validity of the assessment information could be enhanced through future teacher training related to using language related progressions with these specific sub-groups of young children. Kim, Lambert, and Burts (2014) also highlighted the importance of teacher recognition of their own potential biases concerning the cultural backgrounds and learning needs of their students.

The Current Study

The current study was conducted to determine if teachers displayed different patterns of item response when rating specific subgroups of children. In this study, "item response" was defined as placements made by teachers on *GOLD®* developmental progressions. Given that *GOLD®* is an authentic formative assessment, "item response" does not refer to direct answers from children in

response to item stimuli. Rather, it refers to the placements on *GOLD*® developmental progressions assigned to each child by teachers based on evidences produced by the child in the course of normal instructional activities. Specifically, different patterns of response were defined in terms of DIF. DIF focuses on whether test items function in different ways when teachers rate different subgroups of children. Ideally, item functioning is invariant to specific aspects of the children assessed, such as demographic, classroom environmental, or teacher characteristics that are considered irrelevant to measuring the underlying ability or latent construct that is the focus of the assessment.

In order to interpret the findings from this study properly, it is important to recognize exactly what information DIF analyses can and cannot provide. DIF in the context of this study is present when members of different subgroups of children, for a particular *GOLD*® developmental progression, have different probabilities of placement at specific levels on the rating structure, after the overall developmental level of each child on the domain in question has been accounted for. Furthermore, DIF in the context of this study can suggest the possibility that teachers may be biased in their ratings of children in the focal group relative to how they rate the reference group. However, evidence of DIF can suggest a need to investigate potential biases but does not guarantee the presence of item bias.

All previous *GOLD*® DIF studies utilized data yielded by an older version of *GOLD*®, the birth to kindergarten version of the measure. The Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) indicate that if revisions influence the interpretation of scores, assessment developers have the responsibility to offer new validity evidence in support of the use of the new version of a measure. Specifically, Standards 4.24 and 4.25 require assessment developers to provide users with guidelines concerning the appropriate uses of, and the comparability of scores from, earlier and revised version of a measure. Therefore, researchers and the developers of *GOLD*® should provide users with a strong validity argument for the use of the revised *GOLD*® measure

relative to its intended purposes. The current study sought to extend previous research and establish further validity evidence for the birth to third grade version of *GOLD*®.

Specifically, the current study sought to examine potential evidence for DIF in the placements teachers make on *GOLD*® developmental progressions based on gender, primary language spoke in the home, and race / ethnicity. Rasch modeling was used to produce composite interval level scale scores for each domain of development, item difficulties, and child ability estimates that are useful for statistical comparisons in general, and testing for DIF in particular. Furthermore, this study sought to further the validity argument for the use of *GOLD*® with young children by extending and updating previous DIF research through a focus on data collected using the latest version of the measure.

Methods

National Sample

The participants for this study came from the national norm sample used for the 2020 *GOLD*® technical report. For more details on the sample, see Lambert (2020). The sample for this study was drawn from early childhood programs and schools that use *GOLD*® to document the developmental progress of children during at least three assessment periods: fall, winter and spring of the 2018-2019 school year. For each age / grade group (colored band) from birth to kindergarten, the researchers selected a stratified random sample of 5,000 children. The strata were formed using the U.S. Census Bureau subgroups for race / ethnicity in an effort to represent each subgroup in its proper proportion of the U.S. population of children. This process removed the effect of clustering groups of children within their rater or teacher from the data, as whole classrooms of children were not selected. The *GOLD*® assessment system has not been adopted for use in first, second, and third grades at the same rate as it has for the birth to kindergarten years. Therefore, all available data from

children in first, second, and third grade was used. This process resulted in a total norm sample of 32,063 children.

These children from birth to kindergarten received educational services in centers or schools that were located in the all 50 U.S. states and the District of Columbia, Peurto Rico, and schools around the world that serve the families of U.S. Military personnel. The children for the first, second, and third grade sample came from five states: Alabama, Colorado, Georgia, Hawaii, and New Jersey. Therefore, the sample includes a rich geographic diversity from all regions of the U.S. and includes all the types of settings where young children are served (Head Start, private pre-kindergarten, publicly funded preschool programs, and public schools).

The norm sample was very evenly balanced by gender (boys=48.52%, girls=51.48%). Children with an IFSP or IEP comprised 7.24% of the sample. A total of 33.19% of the norm sample qualified for the National School Lunch Program (free or reduced price lunch) as reported by their teacher. This figure is likely an underestimate as all teachers may not accurately report this information. As is, this number reflects an under-representation of these children as the national figure is greater than 50% (at 51.8%). The primary language spoken in the home was distributed as follows: English (78.57%), Spanish (14.47%), and other languages (6.96%). The race / ethnicity of the children in the sample was as follows, shown here with the 2018 national census estimates for U.S. children in parentheses: a.) White – 49.69% (49.39%), b.) African American – 13.89% (13.76%), c.) Native American / Pacific Islander – 1.22% (1.04%), d.) Asian – 4.13% (5.04%), e.) Multiracial / other – 4.64% (4.70%), and f.) Hispanic – 26.43% (26.07%). These values indicate that the sample was approximately nationally representative for all race / ethnicity groups of American children.

Data Analysis

First, interval level measures were created for each of the *GOLD*® domains of development using the Rasch partial credit model (PCM; Masters, 1982), as operationalized through the Winsteps

software (version 4.6.2.1) (Linacre, 2020). Specifically, the resulting measures were created using the Linacre partial credit model, an extension of the Rasch rating scale model (Andrich, 1978) and Masters partial credit model (Masters, 1982). This step allows researchers to access scale scores for each participating child that were measured on the same interval scale. This method also provided invaluable diagnostic information about the fit of the data yielded by each progression to the Rasch measurement model. A separate Rasch analysis was conducted for each of the six domains of development. The Rating Scale (RSM; Andrich, 1978, Bond & Fox, 2001) and the PCM are the two most widely used Rasch model for polytomous response data. The RCM is useful when all items share the same rating scale. In cases where each item has its own rating scale structure, the PCM is the appropriate model to apply. The PCM, rather than the RCM, was chosen because the *GOLD*® progressions do not share the same rating scales (i.e., the number of rating scale categories and category labels varies across progressions).

Specifically, 11 *GOLD*® items include a 0-9 scale, 6 items include a 0-11 scale, 17 items include a 0-13 scale, 25 items include a 0-15 scale, and one item includes a 0-19 scale. For each item, the 0 category represents "Not Yet" and the highest category represents abilities beyond the highest behavioral anchor. In addition, the behavioral anchors for each step on these rating scales are unique for each progression. For more details about these analyses, see the 2020 *GOLD*® technical report (Lambert, 2020).

The following procedures were used to test for the presence of DIF and to examine the magnitude of DIF. First, the difficulty level for each progression was estimated for both the focal and reference groups. The DIF contrast statistic, the difference between the paired Rasch item difficulty estimates for each item, was calculated as the simple difference between the item difficulty estimates for the focal and reference groups. Next, Rasch-Welch *t* tests were examined for statistical significance. These statistics report whether item difficulty estimates for the two groups, across each

of the items, are the same except for measurement error. These tests evaluate a null hypothesis that the DIF contrast value is zero against an alternative hypothesis that the DIF contrast is not equal to zero. Next, the Mantel-Haenszel χ^2 statistics produced by the Winsteps software package (Linacre, 2020) were also used to examine evidence of potential DIF. These tests examine a null hypothesis of no DIF by producing a probability of obtaining differences between the focal and reference groups as large as or larger than those obtained, given that there is no DIF. The groups are stratified into matching ability levels and their relative performance on each item is quantified. Overall ability estimates for each respondent are estimated based on the total scores on the measure. An alpha level of .05 was used for all comparisons.

In addition to statistical significance testing, examination of the magnitude of the DIF contrast, or difference in item difficulty estimates between the focal and reference groups, is critical given the sensitivity of both the Rasch-Welch *t* test and the Mantel-Haenszel χ^2 to small differences. This is especially important in large sample size studies like this one where statistical significance is easy to obtain even when the observed differences do not have practical implications. Magnitude of the DIF contrast was determined according to the criteria set forth by Zwick, Thayer, and Lewis (1999). If both the *t* and χ^2 statistics were statistically significant, and the magnitude of the DIF contrast was less than .43, the DIF magnitude was considered negligible. If both the *t* and χ^2 statistics were statistically significant, and the magnitude of the DIF contrast was greater than or equal to .43 and less than .64, the DIF magnitude was considered intermediate. If both the *t* and χ^2 statistics were statistically significant, and the magnitude of the DIF contrast was greater than .64, the DIF magnitude was considered large. To further aid interpretation, items with difficulty measures were classified as follows: below −.5 were considered easy, −.5 to .5 were considered average, and those with values above .5 were considered difficult. For each domain of development, the average item difficulty was set to a value of zero.

Specifically, this study posed the following research questions:

- Is there evidence of DIF regarding the placements made by teachers on *GOLD*® developmental progressions for male versus female children?
- Is there evidence of DIF regarding the placements made by teachers on *GOLD*® developmental progressions for children living in homes where English versus Spanish is the primary language spoken?
- Is there evidence of DIF regarding the placements made by teachers on *GOLD*® developmental progressions for white versus non-white children?

Results

For the first research question, male children were the focal group and female students were the reference group. This decision was made because the overwhelming majority of teachers are females. For the second research question, children living in homes where Spanish is the primary language were the focal group and children living in homes where English is the primary language were the reference group. This decision was made because most early childhood teachers in the United States are native English speakers and English is the most frequently spoken language. For the third research question, non-white children were the focal group and white children were the reference group. White children were the reference group because the majority of the early childhood workforce consists of white teachers.

Research Question 1 - DIF Analyses by Gender

Social Emotional

For five of the nine progressions within the Social Emotional domain (1.a, 1.c, 2.c, 3.a, and 3.b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one of the progressions within the Social Emotional domain (2.b), only the Mantel-Haenszel χ^2 statistic was statistically significant ($p < .05$). For the remaining three progressions within the Social

Emotional domain (1.b, 2.a, and 2.d), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The estimated item difficulties for males and females were identical for six of the nine progressions (1.a, 1.c, 2.b, 2.c, 3.a, and 3.b). The differences in estimated item difficulties for all progressions were very small (.05 to .11). As seen in Table 1, the item difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all nine progressions. Similarly, the DIF magnitudes were considered negligible for all nine progressions.

Physical

For all five progressions within the Physical domain (4, 5, 6, 7.a, and 7.b), both the Rasch-Welch *t* and the Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The differences in estimated item difficulties for all five progressions were small (-.23 to .25). As seen in Table 2, the item difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across four of the five progressions. One progression (6), met the criteria for difficult for females, and the criteria for average for males. This progression focuses on gross-motor manipulation skills and these results suggest it may be slightly more difficult for teachers to move girls to the next level on the progression as compared to boys. However, the DIF magnitudes were considered negligible for all five progressions.

Language

Across four of the eight progressions within the Language domain (9.a, 9.c, 10.a, and 10.b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one of the progressions within the Language domain (8.b), only the Mantel-Haenszel χ^2 statistics was statistically significant ($p < .05$). For the remaining three progressions within the Language domain (8.a, 9.b, and 9.d), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). Across four of the eight progressions, the estimated items difficulties for males

and females were identical (9.a, 9.c, 10.a, and 10.b). The differences in estimated item difficulties for all progressions were very small (-.05 to .05). As seen in Table 3, the item difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all eight progressions. Similarly, the DIF magnitudes were considered negligible for all eight progressions.

Cognitive

Across four of the 10 progressions within the Cognitive domain (11.c, 12.a, 12.b, and 14.a), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining six progressions within the Cognitive domain (11.a, 11.b, 11.d, 11.e, 13, and 14.b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The estimated item difficulties for males and females were identical across four of the 10 progressions (11.c, 12.a, 12.b, and 14.a). The differences in estimated item difficulties for all progressions were very small (-.13 to .11). As seen in Table 4, the item difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all 10 progressions. Similarly, the DIF magnitudes were considered negligible for all 10 progressions.

Literacy

For six of the 16 progressions within the Literacy domain (15.d, 17.a, 17.b, 18.a, 19.a, and 19.c), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. Across four of the progressions within the Literacy domain (15.b, 15.c, 18.b, and 18.c), only the Mantel-Haenszel χ^2 statistic was statistically significant ($p < .05$). For one of the progressions (18.e), only the Rasch-Welch *t* statistic was statistically significant ($p < .05$). For the remaining five progressions within the Literacy domain (15.a, 16.a, 16.b, 18.d, and 19.b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The estimated item difficulties for males and females were identical for nine of the 16 progressions (15.b, 15.c, 17.a, 17.b, 18.a, 18.b, 18.c, 19.a, and 19.c). The differences in estimated item difficulties for all

progressions were very small (-.05 to .12). As seen in Table 5, the item difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all 16 progressions. Similarly, the DIF magnitudes were considered negligible for all 16 progressions.

Mathematics

For seven of the 12 progressions within the Mathematics domain (20.a, 20.b, 21.b, 22.a, 22.b, 22.c, and 23), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining five progressions within the Mathematics domain (20.c, 20.d, 20.e, 20.f, and 21.a), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The estimated item difficulties for males and females were identical for seven of the 12 progressions (20.a, 20.b, 21.b, 22.a, 22.b, 22.c, and 23). The differences in estimated item difficulties for all progressions were very small (-.07 to .07). As seen in Table 6, the item difficulty classifications (Easy, Average, or Difficult) were the same for both males and females across all 12 progressions. Similarly, the DIF magnitudes were considered negligible for all 12 progressions.

Research Question 2 - DIF Analyses by Primary Language

Social Emotional

For three of the nine progressions within the Social Emotional domain (1.a, 1.b, and 2.b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining six progressions within the Social Emotional domain (1.c, 2.a, 2.c, 2.d, 3.a, and 3.b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The estimated item difficulties for children from homes where English is the primary language (EPL) and children from homes where Spanish is the primary language (SPL) were identical for 1.b. The differences in estimated item difficulties for all progressions were very small (-.15 to .18). As seen in Table 7, the item difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across eight of the nine progressions. The only exception to this finding was for

2.b. The estimated difficulty for EPL children was .49 which is at the top of the Average range and the estimated difficulty for SPL children was .51 which is at the bottom of the Difficult range, resulting in a very small difference. The DIF magnitudes were considered negligible for all nine progressions.

Physical

For one progression within the Physical domain (4), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one progression (7.a), only the Mantel-Haenszel χ^2 statistics was statistically significant ($p < .05$). For the remaining three progressions within the Physical domain (5, 6, and 7.b), both the Rasch-Welch *t* and the Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The differences in estimated item difficulties for all five progressions were very small (-.10 to .18). As seen in Table 8, the item difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all five progressions. Similarly, the DIF magnitudes were considered negligible for all five progressions.

Language

For three of the eight progressions within the Language domain (8.a, 9.a, and 10.a), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one of the progressions within the Language domain (9.d), only the Mantel-Haenszel χ^2 statistic was statistically significant ($p < .05$). For the remaining four progressions within the Language domain (8.b, 9.b, 9.c, and 10.a), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The differences in estimated item difficulties for all progressions were small (-.20 to .23). As seen in Table 9, the item difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all eight progressions. Similarly, the DIF magnitudes were considered negligible for all eight progressions.

Cognitive

For two of the 10 progressions within the Cognitive domain (13 and 14.b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For two of the 10 progressions (11.e and 14.a), only the Mantel-Haenszel χ^2 statistic was statistically significant ($p < .05$). For the remaining six progressions within the Cognitive domain (11.a, 11.b, 11.c, 11.d, 12.a, and 12.b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The differences in estimated item difficulties for all progressions were small (-.10 to .27). As seen in Table 10, the item difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all 10 progressions. Similarly, the DIF magnitudes were considered negligible for all 10 progressions.

Literacy

For three of the 16 progressions within the Literacy domain (15.c, 17.b, and 18.b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For two of the progressions within the Literacy domain (15.b and 18.d), only the Rasch-Welch *t* statistic was statistically significant ($p < .05$). For the remaining 11 progressions within the Literacy domain (15.a, 15.d, 16.a, 16.b, 17.a, 18.a, 18.c, 18.e, 19.a, 19.b, and 19.c), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The estimated item difficulties for both groups were identical for three of the 16 progressions (15.c, 17.b, and 18.b). The differences in estimated item difficulties for all progressions were small (-.12 to .22). As seen in Table 11, the item difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all 16 progressions. Similarly, the DIF magnitudes were considered negligible for all 16 progressions.

Mathematics

For six of the 12 progressions within the Mathematics domain (20.a, 20.c, 20.d, 20.f, 22.c, and 23), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining six progressions within the Mathematics domain (20.b, 20.e, 21.a, 21.b, 22.a, and 22.b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .01$). The estimated item difficulties for EPL and SPL children were identical for three of the 12 progressions (20.a, 20.c, and 23). The differences in estimated item difficulties for all progressions were very small (-.10 to .13). As seen in Table 12, the item difficulty classifications (Easy, Average, or Difficult) were the same for both EPL and SPL children across all 12 progressions. Similarly, the DIF magnitudes were negligible for all 12 progressions.

Research Question 3 - DIF Analyses by Race / Ethnicity

Social Emotional

For one of the nine progressions within the Social Emotional domain (2.b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For two of the nine progressions within the Social Emotional domain (1.a and 2.c), only the Mantel-Haenszel χ^2 statistics were statistically significant ($p < .01$). For the remaining six progressions within the Social Emotional domain (1.b, 1.c, 2.a, 2.d, 3.a, and 3.b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The estimated item difficulties for white and non-white children were identical for 2.b and 2.c. The differences in estimated item difficulties for all progressions were small (-.21 to .14). As seen in Table 13, the item difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all nine progressions. Similarly, the DIF magnitudes were negligible for all nine progressions.

Physical

For two progressions within the Physical domain (4 and 7.a), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For two progressions (5 and 6), only the Mantel-Haenszel χ^2 statistics was statistically significant ($p < .05$). For 7.b, both the Rasch-Welch *t* and the Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The estimated item difficulties for white and non-white children were identical across four of the five progressions (4, 5, 6, and 7.a). The differences in estimated item difficulties for all five progressions were very small (-.06 to .00). As seen in Table 14, the item difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all five progressions. Similarly, the DIF magnitudes were negligible for all five progressions.

Language

For one of the eight progressions within the Language domain (10.a), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining seven progressions within the Language domain (8.a, 8.b, 9.a, 9.b, 9.c, 9.d, and 10.b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .05$). The estimated item difficulties for white and non-white children were identical for one progression (10.a). The differences in estimated item difficulties for all progressions were very small (-.11 to .14). As seen in Table 15, the item difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all eight progressions. Similarly, the DIF magnitudes were negligible for all eight progressions.

Cognitive

For six of the 10 progressions within the Cognitive domain (11.b, 11.c, 11.e, 12.a, 12.b, and 13), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining four progressions within the Cognitive domain (11.a, 11.d, 14.a, and 14.b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The

estimated item difficulties for white and non-white children were identical for five progressions (11.b, 11.e, 12.a, 12.b, and 13). The differences in estimated item difficulties for all progressions were very small (-.13 to .10). As seen in Table 16, the item difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across nine of the 10 progressions. The only exception to this finding was for 11.d. The estimated difficulty for white children was .50, which is at the top of the Average range, and the estimated difficulty for non-white children was .63, which is at the bottom of the Difficult range, resulting in a very small difference. The DIF magnitudes were negligible for all 10 progressions.

Literacy

For nine of the 16 progressions within the Literacy domain (15.b, 15.c, 16.b, 17.b, 18.a, 18.b, 18.c, 19.a, and 19.b), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For the remaining seven progressions within the Literacy domain (15.a, 15.d, 16.a, 17.a, 18.d, 18.e, and 19.c), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .001$). The estimated item difficulties for white and non-white children were identical for seven of the 16 progressions (15.c, 16.b, 17.b, 18.b, 18.c, 19.a, and 19.b). The differences in estimated item difficulties for 14 of the 16 progressions were small (-.13 to .28). The only exceptions to his finding were for 18.d where the estimated difference was .39 and 19.c where the estimated difference was .41. These differences are at the upper end of the negligible range. However, both progressions were very difficult for white and non-white children. As seen in Table 17, the item difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all 16 progressions. Similarly, the DIF magnitudes were considered negligible for all 16 progressions.

Mathematics

For five of the 12 progressions within the Mathematics domain (20.a, 20.c, 20.d, 22.c, and 23), neither the Rasch-Welch *t* nor the Mantel-Haenszel χ^2 statistics were statistically significant. For one progression (20.f), only the Rasch-Welch *t* statistic was statistically significant ($p < .05$). For the remaining six progressions within the Mathematics domain (20.b, 20.e, 21.a, 21.b, 22.a, and 22.b), both the Rasch-Welch *t* and Mantel-Haenszel χ^2 statistics were statistically significant ($p < .01$). The estimated item difficulties for white and non-white children were identical for three of the 12 progressions (20.a, 20.b, and 20.c). The differences in estimated item difficulties for all progressions were small (-.09 to .25). As seen in Table 18, the item difficulty classifications (Easy, Average, or Difficult) were the same for both white and non-white children across all 12 progressions. Similarly, the DIF magnitudes were considered negligible for all 12 progressions.

Discussion

The first research question addressed the potential for DIF based on gender. The estimated average item difficulties for female and male children were identical for many *GOLD*® progressions. Where small differences were found, the magnitude of those differences was considered negligible for all of the progressions. The item difficulty categories (Easy, Average, or Difficult) were the same for female and male children across all but one of the *GOLD*® progressions. Therefore, the results of this study yielded no substantive evidence to support a conclusion that DIF based on gender is present as teachers use the *GOLD*® developmental progressions.

The second research question addressed the potential for DIF based on primary language spoken in the home. The results for research question two were very similar to those reported for research question one. The estimated average item difficulties for NE and ELL children were identical for many *GOLD*® progressions. Where small differences were found, the magnitude of those differences was considered negligible for all of the progressions. The item difficulty categories (Easy, Average, or Difficult) were the same for NE and ELL children for all but one of the *GOLD*®

progressions. Therefore, the results of this study yielded no substantive evidence to support a conclusion that DIF based on primary language is present as teachers use the *GOLD*® developmental progressions.

The third research question addressed the potential for DIF based on race / ethnicity of the children. The results for research question three were similar to those reported for research questions one and two. The estimated average item difficulties for white and non-white children were identical for many *GOLD*® progressions. Where small differences were found, the magnitude of those differences was considered negligible for all of the progressions. The difficulty categories (Easy, Average, or Difficult) for white and non-white were the same across all but one of the *GOLD*® progressions. Therefore, the results of this study yielded no substantive evidence to support a conclusion that DIF based on primary language is present as teachers use the *GOLD*® developmental progressions. However, for two of the progressions the differences in estimated difficulty were at the upper end of the negligible range. These two progressions, 18.d (DIF contrast = .39) and 19.c (DIF contrast = .41) will need to be monitored carefully in future research.

It may be useful to review the content of the behavioral anchors associated with each step on those two progressions, 18.d and 19.c, as some differences were found between the estimated item difficulties for white and non-white children. The content of these progressions could be reviewed for fairness to all racial / ethnic subgroups of children. It is also possible that teachers could benefit from more training, specific to 18.d and 19.c, on how to recognize, elicit, and analyze evidences of child developmental progress from non-white children.

It is important to note the potential limitations to the findings of this study. The analyses that addressed research question two used one specific ELL sub-group as the focal group and NE children as the reference group. Therefore, the results of this study are limited to those specific groups. Future studies could attempt to examine potential DIF with other linguistic and cultural

subgroups of children. In addition, this study treated all ELL children as a unitary group without respect to levels of acculturation or language acquisition. Future research could benefit from a more nuanced examination of sub-groups within the ELL population (Kim, Lambert, Durhams, & Burt, 2018).

Similarly, the analyses that addressed research question three were limited to non-white children as the focal group and white children as the reference group. Future analyses could attempt to examine whether the results of this study extend to additional subgroups based on race or ethnicity. Furthermore, this study did not incorporate any information about the gender, native language, or race / ethnicity of the teachers that made the placements on *GOLD*® progressions. Future research could include demographic information about the teachers along with racial and ethnic congruence between teachers and the children and families they serve.

When the results generated by this study are taken as a whole, they offer very little evidence that the *GOLD*® progressions are measuring different latent constructs for the focal and reference groups compared. These results generally confirm the findings of previous studies (Kim, Lambert, & Burts, 2013; Kim, Lambert, & Burts, 2014; Kim, Lambert, Durhams, & Burt, 2018). Therefore, it is reasonable to conclude that teachers are generally using the *GOLD*® progressions to make ratings of children in a similar manner across the sub-groups compared in this study. At least with respect to the subgroups compared, it appears to be reasonable to interpret the resulting assessment scores in a similar manner for all children. However, it is important to note that teachers can introduce other sources of construct irrelevant variance and rater effects when engaging in the complex response process involved with authentic formative assessment for young children. Therefore, future research can focus on furthering the validity argument for the use of *GOLD*® with young children, including examination of evidence for inter-rater reliability.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-574.

Badia, X., Prieto, L., and Linacre, J. M. (2002). Differential item and test functioning (DIF and DTF). *Rasch Measurement Transactions, 16,* 889.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Englehard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), Large-scale assessment programs for all students: Development, implementation, and analysis (pp. 261-287). Mahwah, N.J.: Erlbaum.

Engelhard, G., & Wind, S. (2018). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. New York: Routledge.

Heritage, M. (2013). *Formative assessment in practice: A process of inquiry and action.* Boston: Harvard Education Press.

Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA/APA/NCME, 2014). *The Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Kim, D.H., Lambert, R.G., & Burts, D.C. (2013). Evidence of the validity of teaching strategies GOLD® Assessment Tool for English language learners and children with disabilities. *Early Education and Development, 24*(4), 574–595. https://doi.org/10.1080/10409289.2012.701500

Kim, D. H., Lambert, R. G., & Burts, D. C. (2014). Validating a developmental scale for young

children using the Rasch model: Applicability of the Teaching Strategies GOLD® assessment system. *Journal of Applied Measurement, 15(4)*, 405-421.

Kim, D., Lambert, R., Durham, S., & Burts, D. (2018). Examining the validity of GOLD® with 4-year-old dual language learners. *Early Education and Development*, *29*(4), 477–493. https://doi.org/10.1080/10409289.2018.1460125

Lambert, M. C., Garcia, A. G., January, S. A., & Epstein, M. H. (2017). The impact of English language learner status on screening for emotional and behavioral disorders: A differential item functioning (DIF) study. *Psychology in the Schools, 55*(3), 229-239. https://doi.org/10.1002/pits.22103

Lambert, R. (2020). Shaping a validity argument for the use of authentic formative assessments to support young children. In Martin, C., Polly, D., & Lambert, R. (Eds.) (2020). *Handbook of Research on Formative Assessment in Pre-K through Elementary classrooms.* IGI-Global.

Lambert, R. (2020). *Technical manual* for the Teaching Strategies GOLD® *assessment (second edition): Birth through third grade*. Center for Educational Measurement and Evaluation, University of North Carolina at Charlotte.

Lambert, R. G., Kim, D-H., & Burts, D. C. (2013). Using teacher ratings to track the growth and development of young children using the *Teaching Strategies GOLD®* assessment system. *Journal of Psychoeducational Assessment*. doi:0734282913485214.

Linacre, J.M. (2020). *Winsteps.* [Computer software]. Beaverton, OR: Winsteps.com.

Masters, G.N. (1982). A Rasch model for partial credit scoring, *Psychometrika*, *47,* 149-174.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American*

Psychologist, 50(9), 741-749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Qi, C.H., Marley, S.C. (2009). Differential item functioning analysis of the preschool language scale-4 between English-speaking hispanic and European American children from low-income families. *Topics in Early Childhood Special Education, 29*(3), 171-180. https://doi.org/10.1177/0271121409332674

Vaezghasemi, M., Eurenius, E., Ivarsson, A., Sundberg, L.R., Silfverdal, S.A., & Lindkvist, M.. (2020). Socio-emotional problems among Swedish three-year-olds: an idem response theory analysis of the ages and stages questionnaire: social-emotional. *BMP Pediatrics*. https://doi.org/10.1186/s12887-020-2000-y

Walker, C. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment, 29,* 364-376.

Webb, M.L., Cohen, A.S., Schwanenflugel, P.J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test-III. *Education and Psychological Measurement, 68*(2), 335-351. https://doi.org/10.1177/0013164407308474

Zwick, R., Thayer, D. T., and Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement, 36,* 1-28.

Table 1
Differential item functioning results by gender for the Social Emotional domain

Developmental Progression	Female Rasch Item Difficulty		Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
1.a	-0.12	Average	-0.12	Average	0.00	Negligible	0.02	0.00		1.09	
1.b	-0.63	Easy	-0.56	Easy	-0.07	Negligible	0.02	-4.44	***	22.90	***
1.c	-0.85	Easy	-0.85	Easy	0.00	Negligible	0.01	0.00		1.54	
2.a	-2.75	Easy	-2.85	Easy	0.11	Negligible	0.02	6.68	***	33.36	***
2.b	0.49	Average	0.49	Average	0.00	Negligible	0.02	0.00		5.25	*
2.c	0.59	Difficult	0.59	Difficult	0.00	Negligible	0.01	0.00		2.13	
2.d	0.81	Difficult	0.76	Difficult	0.05	Negligible	0.01	3.32	***	11.82	***
3.a	1.29	Difficult	1.29	Difficult	0.00	Negligible	0.01	0.00		2.71	
3.b	1.21	Difficult	1.21	Difficult	0.00	Negligible	0.02	0.00		0.11	

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 2
Differential item functioning results by gender for the Physical domain

Developmental Progression	Female Rasch Item Difficulty		Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
4	-0.97	Easy	-1.08	Easy	0.10	Negligible	0.02	5.99	***	40.62	***
5	0.44	Average	0.38	Average	0.06	Negligible	0.02	3.65	***	15.67	***
6	0.51	Difficult	0.26	Average	0.25	Negligible	0.02	14.32	***	100.00	***
7.a	-1.11	Easy	-0.96	Easy	-0.15	Negligible	0.02	-8.55	***	82.44	***
7.b	1.15	Difficult	1.38	Difficult	-0.23	Negligible	0.02	-14.60	***	100.00	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 3
Differential item functioning results by gender for the Language domain

Developmental Progression	Female Rasch Item Difficulty		Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
8.a	0.04	Average	-0.01	Average	0.05	Negligible	0.02	2.43	*	7.56	**
8.b	-2.80	Easy	-2.78	Easy	-0.02	Negligible	0.02	-1.16		4.27	*
9.a	0.25	Average	0.25	Average	0.00	Negligible	0.02	0.00		1.08	
9.b	-0.10	Average	-0.04	Average	-0.05	Negligible	0.02	-2.98	**	11.07	***
9.c	0.52	Difficult	0.52	Difficult	0.00	Negligible	0.02	0.00		1.66	
9.d	1.09	Difficult	1.04	Difficult	0.05	Negligible	0.02	2.89	**	8.50	**
10.a	0.32	Average	0.32	Average	0.00	Negligible	0.02	0.00		3.50	
10.b	0.71	Difficult	0.71	Difficult	0.00	Negligible	0.02	0.00		0.08	

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 4
Differential item functioning results by gender for the Cognitive domain

Developmental Progression	Female Rasch Item Difficulty		Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
11.a	-0.39	Average	-0.26	Average	-0.13	Negligible	0.02	-6.95	***	40.69	***
11.b	-1.63	Easy	-1.57	Easy	-0.06	Negligible	0.02	-3.06	**	10.19	**
11.c	-0.96	Easy	-0.96	Easy	0.00	Negligible	0.02	0.00		0.30	
11.d	0.59	Difficult	0.52	Difficult	0.07	Negligible	0.02	3.63	***	14.55	***
11.e	0.96	Difficult	0.94	Difficult	0.02	Negligible	0.02	1.32	*	5.74	*
12.a	0.31	Average	0.31	Average	0.00	Negligible	0.02	0.00		2.97	
12.b	0.97	Difficult	0.97	Difficult	0.00	Negligible	0.02	0.00		3.05	
13	0.00	Average	-0.11	Average	0.11	Negligible	0.02	5.90	***	34.11	***
14.a	-0.06	Average	-0.06	Average	0.00	Negligible	0.02	0.00		2.01	
14.b	0.18	Average	0.28	Average	-0.10	Negligible	0.02	-5.77	***	39.41	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 5
Differential item functioning results by gender for the Literacy domain

Developmental Progression	Female Rasch Item Difficulty		Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
15.a	-1.71	Easy	-1.66	Easy	-0.05	Negligible	0.01	-3.77	***	13.73	***
15.b	-2.20	Easy	-2.20	Easy	0.00	Negligible	0.01	0.00		5.22	*
15.c	0.94	Difficult	0.94	Difficult	0.00	Negligible	0.01	0.00		3.99	*
15.d	1.82	Difficult	1.80	Difficult	0.02	Negligible	0.02	1.22		3.84	
16.a	-2.04	Easy	-2.16	Easy	0.12	Negligible	0.01	9.74	***	76.60	***
16.b	-0.97	Easy	-1.03	Easy	0.06	Negligible	0.01	4.09	***	17.82	***
17.a	-0.32	Average	-0.32	Average	0.00	Negligible	0.02	0.00		1.18	
17.b	-1.19	Easy	-1.19	Easy	0.00	Negligible	0.01	0.00		0.06	
18.a	0.78	Difficult	0.78	Difficult	0.00	Negligible	0.01	0.00		0.53	
18.b	-2.18	Easy	-2.18	Easy	0.00	Negligible	0.01	0.00		10.01	**
18.c	1.07	Difficult	1.07	Difficult	0.00	Negligible	0.01	0.00		5.40	*
18.d	3.04	Difficult	2.95	Difficult	0.09	Negligible	0.03	3.50	***	6.58	*
18.e	2.93	Difficult	2.88	Difficult	0.05	Negligible	0.03	2.04	*	3.31	
19.a	-2.88	Easy	-2.88	Easy	0.00	Negligible	0.01	0.00		2.85	
19.b	0.08	Average	0.13	Average	-0.05	Negligible	0.01	-4.39	***	13.68	***
19.c	2.97	Difficult	2.97	Difficult	0.00	Negligible	0.03	0.00		0.02	

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 6
Differential item functioning results by gender for the Mathematics domain

Developmental Progression	Female Rasch Item Difficulty		Male Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
20.a	-0.87	Easy	-0.87	Easy	0.00	Negligible	0.01	0.00		3.44	
20.b	-0.92	Easy	-0.92	Easy	0.00	Negligible	0.01	0.00		3.05	
20.c	-0.94	Easy	-1.01	Easy	0.07	Negligible	0.01	5.47	***	37.62	***
20.d	2.32	Difficult	2.23	Difficult	0.09	Negligible	0.03	3.21	**	6.79	**
20.e	2.41	Difficult	2.34	Difficult	0.07	Negligible	0.03	2.67	**	6.31	*
20.f	2.06	Difficult	1.96	Difficult	0.09	Negligible	0.02	3.75	***	9.69	**
21.a	-4.15	Easy	-4.08	Easy	-0.07	Negligible	0.01	-4.48	***	17.50	***
21.b	-0.75	Easy	-0.75	Easy	0.00	Negligible	0.01	0.00		0.23	
22.a	-0.14	Average	-0.14	Average	0.00	Negligible	0.01	0.00		1.82	
22.b	0.79	Difficult	0.79	Difficult	0.00	Negligible	0.02	0.00		1.74	
22.c	1.04	Difficult	1.04	Difficult	0.00	Negligible	0.02	0.00		0.04	
23	-0.71	Easy	-0.71	Easy	0.00	Negligible	0.01	0.00		1.56	

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 7
Differential item functioning results by primary language for the Social Emotional domain

Developmental Progression	English Rasch Item Difficulty		Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
1.a	-0.12	Average	-0.14	Average	0.02	Negligible	0.02	0.96		0.50	
1.b	-0.60	Easy	-0.60	Easy	0.00	Negligible	0.02	0.00		0.54	
1.c	-0.85	Easy	-0.90	Easy	0.05	Negligible	0.02	2.19	*	11.49	***
2.a	-2.83	Easy	-2.67	Easy	-0.15	Negligible	0.02	-6.91	***	49.23	***
2.b	0.49	Average	0.51	Difficult	-0.02	Negligible	0.02	-0.96		0.66	
2.c	0.59	Difficult	0.72	Difficult	-0.13	Negligible	0.02	-6.04	***	50.63	***
2.d	0.78	Difficult	0.87	Difficult	-0.08	Negligible	0.02	-3.88	***	21.52	***
3.a	1.31	Difficult	1.19	Difficult	0.13	Negligible	0.02	6.03	***	47.33	***
3.b	1.24	Difficult	1.06	Difficult	0.18	Negligible	0.02	8.21	***	74.64	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 8
Differential item functioning results by primary language for the Physical domain

Developmental Progression	English Rasch Item Difficulty		Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
4	-1.02	Easy	-0.98	Easy	-0.04	Negligible	0.03	-1.61		3.32	
5	0.43	Average	0.26	Average	0.18	Negligible	0.02	7.15	***	61.19	***
6	0.38	Average	0.44	Average	-0.05	Negligible	0.03	-2.07	*	5.96	*
7.a	-1.03	Easy	-1.08	Easy	0.04	Negligible	0.03	1.60		5.27	*
7.b	1.27	Difficult	1.36	Difficult	-0.10	Negligible	0.02	-4.10	***	23.60	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 9
Differential item functioning results by primary language for the Language domain

Developmental Progression	English Rasch Item Difficulty		Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
8.a	0.01	Average	0.04	Average	-0.02	Negligible	0.03	-0.81		0.81	
8.b	-2.77	Easy	-3.00	Easy	0.23	Negligible	0.03	8.28	***	54.90	***
9.a	0.25	Average	0.29	Average	-0.04	Negligible	0.03	-1.34		0.42	
9.b	-0.09	Average	0.11	Average	-0.20	Negligible	0.03	-7.11	***	59.30	***
9.c	0.52	Difficult	0.68	Difficult	-0.17	Negligible	0.03	-6.35	***	64.51	***
9.d	1.06	Difficult	1.02	Difficult	0.04	Negligible	0.03	1.67		5.71	*
10.a	0.32	Average	0.35	Average	-0.02	Negligible	0.03	-0.95		0.12	
10.b	0.73	Difficult	0.55	Difficult	0.18	Negligible	0.03	6.73	***	41.92	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 10
Differential item functioning results by primary language for the Cognitive domain

Developmental Progression	English Rasch Item Difficulty		Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
11.a	-0.33	Average	-0.27	Average	-0.06	Negligible	0.03	-2.11	*	8.77	**
11.b	-1.58	Easy	-1.75	Easy	0.18	Negligible	0.03	6.63	***	48.33	***
11.c	-0.92	Easy	-1.19	Easy	0.27	Negligible	0.03	10.54	***	100.00	***
11.d	0.55	Difficult	0.65	Difficult	-0.10	Negligible	0.03	-3.74	***	17.38	***
11.e	0.94	Difficult	0.99	Difficult	-0.05	Negligible	0.02	-1.96		4.72	*
12.a	0.31	Average	0.37	Average	-0.06	Negligible	0.03	-2.37	*	7.85	**
12.b	0.97	Difficult	1.08	Difficult	-0.11	Negligible	0.03	-4.36	***	28.07	***
13	-0.06	Average	-0.02	Average	-0.04	Negligible	0.03	-1.44		2.91	
14.a	-0.06	Average	-0.10	Average	0.05	Negligible	0.03	1.74		5.71	*
14.b	0.23	Average	0.28	Average	-0.05	Negligible	0.03	-1.80		3.83	

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 11
Differential item functioning results by primary language for the Literacy domain

Developmental Progression	English Rasch Item Difficulty		Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
15.a	-1.68	Easy	-1.55	Easy	-0.14	Negligible	0.02	-5.71	***	36.02	***
15.b	-2.20	Easy	-2.15	Easy	-0.05	Negligible	0.02	-2.04	*	3.13	
15.c	0.94	Difficult	0.94	Difficult	0.00	Negligible	0.02	0.00		0.03	
15.d	1.80	Difficult	1.71	Difficult	0.09	Negligible	0.03	2.51	*	5.06	*
16.a	-2.10	Easy	-1.97	Easy	-0.13	Negligible	0.02	-6.21	***	41.19	***
16.b	-1.01	Easy	-0.96	Easy	-0.05	Negligible	0.02	-1.98	*	4.61	*
17.a	-0.32	Average	-0.22	Average	-0.10	Negligible	0.03	-3.92	***	16.59	***
17.b	-1.19	Easy	-1.19	Easy	0.00	Negligible	0.02	0.00		2.05	
18.a	0.78	Difficult	0.90	Difficult	-0.12	Negligible	0.02	-5.20	***	24.70	***
18.b	-2.18	Easy	-2.18	Easy	0.00	Negligible	0.02	0.00		0.09	
18.c	1.07	Difficult	1.11	Difficult	-0.05	Negligible	0.02	-1.98	*	5.00	*
18.d	2.99	Difficult	2.83	Difficult	0.16	Negligible	0.05	3.17	**	1.06	
18.e	2.90	Difficult	2.74	Difficult	0.16	Negligible	0.05	3.13	**	9.82	**
19.a	-2.88	Easy	-2.97	Easy	0.09	Negligible	0.02	5.92	***	22.62	***
19.b	0.12	Average	-0.06	Average	0.18	Negligible	0.02	10.14	***	75.36	***
19.c	2.99	Difficult	2.76	Difficult	0.22	Negligible	0.05	4.56	***	10.28	**

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 12
Differential item functioning results by primary language for the Mathematics domain

Developmental Progression	English Rasch Item Difficulty		Spanish Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
20.a	-0.87	Easy	-0.87	Easy	0.00	Negligible	0.02	0.00		0.64	
20.b	-0.92	Easy	-0.98	Easy	0.06	Negligible	0.02	2.87	**	30.01	***
20.c	-0.98	Easy	-0.98	Easy	0.00	Negligible	0.02	0.00		0.36	
20.d	2.27	Difficult	2.20	Difficult	0.07	Negligible	0.05	1.33		0.12	
20.e	2.37	Difficult	2.24	Difficult	0.13	Negligible	0.05	2.83	**	7.96	**
20.f	2.01	Difficult	1.92	Difficult	0.09	Negligible	0.04	1.92		0.58	
21.a	-4.11	Easy	-4.04	Easy	-0.07	Negligible	0.02	-3.29	***	17.97	***
21.b	-0.75	Easy	-0.67	Easy	-0.08	Negligible	0.02	-3.86	***	22.14	***
22.a	-0.14	Average	-0.25	Average	0.11	Negligible	0.02	5.09	***	34.37	***
22.b	0.79	Difficult	0.89	Difficult	-0.10	Negligible	0.02	-4.10	***	29.65	***
22.c	1.04	Difficult	1.01	Difficult	0.03	Negligible	0.03	1.11		0.43	
23	-0.71	Easy	-0.71	Easy	0.00	Negligible	0.02	0.00		1.57	

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 13
Differential item functioning results by race / ethnicity for the Social Emotional domain

Developmental Progression	White Rasch Item Difficulty		Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
1.a	-0.12	Average	-0.10	Average	-0.03	Negligible	0.02	-1.60		9.25	**
1.b	-0.63	Easy	-0.56	Easy	-0.07	Negligible	0.02	-4.77	***	24.52	***
1.c	-0.88	Easy	-0.81	Easy	-0.08	Negligible	0.01	-5.10	***	25.96	***
2.a	-2.90	Easy	-2.70	Easy	-0.21	Negligible	0.02	-12.70	***	100.00	***
2.b	0.49	Average	0.49	Average	0.00	Negligible	0.02	0.00		0.02	
2.c	0.59	Difficult	0.59	Difficult	0.00	Negligible	0.01	0.00		8.31	**
2.d	0.82	Difficult	0.74	Difficult	0.08	Negligible	0.01	5.40	***	32.29	***
3.a	1.36	Difficult	1.22	Difficult	0.14	Negligible	0.01	9.73	***	100.00	***
3.b	1.26	Difficult	1.16	Difficult	0.10	Negligible	0.02	6.32	***	41.39	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 14
Differential item functioning results by race / ethnicity for the Physical domain

Developmental Progression	White Rasch Item Difficulty		Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
4	-1.02	Easy	-1.02	Easy	0.00	Negligible	0.02	0.00		1.59	
5	0.41	Average	0.41	Average	0.00	Negligible	0.02	0.00		5.84	*
6	0.38	Average	0.38	Average	0.00	Negligible	0.02	0.00		5.98	*
7.a	-1.03	Easy	-1.03	Easy	0.00	Negligible	0.02	0.00		0.48	
7.b	1.24	Difficult	1.30	Difficult	-0.06	Negligible	0.02	-3.44	***	18.62	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 15
Differential item functioning results by race / ethnicity for the Language domain

Developmental Progression	White Rasch Item Difficulty		Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
8.a	-0.02	Average	0.05	Average	-0.07	Negligible	0.02	-3.59	***	14.85	***
8.b	-2.77	Easy	-2.84	Easy	0.07	Negligible	0.02	3.71	**	8.37	**
9.a	0.21	Average	0.32	Average	-0.11	Negligible	0.02	-5.87	***	33.92	***
9.b	-0.11	Average	-0.02	Average	-0.09	Negligible	0.02	-4.68	***	25.58	***
9.c	0.52	Difficult	0.55	Difficult	-0.03	Negligible	0.02	-1.70	*	5.06	*
9.d	1.09	Difficult	1.03	Difficult	0.06	Negligible	0.02	3.51	***	13.43	***
10.a	0.32	Average	0.32	Average	0.00	Negligible	0.02	0.00		0.45	
10.b	0.77	Difficult	0.63	Difficult	0.14	Negligible	0.02	8.08	***	58.85	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 16
Differential item functioning results by race / ethnicity for the Cognitive domain

Developmental Progression	White Rasch Item Difficulty		Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
11.a	-0.37	Average	-0.27	Average	-0.10	Negligible	0.02	-5.17	***	33.33	***
11.b	-1.60	Easy	-1.60	Easy	0.00	Negligible	0.02	0.00		2.10	
11.c	-0.96	Easy	-0.93	Easy	-0.02	Negligible	0.02	-1.29		2.72	
11.d	0.50	Average	0.63	Difficult	-0.13	Negligible	0.02	-6.87	***	49.39	***
11.e	0.94	Difficult	0.94	Difficult	0.00	Negligible	0.02	0.00		2.04	
12.a	0.31	Average	0.31	Average	0.00	Negligible	0.02	0.00		0.35	
12.b	0.97	Difficult	0.97	Difficult	0.00	Negligible	0.02	0.00		0.03	
13	-0.06	Average	-0.06	Average	0.00	Negligible	0.02	0.00		1.03	
14.a	-0.01	Average	-0.11	Average	0.10	Negligible	0.02	5.66	***	38.19	***
14.b	0.28	Average	0.18	Average	0.10	Negligible	0.02	5.51	***	33.41	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 17
Differential item functioning results by race / ethnicity for the Literacy domain

Developmental Progression	White Rasch Item Difficulty		Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
15.a	-1.73	Easy	-1.62	Easy	-0.10	Negligible	0.01	-7.22	***	49.40	***
15.b	-2.20	Easy	-2.17	Easy	-0.02	Negligible	0.01	-1.66		3.82	
15.c	0.94	Difficult	0.94	Difficult	0.00	Negligible	0.01	0.00		0.47	
15.d	1.87	Difficult	1.69	Difficult	0.18	Negligible	0.02	9.42	***	86.50	***
16.a	-2.13	Easy	-2.05	Easy	-0.08	Negligible	0.01	-6.23	***	32.95	***
16.b	-1.01	Easy	-1.01	Easy	0.00	Negligible	0.01	0.00		4.63	
17.a	-0.37	Average	-0.24	Average	-0.13	Negligible	0.02	-8.19	***	70.71	***
17.b	-1.19	Easy	-1.19	Easy	0.00	Negligible	0.01	0.00		1.47	
18.a	0.78	Difficult	0.81	Difficult	-0.03	Negligible	0.01	-1.94		6.55	
18.b	-2.18	Easy	-2.18	Easy	0.00	Negligible	0.01	0.00		1.98	
18.c	1.07	Difficult	1.07	Difficult	0.00	Negligible	0.01	0.00		0.05	
18.d	3.14	Difficult	2.75	Difficult	0.39	Negligible	0.03	14.62	***	100.00	***
18.e	3.01	Difficult	2.73	Difficult	0.28	Negligible	0.03	10.65	***	100.00	***
19.a	-2.88	Easy	-2.88	Easy	0.00	Negligible	0.01	0.00		8.62	
19.b	0.10	Average	0.10	Average	0.00	Negligible	0.01	0.00		4.77	
19.c	3.12	Difficult	2.72	Difficult	0.41	Negligible	0.03	15.65	***	100.00	***

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.

Table 18
Differential item functioning results by race / ethnicity for the Mathematics domain

Developmental Progression	White Rasch Item Difficulty		Non-white Rasch Item Difficulty		DIF Contrast	DIF Magnitude	Joint *se*	Rasch-Welch *t*	*p*	Mantel Haenszel χ^2	*p*
20.a	-0.87	Easy	-0.87	Easy	0.00	Negligible	0.01	0.00		5.53	
20.b	-0.92	Easy	-0.92	Easy	0.00	Negligible	0.01	0.00	**	1.86	***
20.c	-0.98	Easy	-0.98	Easy	0.00	Negligible	0.01	0.00		0.01	
20.d	2.35	Difficult	2.14	Difficult	0.21	Negligible	0.03	7.56		41.37	
20.e	2.47	Difficult	2.22	Difficult	0.25	Negligible	0.03	9.43	**	82.58	**
20.f	2.09	Difficult	1.87	Difficult	0.22	Negligible	0.03	8.62	*	51.89	
21.a	-4.14	Easy	-4.07	Easy	-0.07	Negligible	0.01	-4.81	***	30.87	***
21.b	-0.79	Easy	-0.70	Easy	-0.09	Negligible	0.01	-6.09	***	48.23	***
22.a	-0.16	Average	-0.11	Average	-0.05	Negligible	0.01	-3.83	***	14.39	***
22.b	0.82	Difficult	0.76	Difficult	0.05	Negligible	0.02	3.38	***	3.36	***
22.c	1.08	Difficult	0.99	Difficult	0.09	Negligible	0.02	5.02		20.88	
23	-0.74	Easy	-0.67	Easy	-0.07	Negligible	0.01	-5.20		26.31	

Note. *** = $p < .001$, ** = $p < .01$, * = $p < .05$.