2010-08-06

Fighting Bias with Statistics: Detecting Gender Differences in Responses on Items on a Preschool Science Assessment

Ariela Caren Greenberg
University of Miami, ac_greenberg@yahoo.com

UNIVERSITY OF MIAMI

FIGHTING BIAS WITH STATISTICS: DETECTING GENDER DIFFERENCES IN RESPONSES TO ITEMS ON A PRESCHOOL SCIENCE ASSESSMENT

By

Ariela Caren Greenberg

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

August 2010

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

FIGHTING BIAS WITH STATISTICS: DETECTING GENDER DIFFERENCES IN RESPONSES TO ITEMS ON A PRESCHOOL SCIENCE ASSESSMENT

Ariela Caren Greenberg

Approved:

_____ Daryl Greenfield, Ph.D. Professor of Psychology	_____ Terri A. Scandura, Ph.D. Dean of the Graduate School
_____ Randall Penfield, Ph.D. Associate Professor of Education	_____ Erin Kobetz, Ph.D. Assistant Professor of Epidemiology and Public Health
_____ Maria Llabre, Ph.D. Professor of Psychology	_____ Rebecca Bulotsky-Shearer, Ph.D. Assistant Professor of Psychology

GREENBERG, ARIELA CAREN (Ph.D., Psychology)

Fighting Bias with Statistics: Detecting Gender Differences in Responses to Items on a Preschool Science Assessment (August 2010)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Daryl Greenfield.
No. of pages in text. (94)

Differential item functioning (DIF) and *differential distractor functioning* (DDF) are methods used to screen for item bias (Camilli & Shepard, 1994; Penfield, 2008). Using an applied empirical example, this mixed-methods study examined the congruency and relationship of DIF and DDF methods in screening multiple-choice items. Data for Study I were drawn from item responses of 271 female and 236 male low-income children on a preschool science assessment. Item analyses employed a common statistical approach of the Mantel-Haenszel log-odds ratio (MH-LOR) to detect DIF in dichotomously scored items (Holland & Thayer, 1988), and extended the approach to identify DDF (Penfield, 2008). Findings demonstrated that the using MH-LOR to detect DIF and DDF supported the theoretical relationship that the magnitude and form of DIF and are dependent on the DDF effects, and demonstrated the advantages of studying DIF and DDF in multiple-choice items. A total of 4 items with DIF and DDF and 5 items with only DDF were detected. Study II incorporated an item content review, an important but often overlooked and under-published step of DIF and DDF studies (Camilli & Shepard). Interviews with 25 female and 22 male low-income preschool children and an expert review helped to interpret the DIF and DDF results and their

comparison, and determined that a content review process of studied items can reveal reasons for potential item bias that are often congruent with the statistical results. Patterns emerged and are discussed in detail. The quantitative and qualitative analyses were conducted in an applied framework of examining the validity of the preschool science assessment scores for evaluating science programs serving low-income children, however, the techniques can be generalized for use with measures across various disciplines of research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

Test validity, the extent to which evidence and theory support the purpose and interpretation of test scores, can be substantiated by analysis of content, construct validity, external criteria, and analysis of internal structure (AERA, APA, NCME, 1999; Osterlind, 2006). Internal structure analyses generally examine the relation between the test items and test score interpretations, such as the degree to which items measure the intended constructs, or function the same way for all examinees (AERA, APA, NCME, 1999). When items do not function the same way for all test takers, subgroup differences related to irrelevant constructs may signal potential bias or a threat to test fairness (Camilli & Shepard, 1994).

According to Camilli and Shepard (1994), bias is "systematic error that distorts the meaning of these inferences for members of a particular group" (p. 1). To evaluate items for bias, both quantitative procedures and follow-up qualitative *content review* investigations are necessary to determine if any differential difficulty is related to relevant or irrelevant constructs. Quantitative approaches to examining these conditional between-group differences include *differential item functioning* and its extension - *differential distractor functioning*. *Differential item functioning* (DIF) occurs when, after controlling for ability level, certain subgroups have a different probability of answering an item correctly (Hambleton, Swaminathan, & Rogers, 1991; Holland & Thayer, 1988). *Differential distractor functioning* (DDF) occurs when, after controlling for ability, certain subgroups have a different chance of selecting particular distractor choices in an item (Penfield, 2008). To explain statistical results, c*ontent reviews* include further

examination of items by gathering information from experts or test takers to provide judgments regarding item bias (Camilli & Shepard, 1994).

Issues of validity associated with potential item bias have been studied using statistical methods of DIF (e.g., Camilli & Shepard, 1994; Dorans & Holland, 1992; Gelin, Carleton, Smith, & Zumbo, 2004; Hidalgo & Lopez-Pina, 2004; Holland & Thayer, 1988; Li, Cohen, & Ibarra, 2004; Poggio, Glasnapp, Yang, & Poggio, 2005; Rogers & Swaminathan, 1993). Given that DIF is more widely used, less is known about the relation of DIF and DDF, the utility of DDF to detect item bias, and whether conducting DDF analysis in combination with DIF yields similar, unique, or complementary results in applied datasets (Penfield, 2010). More research is needed to clarify how to ensure a comprehensive yet efficient statistical screening process to accurately detect item bias in multiple-choice items. Although research on non-parametric methods to detect DIF and DDF has been conducted with assessments in higher grade levels (e.g, Hidalgo & Lopez-Pina, 2004), more research on these methods is needed in early education assessment contexts. Thus, the current two-part study examines the use of DIF, DDF, and content reviews in an applied setting using a non-parametric approach.

Differential Item Functioning (DIF)

In DIF studies, the two groups being compared are designated as reference and focal; the *reference* group typically represents the majority (or advantaged) group and the *focal g*roup typically represents the protected (or disadvantaged) group. Thus, consider for example a science test item that relies heavily on verbal skills. English language learners (focal group) may be less likely to answer correctly than native English speakers

(reference group) with similar science abilities. This item may demonstrate DIF, because the group difference is related to language rather than science ability, and the group difference in probability of correct response exists at matched levels of science ability.

There are two general forms of DIF for dichotomously scored items: uniform and nonuniform (Camilli & Shepard, 1994). Uniform DIF exists when the relative advantage of one group over the other is constant across ability levels (see Figure 1). Nonuniform DIF exists when this relative advantage is not consistent across ability levels, for example, when there is greater disparity at lower ability levels than at higher ability levels (see Figure 2). A special case of nonuniform DIF shown in Figure 3 is called crossing DIF, when the reference group is favored at one end of ability, yet the focal group is favored at the other end of ability (Camilli & Shepard, 1994). The extent to which these two forms of DIF relate to the selection of distractors will be addressed.

Statistical methods for DIF. Common statistical methods for examining DIF include logistic regression (LR; Jodoin & Gierl, 2001; Swaminathan & Rogers, 1990), item response theory (IRT; Camilli & Shepard, 1994; Clauser & Mazor, 1998; Thissen, Steinberg, & Wainer, 1993), and the Mantel-Haenszel log-odds ratio (MH-LOR; Mantel & Haenszel, 1959). An extensive comparison of these and other methods can be found in Penfield and Camilli (2007). Although LR can detect both uniform and nonuniform DIF, and IRT estimates trait level parameters for both persons' responses and item properties, LR and IRT methods require large samples and are more costly and time consuming procedures compared to other methods due to the iterative process of model fitting. In contrast, the advantages to the MH-LOR, considered one of the most powerful DIF

methods (Haladyna, 2004), are that it requires modest sample sizes, is more efficient, and provides meaningful results (Penfield, 2008).

MH-LOR to assess DIF. For dichotomously scored items, the MH-LOR method compares the *odds of a correct response for the reference group* to the *odds of a correct response for the focal group* at each level of ability (i.e., the total observed score; Holland & Thayer, 1988). Then, a weighted average across ability levels is calculated to obtain the MH-LOR for each item, denoted here as β_{MH} (see Appendix A). Take for example how the difference between two item characteristic curves represents the DIF effect (see Figures 1-3). At each level of ability, the log odds ratio is calculated and then averaged (weighted proportionately to the sample size). The figures illustrate how that average log-odds ratio is dependent on the difference in distribution of item responses among the two groups being compared. When the log-odds ratio is relatively equal across ability levels, there is uniform DIF (see Figure 1). When the log-odds ratio varies across ability levels, there is nonuniform DIF (see Figure 2). When the log-odds ratio is positive for one group at one extreme and negative at the other extreme there is crossing DIF; the positive and negative odds ratios cancel out (see Figure 3).

The MH-LOR approach has demonstrated power equivalent to other methods such as logistic regression for evaluating DIF in dichotomously scored items (e.g., Greenberg, Penfield, & Greenfield, 2007; Hidalgo & Lopez-Pina, 2004). Additionally, β_{MH} indicates the direction and magnitude of the differential effect. The direction of the DIF effect is indicated by the sign of β_{MH}, where a positive effect indicates DIF favoring the reference group (e.g., males) and a negative effect indicates DIF favoring the focal group (e.g., females). For example, for an item where β_{MH} = -0.80, females are more

likely to answer correctly than males. The magnitude of the effect size is determined by the absolute value of β_{MH}, where zero equals no DIF, and effects can theoretically range from $-\infty$ to ∞. This effect is often evaluated based on established criteria commonly used by Educational Testing Service (ETS) where a small effect is $|\beta_{MH}| < 0.43$; a moderate effect is, $0.43 \leq |\beta_{MH}| \leq 0.64$, and a large effect is $|\beta_{MH}| > 0.64$ (Penfield & Camilli, 2007). Although the effect size itself can be meaningful without tests of significance, the β_{MH} is often evaluated for statistical significance by using a *z*-test of the null hypothesis of no differential item functioning.

Limitations of DIF. DIF can be sufficient in some situations, such as examining dichotomously designed items (e.g., true-false, open-ended correct or incorrect; skill attained or unattained). DIF can also be useful in examining multiple-choice items when they are dichotomously scored. Detecting only DIF in multiple-choice items, however, can limit the capacity to detect the location of the effect in those items (Penfield, 2008). For instance, if one subgroup selects equally or randomly among all incorrect options, the effect is likely due to the questions or an overall problem with the item. In another instance, one subgroup might disproportionately select a particular distractor, and that distractor option might be a cause of the DIF effect. Thus, even with a DIF effect, without separately examining how distractors are functioning for each item, it is not possible to distinguish between these two patterns and possible reasons for DIF.

In yet another situation, one group might disproportionately choose one incorrect answer and the other group disproportionately chooses another incorrect answer. DIF alone will not detect this pattern: a difference in answer patterns that does not directly

relate to a difference the probability of correct response. Furthermore, when this pattern occurs, each groups' relative advantage is affected and scores are misestimated. Due to the possibility each group is choosing different distractors for different underlying reasons, the interpretation of scores may be different for each group. Therefore, a method to detect each of these instances of distractor effects is essential to ensure validity of score interpretations.

Differential Distractor Functioning (DDF)

To further examine multiple-choice items, DIF concepts and their associated statistical methods have been extended to examine *differential distractor functioning* (DDF; Schmitt & Bleistein, 1987; Schmitt & Dorans, 1990; Veale & Foreman, 1983). DDF examines group differences in the probability of selecting among more than one incorrect distractor options. Thus DDF theoretically provides information above and beyond findings from DIF analyses about differential responses to multiple-choice items, because the probability of choosing a correct answer can be studied as the dependency of selecting from three or more options (Penfield, 2008). DDF methods can identify whether distractor options may be contributing to group differences in selection of multiple-choice responses and whether those responses lead to differences in the probability of a correct response.

Statistical Methods to assess DDF. Methods for identifying DDF include a log-linear method, standardization, IRT, and MH-LOR. According to the log-linear method, a DDF effect exists when, controlling for the main effects of ability and subgroup, there is an interaction effect where members of different subgroups with the same ability are choosing different option choices at different rates (Green, Crone, & Folk, 1989).

Although the log-linear method detects the presence of interaction effects, it indicates neither the magnitude nor location of specific DDF effects. One nonparametric approach for DDF is standardization (Dorans, Schmitt, & Bleistein, 1992). Standardization results indicate a more meaningful measure of the DDF effect size for each distractor; however, this DDF effect is not theoretically equivalent to parametric models, and therefore does not provide a clear interpretation of effect size (Penfield, 2008). DDF can also be evaluated using a multiple-choice model or nominal response IRT framework (Thissen et al., 1993), but this method has several disadvantages: the models become increasingly more complex (Penfield, 2008; Thissen et al., 1993); necessitate a working knowledge of the models, and parameter estimation, by analysts and other individuals reviewing results (Haladyna, 2004); and require special software to analyze the data (Penfield, 2008). Another disadvantage of these three methods is that they require large sample sizes for stable estimates of effect sizes, in particular IRT, which requires very large sample sizes - ideally 1000 per sub group to account to estimate differences in difficulty, discrimination, and guessing (Embretson & Reise, 2000). Another nonparametric approach is the MH-LOR. According to Penfield (2008), MH-LOR is an advantageous method for estimating DDF because it (a) provides a meaningful effect size (i.e., it provides the magnitude, direction, and location of effect), (b) does not require large sample sizes for iterative procedures to estimate model parameters, and (c) is relatively straightforward in conveying and sharing results with other researchers.

MH-LOR for DDF. To extend the MH-LOR method of studying DIF, Penfield (2008) proposed using the MH-LOR for DDF to compare the *likelihood of correct response to the likelihood of choosing each distractor, after controlling for ability.*

Further, Penfield proposed that a theoretical relationship exists where the MH-LOR DIF effect is equivalent to a weighted aggregate of the MH-LOR DDF effects (see Penfield, 2010). More specifically, the patterns of DDF help explain and define the presence and type of DIF in an item (e.g., uniform, nonuniform, crossing), because this aggregate is dependent on the item properties (e.g., difficulty, discrimination).

To calculate DDF, the MH-LOR method for DIF is extended to compare the odds of selecting a correct answer given the choice between the correct option and each distractor. A separate MH-LOR value is calculated for each distractor *j* and denoted here as β_{MH_j}. For example, in a multiple-choice item with distractors denoted by *j* = 1, 2,…*n*, the β_{MH_j} would be computed *n* times, or once for each distractor *j*. Thus, for the j^{th} distractor, β_{MH_j} describes between-group differences in the probability of choosing the correct response over choosing that particular j^{th} distractor, after controlling for ability level (see Appendix B).

One of the advantages of MH-LOR is that the statistic is clear to understand: it indicates the magnitude, direction, and the location of DDF effects. First, the magnitude of effect is determined by the absolute value of β_{MH_j} and serves as an estimate of the DDF effect for the j^{th} distractor, where zero equals no DDF for that distractor. This magnitude is evaluated based on the same ETS effect size criteria established for examining DIF in dichotomous items where: small is $\left|\beta_{MH_j}\right| < 0.43$, moderate is $0.43 \leq \left|\beta_{MH_j}\right| \leq 0.64$, and large is $\left|\beta_{MH_j}\right| > 0.64$ (Penfield & Camilli, 2007). Again,

β_{MH_j} can be evaluated by effect size only or along with statistical significance by using a *z*-test of the null hypothesis of no differential distractor functioning.

Second, the direction of the DDF effect indicates which subgroup (e.g., gender) is favored and which is disadvantaged for the j^{th} distractor. This direction is indicated by the sign of β_{MH_j}, where a positive effect indicates that a distractor is disproportionately less attractive to the reference group (e.g., males more likely to choose correct option, females more likely to choose distractor). A negative effect indicates that a distractor is disproportionately less attractive to the focal group (e.g., females more likely to choose correct option, males more likely to choose distractor). Consider the Sample Item in Appendix D, with the leaf, robot, fish, and a prompt that reads, "Point to the animal." If for the robot $\beta_{MH_3} = -0.72$, males are disadvantaged by this item because they are more likely to choose the distractor (robot) and females are more likely to choose the correct answer (fish).

Third, in multiple-choice items, β_{MH_j} is calculated separately for each distractor *j*, to identify which one (or more) is the location of the DDF effect and possibly the source of the DIF effect and potential bias. In the earlier example (see Appendix D, Sample Item), if it is known the β_{MH} value of DIF is $\beta_{MH} = -0.80$, it is known that the item favors girls. By also knowing the DDF for the robot is $\beta_{MH_3} =$ -0.72, further investigation of DIF can focus on why male children are more likely to select the incorrect answer (robot). In summary, a large and significant effect for an item, or a certain distractor, in either direction, indicates differential functioning is detected and the item should be further reviewed.

Relationship of DIF and DDF

In the MH-LOR framework, the DIF effect is a complex weighted aggregate of the DDF effects. The items with DDF effects can be examined in terms of how they provided unique or complementary information to DIF in multiple-choice items according to three potential patterns discussed by Penfield (2010). First, when all MH-LOR DDF effects are in the same directions and relatively equivalent in magnitude, they are considered constant and the item exhibits uniform DIF effects. As an ideal example, if DDF effects are exactly equal across options, that value is the DIF effect. See Appendix C for a mathematical derivation of this relationship.

The second potential outcome occurs when DDF effects vary in magnitude but not in sign and the DDF is considered nonconstant. One distractor may have a moderate or large effect and one distractor may have a very small or no effect. This presence of nonconstant DDF leads to nonuniform DIF, which may or may not be detected using MH-LOR method. To understand how this occurs, most multiple-choice distractors are chosen to represent different levels of ability (e.g., full knowledge, partial knowledge and little/no knowledge). The systematic selection of different options by the different groups (e.g., females choose partial knowledge and males choose little/no knowledge) in a multiple-choice items, also produces a systematic difference in each group's ability estimates for that item, but that is unrelated to their total ability level. This pattern manifests as a relationship between DDF effects and a nonuniform DIF effect (Penfield, 2010), where relative differences in probability that vary across ability levels also vary by gender (Camilli & Shepard, 1994; see Figure 2). However, answer responses should

ideally only vary by ability level. Thus, when the form of DIF in an item is nonuniform, distractors must be evaluated for DDF effects.

Third, when two or more DDF effects correspond in magnitude, but differ in sign (i.e., each distractor is disproportionately selected by a different subgroup), the effects may cancel out to a non-significant uniform DIF effect (Penfield, 2010). This can be categorized as divergent DDF, where answer patterns can lead to nonuniform DIF. In some cases, divergent DDF can also create crossing DIF. These item characteristic patterns are summarized in Table 1. See also Figures 1-3.

Content Reviews to Explain DIF and DDF

Statistical methods such as MH-LOR for DIF and DDF are helpful in detecting group differences in rates of correct responses. After DIF and DDF are statistically detected, follow-up studies of item content, often called *content reviews*, are necessary to determine if the item is measuring irrelevant constructs (Dorans, 1989 as cited in Camilli & Shepard, 1994). Such studies, referred to here as content review, have been conducted on SAT tests to determine which items had evidence of gender-related DIF (Dorans & Kulick, 1983), and whether cultural factors influenced African-American students' answer choices (McGurk, 1951 as cited in Camilli & Shepard, 1994). Content reviews can include reviews by test developers, opinions from experts in the content area, and/or feedback from think-aloud conversations with test takers. In applying content reviews to DIF/DDF studies, for items with only an overall DIF effect, the content review should primarily examine the stem or overall item. For items with DDF, it follows that the content review should primarily explore those distractors with significant effects.

This process may be different for preschool children because it can be challenging to obtain consistent and accurate information of how they process information, especially with an unfamiliar assessor. However, information on conducting this second step with preschool level assessments has not been discussed in the literature and developing methods for this process is currently necessary.

Previous Research on MH-LOR for DIF and DDF in Early Childhood

While the literature regarding DIF and DDF has recently expanded, studies often address contexts beyond the early childhood years (e.g., Abedi et al., 2007; Banks, 2009; Kato et al., 2009). An important reason to expand research into early childhood assessment is the growing accountability for early education programs (e.g., Grisham-Brown, Hallam, & Brookshire, 2006; NRC, 2008). Studies that reviewed preschool level assessments generally used other statistical methods, lacked a DDF component, or did not incorporate content reviews. In particular there has been very limited use of MH-LOR for DDF and content reviews in the early childhood assessment, leaving a gap to be addressed by this study.

Limited contexts. Initial studies using DDF to evaluate test fairness began with examining whether gender and race subgroup differences in responses to college entrance exam items went beyond correct or incorrect item responses and were dependent on test takers' selection from among several multiple-choice options (Schmitt & Bleistein, 1987; Schmitt & Dorans, 1990; Veale & Foreman, 1983). More recently, DDF has been used to examine gender and race differences (Banks, 2009), dual language learners (Douglas, Atkins-Burnett & Vogel, 2008; Qi & Marley, 2009), as well as used sparingly in other contexts. For example, Abedi, Leon, and Kao (2007) detected DDF with LR in items

only on the second half of elementary and high school state reading assessments when comparing students with and without learning disabilities. This finding led to consideration of fatigue effects on long multiple-choice tests for students with learning disabilities (items were ordered by test publishers to alternate between easy and hard questions to encourage test completion); however other research on this data found the item type in the second session to be the possible source of DIF and DDF (Kato, Moen, & Thurlow, 2009). Poggio et al. (2005) compared item responses between students who took a paper-and-pencil format and students who took a computer-based format. They screened for DDF using IRT models. Further, the Banks and Poggio et al. studies focused DDF analyses only on items that displayed DIF. Without examining all of the items for DDF, items with only DDF effects (nonconstant, divergent) may have gone undetected.

Even though DIF and DDF have been used to make conclusions about various group differences in elementary, middle, and high school level tests (Abedi et al., 2007; Banks, 2009; Poggio et al., 2005; Willingham & Cole, 1997), as well as college entrance exams (Li, Cohen, & Ibarra, 2004; Schmitt & Bleistein, 1987; Schmitt & Dorans, 1990), there has been less research at the early education level. Screening for item bias has been considered important primarily in high-stakes or large scale testing (e.g., Camilli & Shepard, 2004; Ryan, 2008). Although early childhood assessments have not historically been used for high-stakes purposes, the growing national and state accountability requirements have led to legislation that mandates increased accountability for publicly-funded early childhood programs through performance measures (e.g., Grisham-Brown, Hallam, & Brookshire, 2006). Due to increased accountability, emphasis has been placed

on the quality of early child assessments, including validity issues such as test fairness, such as the absence of DIF as evidence of internal structure (NRC, 2008, chapter 7). Furthermore, few of these studies examined differences based on gender. Therefore, evaluating item fairness in preschool-level direct measures on demographics such as gender has become more important in order to bridge the gap between early childhood test development and emerging psychometric techniques designed for this purpose.

Limited methods. The existing but limited research on early assessment primarily includes evaluating teacher rating scales, and rarely uses MH-LOR, DDF, and content reviews. For instance, some studies on preschool assessment have included methods such as Structural Equation Modeling to detect DIF on teacher rating scales of preschool behavior (SEM; Douglas et al., 2008). Other researchers examined DIF in a teacher report of language skills with LR and IRT (Qi & Marley, 2009). Webb, Cohen, and Schwanenflugel (2008) conducted an initial DIF analysis, but not DDF, using MH-LOR, and then a follow up study with Latent Class Analysis and expert reviews on the Peabody Picture Vocabulary Test (PPVT-III; Dunn & Dunn, 1997). Other research on the PPVT entailed IRT analysis of DIF for item responses from prekindergarten (Restrepo, Schwanenflugel, Blake, Neuharth-Pritchett, Cramer, & Ruston, 2006). Studies on teacher reports of attained kindergarten readiness skills explored DIF with MH-LOR but could not explore DDF, as the answer type was dichotomous (Greenberg, Penfield, & Greenfield, 2008). MH-LOR is an advantageous method for screening items; however, questions still remain about the specific use of MH-LOR for DIF and DDF in early education contexts.

Limited content reviews. Although various statistical methods have been explored, content reviews have rarely been published to explain the statistical results. Those content reviews that were conducted, tended to examine assessments for elementary school and above (e.g., Ercikan, Arim, Law, Domene, Gagnon, & Lacroix, in press; Morell & Tan, 2009) by conducting think-aloud protocols with older test takers. Webb, Cohen, and Schwanenflugel (2008) conducted an expert review, but did not gather input from the young test takers in their study. Kato, Moen, and Thurlow (2009) and Haladyna (2004) asserted that more research is needed regarding the use of think-aloud and other follow-up procedures to interpret DIF and DDF results.

In summary, previous research on DIF and DDF has been limited because it has not: (a) evaluated multiple-choice test items for both DIF and DDF and comparing the results, (b) employed the MH-LOR method, (c) conducted and published a content review on flagged items, and (d) applied these methods to real data from an early education assessment.

Study Purpose and Research Questions

To review and summarize, an advantage of DIF is that it can detect group differences in correct and incorrect rates, and the advantage to DDF is that it can detect group differences in selection of each distractor. The two methods are related in that DIF is determined by the magnitude and direction of the DDF effects and an aggregate of DDF effects (Penfield, 2010). Previous research has used simulated data, this is the first study to explore the DIF and DDF relationship in multiple-choice items for real data in preschool assessment contexts.

Therefore, Study I comprehensively screened for item bias using both DIF and DDF to determine the congruency of the DIF and DDF analyses and identify potential problem items. This aspect was addressed empirically by comparing the results of using MH-LOR to detect DIF and DDF in the same multiple-choice items. To expand the literature in this area, Study I addressed the following two primary research questions: (a) Does MH-LOR provide similar, unique, or complementary information in regards DIF and DDF when screening for item bias in a real data set? and (b) Is the relationship between DIF and DDF using MH-LOR observed in a real data set consistent with the expected theoretical relationship that DIF is a weighted aggregate of DDF effects?

Study II evaluated how a content review of children's knowledge, preferences, and gender-based perspectives of the distractors may have contributed to the source of DIF. Follow-up interviews and expert panel reviews were conducted to inform interpretations of the statistical findings, and to help disentangle results of the DIF and DDF analyses. Creating and employing a content review process was warranted as this type of examination of the test-taking processes of low-income preschool children based on their responses to science items has been minimally explored. Study II aimed to answer the following research questions: (c) Does the content review process, used in conjunction with a DIF or DDF study, aid in identifying potential causes of bias? and (d) Does the content review process, used in conjunction with a DIF or DDF study, aid in determining whether DIF or DDF is the preferred approach, or if both are needed?

Data source for applied example. This methodological study used DIF and DDF to determine whether gender differences existed in the items of a newly developed preschool science assessment. There were several reasons for choosing the preschool

science assessment, which was developed for children in programs serving low-income families as a data example for the methods. First, validity of the content has otherwise been found to be appropriate for a test of preschool science (see Greenfield, Jirout, Dominguez, Greenberg, Maier, & Fuccillo, 2009). Second, during the time of the analysis, the measure was under development and therefore suggested revisions based on this study were feasible and consequential. Third, the science content and picture format of the items had potential to demonstrate gender-based preferences for images that were unrelated to ability. Finally, because children in early education programs for low-income families have demonstrated poorer performance in science than in other academic areas (Greenfield et al., 2009), and due to a recent increased focus on early science education (e.g., Cech, 2008; French, 2004, Gelman & Brenneman, 2004), there is a need to develop a direct measure of preschool science that is as reliable and valid as possible to evaluate science curricula interventions for low-income children at risk for poor academic achievement (Chen & McNamee, 2007; Chittenden & Jones, 1999; Popham, 2003). It is important to determine that the science constructs are being measured and scores interpreted in the same way for male and females, as to not further disadvantage one subgroup of this population already at-risk for lower academic achievement.

Thus, in addition to answering the main questions, together, Study I and II answered supplementary research questions that could aid in the test development process: (e) Is the detected gender DIF/DDF due to actual bias? (f) Where, how, and why does the item actually contain gender bias? and (g) Can the items be revised to reduce

bias, and how, or should they be removed? Finally, this study provided a generalizable way to find partial evidence of validity based on internal consistency in multiple-choice test items.

CHAPTER 2: STUDY I - QUANTITATIVE ANALYSIS

There is a mathematical relationship between the DIF and DDF using MH-LOR that implies that both methods are necessary for multiple choice items (Penfield, 2010). The purpose of this study is to quantitatively explore this relationship and the utility in conducting both methods in real data. The data for this DIF/DDF study were from a direct assessment of preschool science that was administered to children attending early childhood centers for low-income families. Comparison groups for the DIF and DDF were assigned so that male children were the reference group and female children were the focal group.

Method

Participants. Participants were 507 children (271 female, 236 male, $M = 48$ months, $SD = 6.7$, age range: 35 to 68 months) who attended Head Start centers in Miami-Dade County, Florida. Among the 507 children, 297 were recruited from 30 classrooms across six Head Start centers that were participating in a larger university-community partnership study regarding early childhood science. Given that a larger sample was required for these analyses, the remaining 210 children were recruited from an additional 14 classrooms from three Head Start centers in the same district. The overall ethnic/racial composition of the children in the final sample was approximately 77% African American ($n = 384$), 16% Hispanic ($n = 90$), and less than 7% from Caucasian, Asian, and other backgrounds ($n = 33$).

Participants were randomly sampled from the pool of children in the classrooms. First subgroups were created by stratifying classroom lists by age and gender, so that an equal number of 'older' (i.e., 48 months by September 1) and 'younger' preschoolers,

and an equal number of males and females were included. Then, about two to three children were randomly sampled from each stratum.

Eligible participants must have completed the Preschool Science Assessment (i.e., not incomplete due to lunch, fatigue, nonresponsiveness), and demonstrated native or sufficient levels of English (to eliminate the confounding effects of variation in English language skills). Children in the larger study were also assessed with the PPVT and those who could not pass the training items or establish a basal were not included. From an original sample of 640, a total of 133 children were dropped from the study: 46 due to low English proficiency (5 failed the PPVT-III), 30 due to non-compliance, 11 due to chronic absenteeism or tardiness, 8 due to lack of assent, 31 due to relocation, and 7 due to lack of parental consent, and 1 had a developmental language delay.

Measures. Primarily the preschool science assessment was used in this study. The receptive vocabulary scores were not analyzed but failing scores were used as an indication to exclude participants from the analysis.

Early science education. The *preschool science assessment* (Greenfield, Dominguez, Greenberg, Fuccillo, & Maier, 2010) is an IRT-based preschool science assessment that covers three science subcontent areas: life sciences; Earth and space sciences; and physical and energy sciences. The assessment also covers process skills such as observing, comparing, and reflecting. The measure contains 80 items, including 56 multiple-choice items, which were selected from a pool of 180 piloted items. Items were selected to cover a wide range of difficulty, the extent of subcontent areas and process skills, and higher point-biserial values (a measure of discrimination and item quality). There are 48 multiple-choice test items that have three possible options, seven

that have four possible options, and one item with five options. The non-multiple choice formats include six sorting, two classifying, one matching, one measuring, one ordering, and 13 open-ended verbal items. The assessment takes approximately 35-45 minutes to administer. A trained, independent assessor reads the prompts as the child points to the picture options, sorts card, or answers verbally. The assessor records all correct and incorrect responses.

The 80-item form demonstrated good reliability with this sample (Cronbach's α = 0.92). When examining only multiple-choice items, reliability was slightly lower (Cronbach's $\alpha = 0.89$). Scores were highly correlated with scores in mathematics (r = .66, $p < .01$), vocabulary ($r = .72$, p , .01), and listening comprehension ($r = .6$, $p <.01$), and moderately with teacher reports of science skills ($r = .37$, $p < .01$).

Receptive vocabulary. The Peabody Picture Vocabulary Test (PPVT-III) was used to screener for sufficient English skills among the 297 children from the larger study. The PPVT-III is a 5-15 minute receptive vocabulary assessment that is correlated ($r = .90$) with full-length intelligence assessments, such as Wechsler Intelligence Scale for Children-III, making it an efficient proxy measure (Dunn & Dunn, 1997). PPVT-III is also correlated ($r = .63$, $p < .01$) with other measures of vocabulary developed for use with low-income children (Greenfield et al., 2010). The child is presented with picture panels of four pictures and assessor says "Point to [target picture]." The item is correct if the child points to the picture of the target word.

Procedures. First, approval for these procedures was obtained from the university IRB. Next, classroom teachers were approached for consent to participate. Teachers from 44 classrooms agreed to participate. Consent was obtained for children in

participating classrooms, except two families who declined. At the start of the school year as part of the Head Start enrollment process, parents signed consent forms giving permission for their children to participate in program evaluation research. As part of an agreement in the larger research project to evaluate a science curriculum, a letter describing this specific project was sent to each parent with an option to decline participation in this particular study before commencing data collection. Child assent was obtained at the time of testing as described below.

To initiate testing, a trained graduate student or undergraduate asked each participating child if he or she wanted to play some games with pictures; those who agreed participated in a 35-minute assessment session. As the test was administered, all correct and incorrect responses were recorded in detail by the trained assessors. Children received neutral, consistent praise as they participated to encourage completion of the game-like test. For the multiple-choice questions, the child only needed to point to the correct answer. When children did not respond, said, "I don't know," or pointed to more than one answer, those response types were recorded. Children were then re-prompted once and their final answer was recorded. If, for the final answer, the child still did not respond, said, "I don't know," or pointed to more than one answer, then that item was not included for that child in the DIF or DDF analyses.

Data analysis. After data were organized and recoded, they were analyzed using the MH-LOR approach (Dorans & Holland, 1992; Holland & Thayer, 1988). Both DIF and DDF analyses were conducted in the DIFAS computer program (Penfield, 2005). For the DIF analysis, response data were dichotomized so that a correct answer was assigned a 1 and an incorrect answer was assigned a 0. For the DDF analyses, response

variables were first dichotomized for each distractor. This recoding process was programmed in SPSS syntax such that each correct answer equaled 1, the distractor *j* equaled 0 and all other responses were null. Then, a MH-LOR analysis was conducted simulating a 150-item DIF analysis to obtain MH-LOR values for each distractor. Among the 48 multiple-choice items that have three options, 28 were designed in a format where there was a top picture (see Appendix D, Item B), which was not intended to be an option. However, since a meaningful number of children selected this "top" picture in many items, it was considered a fourth option and coded accordingly, and therefore it appears that 35 items have four options.

The MH-LOR analysis controls for ability using a total score approach. For the DIF and DDF analyses, total score ability was calculated in two ways: total correct out of 80 items and total correct out of the 56 multiple-choice items. This procedure created four total sets of analyses and was helpful in examining multiple-choice items embedded in a test with other formats.

DIF analyses. In the first analysis, DIF effects were calculated by estimating ability as number of total correct items among all 80 multiple-choice and non multiple-choice items. In the second DIF analysis, DIF effects were estimated with ability based on the total correct among only the 56 multiple-choice items. The results for the DIF analyses were evaluated based on established effect size *and* significance criteria commonly used by Educational Testing Service (ETS) as described above. These criteria were chosen so that results could be compared to other studies that have used or will use MH-LOR procedures. Items with at least a moderate ($|\beta_{MH}| \geq 0.43$ and statistically

significant (different from zero) DIF effect were considered among the flagged or "studied" items.

DDF analyses. In the third analysis, DDF was estimated with ability as a total correct of all 80 items. Finally, the fourth analysis screened for DDF using the total correct of only the 56 multiple-choice items as an estimate of ability. To determine flagged items for content review, ETS effect size criteria were used to evaluate DDF results, but determination of flagged effects was modified. The effect size criteria were chosen so that results would be comparable to the DIF analysis in this study, and to other studies using MH-LOR. For some flagged DIF items, however, moderate effect sizes were just below the threshold for significance using ETS criteria. Theoretically, for an item to have DIF, at least one distractor must have a DDF effect, but small subsample sizes of children selecting certain responses limit the power to detect significance using the *z*-tests associated with ETS criteria. Therefore, for items with DIF and one other significant DDF effect, remaining DDF results were considered flagged for review using less conservative criteria: moderate DDF effects were $0.43 \leq \left|\beta_{MH}\right| \leq 0.64$ and had a low standard error (*SE; SE* ≤ 0.35); and large effect were $\left|\beta_{MH}\right| \geq 0.64$ with $SE \leq 0.35$.

Relating DIF to DDF. As noted, in the MH-LOR framework, the DIF effect is equivalent to a complex weighted aggregate of the DDF effects. The items were examined in terms of how the DDF effect sizes related to the form and magnitude of DIF according to the three potential patterns discussed by Penfield (2010): (a) constant DDF and uniform DIF, (b) nonconstant DDF and nonuniform DIF, and (c) divergent DDF and nonuniform (possibly crossing) DIF. When examining multiple-choice items with MH-LOR, potential bias in the latter two patterns of nonuniform DIF might be underestimated

without the additional DDF analysis. Thus, to evaluate this theoretical relation, results of the DIF and DDF analyses were compared and synthesized to determine the comprehensive utility of screening multiple-choice items for item bias using both methods.

Results

Descriptives. Mean raw scores based on the full 80 items (M_{Male} = 33.64, SD = 13.16; M_{Female} = 31.92, SD = 13.93) were not significantly different, t (505) = 1.43, p = .16. Reliability for all participants and for each gender group was .92 for the full 80-item scale. Mean raw scores based only on the 56 multiple-choice items (M_{Male} = 28.63, SD = 9.43; M_{Female} = 27.28, SD = 10.33) were not significantly different (t (505) = 1.53, p = .13). Reliability for only the 56 multiple-choice items was .89 for the entire sample, .88 for males, and .90 for females.

DIF results. The DIFAS (Penfield, 2005) program was used to estimate β_{MH} values as the measure of DIF for all 56 multiple-choice items. A large majority of the items had negligible effects at the "A" level. All DIF effects described below were flagged at the "B" level with moderate and statistically significant effects (i.e., ≥ .43 and significantly different than 0). A summary of significant item statistics are reported in Table 2.

In the first DIF analysis, where ability was estimated using total score correct on the 80 item test, DIF effects were detected in four items, one that favored females (Item 4) and three that favored males (Items 35, 44, 46). See Appendix D for item prompts and images. In other words, controlling for overall science ability level, female children were more likely than male children to answer Item 4 correctly (β_{MH} = -0.64, p < .001). Male

participants were significantly more likely than females to answer correctly on Item 35 ($\beta_{MH} = 0.49$, $p = .027$), Item 44 ($\beta_{MH} = 0.61$, $p = .03$), and Item 46 ($\beta_{MH} = 0.54$, $p = .032$).

In the second DIF analysis ability was conditioned on total correct multiple-choice items (see Table 2). The same DIF effects were detected in three of the same items (Items 4, 35, 46) with comparable effect sizes to the first analysis ($\beta_{MH} = -0.65$, $p < .001$; $\beta_{MH} = 0.60$, $p = .011$; and $\beta_{MH} = 0.55$, $p = .021$, respectively). Statistics for significant items are summarized in Table 2.

DDF results. In the first DDF analysis, where ability was conditioned on total correct among the 80-item scale, eight items had distractors that demonstrated DDF effects. Three items had incorrect images that were relatively more attractive to males (Items 4, 21, and 31). Female children were more likely to choose the correct answer, while male children were more likely to select certain incorrect options. For example, Item 4, where the child was asked to select two animals of the same species, males were significantly more likely to select option 1 of the dog and cat ($\beta_{MH_1} = -0.76$, $p = .003$) or option 3 of the fish and cat ($\beta_{MH_2} = -0.59$, $p = .032$), and females were more likely to select the correct option of the two cats. In Item 21, where children needed to select the material from which a wooden box was made, when answering incorrectly, males were more likely than females to select the top picture of the wooden box ($\beta_{MH_T} = -1.20$, $p = .004$) or option 1 of the small blocks ($\beta_{MH_1} = -0.49$, $SE = .25$, $p = .052$) than the correct answer of the wooden logs in option 3. When answering Item 31 incorrectly, males were

more likely to select option 1 of the microphone (β_{MH_1} = -0.52, *SE* = .25, *p* = .066) than the correct answer of the flower in option 3.

Five items contained images that were relatively more attractive to females (Items 35, 44, 46, 68, 74, and 77). Male children were more likely to choose the correct answer, while female children were more likely to choose from among the incorrect options. Three items were significant according to the ETS criteria (Items 35, 68, and 77). In Item 35, when answering incorrectly, females were significantly more likely than males to select option 2 of the leaf (β_{MH_2} = 0.67, *p* = .014) than the correct option 1 of the bird. In Item 68, female children were significantly more likely than males to select option 1 of the paper towels than the correct option 3 of the brick wall (β_{MH_3} =.58, *p* = .032). For Item 77, females were significantly more likely to select option 2 of the glass full of marbles (β_{MH_2} = 0.55, *p* = .014) than the correct option of the juice. Two additional DDF effects were flagged because they fit the criteria for low standard error (Items 44 and 46). In item 44, female children were more likely to select the top picture of the elephant (β_{MH_T} = 0.50, *SE* = .35, *p* = .153). In item 46, females were more likely to choose the top picture of fish bowl (β_{MH_T} = 0.48, *SE* = .32, *p* = .077). The option 3 image of the forest also had a large effect size, but was not significant due to a high standard error, and small subsample sizes (β_{MH_3} = 1.05, *SE* = .61, *p* = .084).

In the second DDF analysis, ability was conditioned on total correct multiple-choice items (see Table 2). Three of the same items (4, 35 and 46) demonstrated DDF effects. See details above and Table 2 for statistical details. An additional item, 74, was the only item to demonstrate divergent distractor DDF, where effects were of about the

same magnitude but in different directions. Item 74 asked children to choose the object that made light. In particular, male children were more likely to choose the game die in option 2 (β_{MH_2} = -0.57, *SE* = .40, *p* = .177) or to choose the fan in option 3 (β_{MH_3} = -.60, *SE* = .39, *p* = .111), while females were more likely to select option 4 of the crayon (β_{MH_4} = .58, *p* = .022). The DIF effect was therefore negligible (β_{MH} = 0.14, *p* = .29), indicating no direct gender difference in the likelihood of selecting the correct option of the flashlight.

Relating DIF to DDF. As predicted, the three proposed patterns of DDF effects surfaced: three items with constant DDF, five items with nonconstant DDF, and one item with divergent DDF. The three items that had constant DDF effects, also had moderate, significant DIF effects (Items 4, 44, and 46; see Appendix D). For example, Item 4 had a moderate, significant DIF effect (β_{MH_3} = -0.64, *SE* = .20, *p* < .001), and DDF effects that were moderate, significant, and consistent in magnitude and direction [(β_{MH_1} = -0.76, *SE* = .25, *p* = .001) and (β_{MH_2} = -0.59, *SE* = 36, *p* = .02), see Table 2]. In this item, female participants were more likely to choose the correct answer, while males were disproportionately attracted to both incorrect options. For items 44 and 46, DDF effects were relatively consistent in magnitude and at least one effect was sufficient in magnitude to result in a weighted aggregate DIF effect that was flagged at the moderate level. For example, in Item 44, effect size were relatively equivalent: a DDF effect that was moderate and significant (β_{MH} = 0.60, *SE* = .27, *p* = .03), a DDF that effect met the less stringent criteria at .35 (β_{MH_3} = 0.50, *SE* = .35) and two that did not (β_{MH_1} = 0.37, *SE* = .43 and β_{MH_2} = 0.61, *SE* = 50). In other words, controlling for ability, there was a gender

difference in the rate of correct response, where females were choosing among all distractors disproportionately more than males. For these three items, results indicate uniform DIF and that the source of DIF may rest in the question, the correct answer choice, or possibly a larger at-random guessing rate for the focal group(i.e., less discrimination; Penfield, 2010).

In the five items with nonconstant DDF effects (Items 21, 31, 35, 68, and 77), even though DDF effects were in the same direction within each item, the magnitudes of some DDF effects were moderate, while others were small. This reduced the weighted aggregate to a small or non-significant overall DIF effect as detected by MH-LOR, except for Item 35 which had a significant DIF effect because one distractor was sufficiently higher. For example, in item 21, the overall DIF effect (β_{MH} = -0.40, *SE* = .22, *p* = .04) was small, but one distractor had a moderate, significant DDF effect (β_{MH_3} = -1.20, *p* = .004), one had a moderate effect size and *SE* ≤ 0.35 (β_{MH_1} = -0.49, *SE* = .25, *p* = .066) and the other distractors did not have significant effects (β_{MH_2} = -0.40, *p* = .04). In these cases, although no statistically significant difference in a correct or incorrect response was detected, the values did trend toward a moderate effect in part due to one distractor with a moderate effect (large effect size, but high SE). Based on these results, a potential source of item bias that should be explored are the incorrect options with flagged DDF effects.

One item had divergent DDF effects in the analysis that considered ability score based only on the multiple-choice items. The results of Item 74 exemplify the possibility of distractor effects being in approximately equal magnitude, but opposite directions. These divergent moderate DDF effects (β_{MH_2} = -0.57, β_{MH_3} = -.60, and β_{MH_4} = .58) cancel

out the DIF effect ($\beta_{MH} = 0.14$, $p = .29$). This item has meaningful DDF effects, even approaching large effects, but this item would not have been detected using on a MH-LOR measure of DIF. Rather than DIF explained by DDF, for this item, the DDF is causing a nonuniform DIF effect. According to these results, there is a potential source of item bias from the incorrect distractors.

Discussion

Study I demonstrated the utility and importance of conducting both DIF and DDF in terms of overall test bias. With DIF alone, there was only one item in favor of females (item 4) and three in favor of males (items 35, 44 and 46), the test appeared to be functioning equally overall, yet very slightly in favor of males. With a DDF analysis in isolation, results would have revealed three items that favored females and four items that favored males. Even though this is slightly more items (seven), they are still quite balanced in favoring each group. When examining results of both DIF and DDF, in total, there were three items favoring females and six favoring males. This combination revealed nine items (versus four using DIF alone) that contributed to misestimating scores for both male and female children. There were also three more items favoring males than females. These results have implications regarding interpretation of test scores in the same way for both groups. Here it relates to preschool science ability, but it can relate the importance regarding other multiple-choice measures the importance of verifying that scores are measuring the same way for all test takers beyond correct/incorrect responses. The comprehensive analysis provided a more thorough and conservative screening of the items.

Study I also demonstrated two advantages for the utility of including a DDF analysis along with a DIF analysis. First, the results for three items with DIF were partially explained by their constant DDF effects. Items with this pattern gain complementary information from DIF and DDF. Although they must be further examined by the content review, the relatively consistent distribution of DDF across the response options provided theory-based reason to evaluate the overall item, stem, and correct option. The gender difference in rates of correct response is a function of one group selecting among all the distractors disproportionately to ability level.

Second, the DDF analysis detected an additional six items (than DIF alone) that had nonconstant DDF effects among distractors. Differences in the rate of correct response were a function of one subgroup choosing a certain distractor disproportionately, or in one case each subgroup choosing different distractors. Thus, the DDF analysis provided unique information in that it detected effects in the items that would not have been flagged for bias using only DIF. These items should undergo a content review process, in which the DDF effects can be examined for construct relevant patterns that indicate whether the incorrect options should be revised (or omitted). For example, in Item 21, the top picture was not an intended option; however, it was included as on option in the DDF analysis, because children were choosing it. Even though there was a small DIF effect size ($\beta_{MH} = 0.40$, $p = .13$), a large DDF effect size was found ($\beta_{MH_3} = -1.20$, $p = .004$), where male children were disproportionately more likely to select the top picture. While 37 males and 35 females selected the top picture, since DDF controls for ability, results indicate that more of the males at higher levels of ability were likely to choose the top picture for reasons unrelated to their science ability. Therefore,

quantitative analyses supported that the three DDF patterns exist in real data and supported the recommendation to supplement DIF analyses of multiple-choice items with DDF Penfield (2010).

However, a detailed content review was necessary to further evaluate why these patterns occurred and reasons for bias in this applied example because statistical results alone are not conclusive of bias. A second step of conducting a content review of items is necessary (Camilli & Shepard, 2004; Penfield & Camilli, 2007). Each item that had a significant and moderate to large DIF or DDF effect size by the statistical procedures was examined in Study II to further evaluate the source of the difference according to these patterns, and substantiate any reasons for bias in this applied example.

CHAPTER 3: STUDY II - CONTENT REVIEW

Each item that was flagged by the DIF and/or DDF statistical procedures was examined in detail to determine the sources of the difference and whether the item was possibly biased toward one gender or the other. This two-part procedure consisted of a set of interviews with participating children and an expert panel review. The two fold purpose to the content review was to evaluate the conceptual framework of the DIF/DDF relationship theory and to do so, analyze the items under study in the applied example.

Method

Participants. Interview participants were 48 children who were not previously tested in Study I (M = 56 months, SD = 6.5; 25 female, 23 male, age range 42 to 65 months) and were sampled from 25 classrooms and across seven centers that participated in Study I. In most cases only 2-3 children in each age/gender subgroup were available for interviews, so their selection was predetermined. When classrooms had more than four children per subgroup available for selection, the same sampling method as Study I was used. Sixty-four interviews were commenced or conducted and a total of 48 interviews were retained for analysis. Fourteen interviews were excluded because the child had difficulty responding to the interview questions, or because the child's responses would provide limited validity – the child gave the same answer to every question, even after the interviewer re-prompted.

Measures. The *interview protocol* was designed to assess the children's (a) knowledge, (b) preferences, and (c) perspectives of the items. To assess their knowledge, first children were asked to identify images and tell what they knew about them or where they had seen it before. For some items, they were asked about how that image related to

the construct on the item (i.e., relation of the prompt to answer choices). For example, in Item 31, for each image, children were asked, "Is it living?" Then, to assess their preferences, children were asked to rank the images from favorite to least favorite, and why they chose each image. Finally, children were asked to classify images as *for boys*, *for girls*, or *for boys and girls*. These classifications were intended to provide a measure of how they associated gender with each image. When a child indicated an image was *for boys* or *for girls*, they were asked why and those responses were coded.

Interview Procedures. In order to establish objectivity, the primary author trained a second researcher to conduct interviews. The children were interviewed by the second researcher who was familiar with the test items, testing procedures, and working with Head Start children, but not familiar with the statistical results of the DIF and DDF analyses. This level of insight ensured the interviewer would expand on questioning in a consistent, concept-relevant, yet unbiased approach.

Interviews lasted from 30-60 minutes; longer sessions were conducted over two mornings. During the first few pilot interviews, multiple techniques for asking questions and obtaining the necessary information were tested with several children before selecting a protocol. After several children were interviewed, the protocol was further adapted to improve the quality of responses. These questions were scripted yet flexible enough to accommodate the variation in information the children supplied, because it is recommended that the questioning and tasks with preschoolers be conducted in a play-style manner (Gullo, 1994). The final protocol included both closed- and open-ended questions (see Appendix F). The primary researcher was also present at the interviews, and observed and recorded the interview responses.

The session began with a shortened version of the regular assessment, that consisted of the nine studied items (those with DIF/DDF) and eight additional randomly selected items from the remaining 71 items. All the item types were included to help the child establish rapport with the interviewer, and to practice expressing answers and sorting images. Immediately after each studied item, the child was asked only, "Why did you choose the ____?" in order to get an initial response justification without disrupting the pace of the test administration. Since this question may have caused the child to think they answered incorrectly, extra non-studied items were included so that the children did not feel they answered all items incorrectly. In addition, neutral feedback was provided throughout the session. This section took about 5-10 minutes. The entire assessment was not administered because the length of time necessary (35-45 minutes) to administer the full 80-item scale, or the 56 multiple-choice items, in addition to the interview time would have exceeded the attention span of average preschool children.

After the 17 items were administered, the interviewer returned to the studied items for more extensive questioning. A script was provided, but if a child's response required further elaboration, the interviewer was allowed to ask additional questions, especially questions relevant to gender-related comments (see Appendix F).

Expert panel procedures. The expert review panel consisted of six early childhood professionals who were experienced in early childhood science, and/or had experience with low-income or at-risk populations, as well as both male and female panelists. The panel included three early childhood science teachers (including one clinical child psychologist) and three directors of early childhood programs. The panelists were two male and four female experts who varied in years of experience

working with preschoolers (M = 11.8 years, SD = 6.11, range 5-20 years) and represented various ethnic backgrounds (e.g., 3 Caucasian, 2 Hispanic, and 1 African-American) and geographical regions (e.g., South Florida, Maryland, New York and Michigan).

The expert panel was conducted in two stages: a blind review and an informed review. First, for the blind review, each reviewer was sent a packet with the nine studied items along with six randomly selected multiple-choice items from the remaining 47 items. Each reviewer was instructed to decide if an item was biased and elaborate in a brief questionnaire for each item (see Appendix E) to provide the reason and suggestions for revisions to reduce bias. Second, after the expert submitted the blind review responses, each received the informed review packet. The informed review packet contained the nine studied items along with the results of the statistical analysis expressed in lay terms and a description of the frequencies from the interview information regarding children's knowledge, preferences, and perspectives. Reviewers were then asked to decide if the interview responses changed their opinion on item bias and elaborate on their opinion and suggest changes (see Appendix G).

Interview data analysis. The first step of the qualitative analysis examined the closed-ended responses regarding the children's (a) knowledge, (b) preferences, and (c) perspectives of the items. Information was gathered in a way that could be analyzed and compared across genders. Questions were developed to obtain both closed answers, which were generally examined as frequencies, and open answer responses, which were coded as described below.

Knowledge. This analysis included determining whether children could correctly identify or name the images, knew about them (e.g., what an object does, a tool's

purpose) and how they fit into the context of the item (e.g., relationships of the prompt pictures to answer choices). In addition, the consistency between children's answers in their initial response and follow up responses was coded, evaluated, and compared across genders.

Preferences. Children's ranking of the images from favorite to least favorite was compared across genders, including consideration of the rank order and the percentage of each gender that chose the image as their first, second, or third choice. The reasons the child chose each image as their favorite were coded and analyzed as described below.

Perspectives. Children's classifications of images as *for boys*, *for girls*, or *for boys and girls* were examined for patterns and compared across genders. The reasons children designated an image as *for boys* or *for girls*, were coded and analyzed as described below.

When possible, results were transformed into quantifiable data in order to organize analyses (e.g., yes or no questions; items where children rank their preferences). However, when feedback could not be directly quantified, the text recorded was coded and analyzed using a program for qualitative data analysis called ATLAS.ti (Atlas**.**ti Scientific Software, 2004). This software aided in recognizing patterns by aggregating coded documents. For each item, a separate file for male and female responses was compiled, with each participant's comments indicated by their ID number to separate the text. All documents were linked as a single ATLAS.ti project so that reports could be complied by code, item, image, or other trend throughout all comments.

Themes that appeared were given a code and highlighted as they appeared in each document. Choosing key themes and patterns was based on the interview question

structure and the responses from the children (e.g., "___ are for boys", "I like ____", "Boys like to _____", etc.). A standard coding pattern was formed, yet codes were allowed to differ slightly by item when appropriate ("Flowers are for mommy" for Item 31). A total of 460 codes and 40 code families were created. Although qualitative coding is often conducted without a quantitative reliability, about 15% of the thematic coding was double-coded to train one additional coder and establish accuracy and consistency. Each code given by both coders was scored as 1 and each code only given by an individual coder was given a 0. Sum of all 1's was divided by the sum of the number of all 1's and 0's to provide a proportion of agreement. The rate of agreement was calculated for each item (M = .67, SD = .08, range .5 to .75). Differences were reconciled among the two coders. Given the complexity of the coding scheme, this rate was considered acceptable and then the remaining quotes were verified by the double coder.

Expert panel analysis. The expert reviewers' categorizations of bias/lack of bias were considered in explaining the quantitative and interview results, and in part to disentangle the study's main question regarding the use of DIF, DDF, or DIF and DDF. Comments were used qualitatively to support interpretations and recommendations.

Patterns from the interview and expert panel responses were used to determine whether theoretical reasons for DIF as determined by DDF were present in real data by explaining how elements were indicative of item bias irrelevant to the construct.

Results

This study incorporated a two-part method for conducting item reviews at the preschool level; first, interviews were conducted with children drawn from the same

population of the 507 participants in Study I. The interviews assessed three main aspects of the children's perspective on the studied items and their images: knowledge, preferences, and gender-based perspectives and compared these three aspects across genders. Overall, based on the interviews and review panel, main patterns emerged that partially explained some of the differences in the likelihood of correct response. Second, a panel of expert reviewers provided an initial judgment on bias and then feedback on DIF/DDF effects and interview responses. Response from both methods were synthesized and aggregated by item, and organized by three statistical outcomes. Some general patterns emerged from the qualitative review; some that parallel the theoretical patterns and others that are inconsistent or do not support the proposed patterns. To clarify, the mention of a group selecting or choosing a distractor is always after controlling for ability. See Appendix D for illustrations of all items.

Items with DIF and constant DDF. Three of the nine flagged items (4, 35, 44 and 46) had DIF effects, that were further explained by constant DDF effects. Further, three of these items (4, 35, and 44) had statistical effects in both sets of ability condition analyses. For example, in Item 4, where the child had to select two animals of the same species, females were significantly more likely to answer correctly (two cats). When answering incorrectly, males were significantly likely to choose from among both incorrect options (constant DDF effects). For the trends from the open-ended questions, girls were twice as likely to say "cats are girls/cat is a girl" (7 female/3 male quotes), which was said more often than "cats are boys/cat is a girl" (4/2). Girls were also more likely to say "I like cats/picture of cats" (18) than boys (11), and to say "girls like cats" (4/3) more than "boys like cats" (2/1). Similarly, when speaking of dogs, children were

more likely to say "dogs are boys/dog is a boy" (5), than the dog is a girl (0); they were also more likely to say "boys like dogs" (3), than "girls like dogs" (1). See also the classifications in Table 6. All together, more comments related girls to cats and dogs to boys. These results support the theory that the problem lies in the overall item and the correct response. Furthermore, almost no children knew the definition to the word "species" in the prompt.

Perhaps children in both gender groups were selecting at random, because additional evidence indicates preferences for the distractors that may have exacerbated the above effects. In the follow up interviews, most of the girls ranked the two cats as their favorite (see Tables 4 and 5), while most of the boys ranked the cat and dog as their favorite image. When asked to sort images by gender, 57% of boys regarded the cat and fish picture as being *for boys*. They may have not only failed to select the two cats, but also favored option 1 of the dog and cat because the dog is 'boyish," or option 3 of the cat trying to get the fish because this predatory action appears "boyish." In this case, the open and closed interview responses provided more information; in both reviews, all experts indicated the item as not biased with one exception. Also, based on the informed review, that one expert commented:

> Since in follow-up outcomes girls ranked the two cats as their favorite image, it is very likely that they were simply choosing it not because it is the right answer but because it is their favorite. Boys did the same things, that is choose their favorite with the exception that their favorites did not happen to be the right. To eliminate this common gender association, the same expert comments, "I would suggest to changing the pictures to other animals such as turtles, hermit crabs, etc."

Item 44, in which the child was asked to point to the missing part of the elephant (trunk), males were significantly more likely to answer correctly than females, who chose

about equally from among all incorrect options. Most of the boys ranked the elephant (top picture) as their favorite image. Most notably, when examining open-ended comments for patterns, there were 40 instances of attributing male gender to the elephant, referring to the elephant as he/his/him, with 25 quotes by male children and 15 quotes by female children. For instance, children made comments such as, "He's looking for his nose" or "Because his nose is missing." Only one female child made a comment referring to the elephant as female, saying, "It's in her mouth." This trend is indicative of a male gender tone for the overall items because similar attributions (i.e., use of "he" or "she" instead of "it") were only made in 18 quotes in total for all other items combined. One expert acknowledged that "as a culture we tend to anthropomorphize animals so much that I am not sure if there are animals for which there will be no construction of gender!"

Minimal evidence pointed to the distractor with the largest effects size (cat tail; *SE* = .51), as fewer than half the children correctly identified the distractor, while most of the girls ranked it as their favorite. In fact, none of the children who selected the cat tail as their favorite identified it as a cat tail (e.g., other labels included "feather," "bird," and "tail") In the informed review, three reviewers indicated that this item was biased, and some suggested reasons for bias in this item, such as, "Although I am less sure about this construction compared to the cats and dogs problem above, it seems that the boys in this case have greater familiarity with the image of the elephant." Qualitative evidence in this study points to the perceived masculinity of the elephant; possibly boys were devoting more attention overall to this item, but not necessarily to any particular distractor, or were more familiar with elephants.

In Item 46, males were significantly more likely to answer correctly. In addition, there was a large, but not significant, effect that females were more likely to select the fishbowl or forest. More girls (80%) than boys (48%) correctly identified the forest/woods, and more girls (72%) correctly identified the fishbowl than boys (48%). Most of the girls ranked the underwater scene as their favorite and most of the boys ranked the fish bowl as their favorite image. There were no trends identified in the open-ended comments to further elaborate. One expert commented in the blind review, "Boys might be more favorable to blue color of the gravel in a fish bowl, which is content irrelevant…so maybe changing it to a neutral color such as grey or burgundy would improve its fairness to boys and girls." In the informed review, one expert noticed a potentially construct irrelevant difference in male and female visual processing related to the overall item:

> The only explanation I see for this difference is construct irrelevant, as it relies on differences between the ways that boys and girls visually process details. Girls tend to see more details than boys. When comparing the fish bowl to the images, it is possible that differences in level of detail affected the choice girls made. Images 1 and 3 have a much higher level of visual detail, while image 2, the desert, has simplicity similar to the image of the fish bowl…make all three images similarly complex. All three should have the same amount of detail to be processed.

In general, the content review responses supported the theoretical view that for items with DIF and constant or limited DDF effects, the source of DIF can be found in the overall item structure, question, or in the properties of the correct response. Items 4 and 44 follow-up analyses indicated an overall item problem, where as the Item 46 follow-up analysis indicated the female children were selecting incorrect options about equally, with slightly more choosing the

fishbowl on top. There were however, some items where distractors with slightly higher DDF effects were contributing to the original flag for possible item bias.

Items with nonconstant DDF. There were four items (21, 31, 68, and 77) that had DDF effects, but no significant DIF effects and one item that had DDF effects large enough to create a DIF effect (Item 35). The nonconstant DDF effects were different in magnitude but in the same direction and the distractors were explored for more information on why they may be potential sources of bias.

In Item 21, children were asked to select the material from which a wooden box was made (small blocks, ropes, or wooden logs). When answering incorrectly, boys were significantly more likely than girls to select the top picture of the box or the picture of the small blocks than the correct option of the wooden logs. The content review revealed that most males ranked the wood box as their favorite image. Open-ended comments about the box were minimal and approximately balanced in terms of gender specification. For example, children said, "boys like to use the box" (4) and "boys like it" (2), as well as, "girls use/play/like the box" (4). Most females ranked the small cement blocks as their favorite, yet most males (62%) indicated that the blocks in image 1 are "for boys." Comments about the blocks were more general; for instance, "boys like to build" (7) and "girls like to build" (4).

In the blind review, one expert commented that the item is biased, because, "A boy who takes an interest in construction may be able to deduce that this [top image] is made of wood….This might be a difficult question for girls who have not taken such an interest." However, in the informed review, the same expert expressed this item was not biased. Another trend that emerged involved the boys' interest in, and enthusiasm for,

the incorrect image of the ropes. More boys (70%) than girls (48%) correctly identified the ropes. Half of the male children (48%) categorized the ropes as "for boys." Further, male children (32%) expressed comments such as "Because boys play with ropes," "Boys like to be cowboys, they need a rope," and "Because boys do like this with the rope on the horsy," (while they made lasso motions). Another expert noted in the blind review that in her opinion the item favors boys, "because all the images look hard and nothing is soft." The same expert continued to support this view in the informed review: "Boys are more likely to be exposed to [real] building materials than girls. The girls picked the small blocks because they might remind them of the building blocks used in school [block area]" Evidence was provided for boys selecting the wooden box based on preferences and the blocks based on perspectives, support detection of DDF effects.

In Item 31, children were asked to select the living thing among a microphone, a teddy bear, and a flower. Males were significantly more likely to select the microphone than the correct option of the flower, this DDF pattern was supported by preferences and perspectives data in the content review. First, boys most often ranked the microphone as their favorite image; although girls also ranked the microphone as their favorite image. Second, and more notably, boys (65%) and girls (32%) classified the microphone as *for boys*. For the correct answer of the flower, girls (56%) and boys (48%) indicated it was for girls (see Table 6). These rankings indicate selection for some boys may have been by preference rather than knowledge, because images were easily identified and identified equally among boys and girls. Another trend for this item related to knowledge base was that when asked if each object was living, only one male and two female participants correctly identified the flower as living. Further, only four boys (18%) and

two girls (8%) were completely consistent between their original item response and their classification of each image as living or nonliving. Only 25 % of boys and 35% of girls were consistent at least regarding the flower between their original answer and classifying as living or nonliving (see Table 3). Open ended responses did not reveal any additional patterns. None of the panelists suggested changing the microphone, but one expert suggested changing the flower to a more neutral plant to narrow the gender disparity. In the blind review, experts indentified the microphone as a potential cause of item bias:

> This is tough because a microphone, which I associate more with boys of that age, might seem alive in its ability to make sound and they may not know that plants are living….There is a chance that a girl would pick the nice, bright flower over the other two objects.

In Item 35, which asked children to point to what might eat a worm among a bird, leaf and butterfly, male participants were more likely to answer correctly (bird), and females were significantly more likely than males to incorrectly select the leaf. Several inconsistencies surfaced between DDF effects in Study I and among information gathered in follow up interviews. For example, , more female than male children correctly identified the images, except for the leaf, which 28% of females incorrectly identified as a 'flower' (see Table 3), yet, female children were more consistent in regards to their original answer choice and indicating whether the three options could 'eat a worm.' Most male children (43%) selected the worm as their favorite image and the leaf as their least favorite; most female children (40%) selected the bird and butterfly as their favorites, the leaf as their second choice, and then the worm as their least favorite (see Table 4 and 5). In categorizing, all children classified the worm is *for boys* (44%), or *for boys and*

girls (45% of males, 43% of females). The majority of females labeled the bird and leaf as *for boys and girls* (60% 52%, respectively). Females indicated the butterfly was primarily *for girls* (44%), or *for boys and girls*, (52%). The males were divided equally across categories for the bird, leaf, and butterfly (see Table 6). For these results, one would guess that boys should pick the worm and girls the bird or butterfly, but the DDF effect was that girls disproportionately chose the leaf. In the open-ended comments, as compared to trends for images, there were little to no trends to further explain why girls favored the leaf. Three girls made a statement related to having a leaf or it being 'for mommy.' Based on the informed review, one expert asserted:

> When girls looked at this problem, they were more likely to see the top item as the agent, and the bottom items as the things that are being acted on. So, the problem to the girls might have been interpreted as 'which of these does the worm eat' rather than the more passive (from the worm's perspective) 'what eats worms?' It seems to be a problem of direction of scanning rather than anything having to do with the individual objects…It would be worth re-arranging the images to see if a change in scan direction changes the results. What happens when the choices are on top?

It is possible that both male and female children were confused by the wording. Boys may have been confused to the extent that they guessed at random, or heard "worm" and chose the worm; while girls were slightly less confused but enough to think they needed to choose what a worm would eat, hence selecting the leaf. The feedback on this item supported the notion that a problem lies in the overall item structure, although some evidence points to knowledge differences regarding the leaf, which had the largest DDF effect.

Item 68 prompted children to choose the item that is "hard" among a teddy bear, paper towels and a cement brick wall. A DDF effect showed that female children were

more likely to select the paper towels than the correct option of the brick wall. Females correctly identified the answer images better than males. For instance, the largest knowledge gap occurred for the distractor with DDF where more females (92%) than boys (52%) correctly identified the paper towels. When asked whether each object was hard or soft, males were more often correct, but females were more consistent with their original response (see Table 3). Most of the girls ranked the teddy bear as their favorite and most of the boys ranked the brick wall as their favorite image. Some open comments were quoted relating to the teddy bear but no trend emerged. However, in the blind and informed review, all experts indicated the item as not biased overall and did not indicate any images as favorable or biased towards a certain gender. For this item, the effect of the paper towel choice is partially explained by greater knowledge by females of that image, and the boys' preference for the brick wall.

For item 77, children were asked to select what will mix with water among a glass of juice, glass of marbles and a candle in a glass. Females were significantly more likely to select the glass full of marbles than the correct option of the juice. There were limited patterns to support the DDF effect of the marbles as the source of DIF. Females were better than males at correctly identifying the juice (96% vs. 78%), the marbles (56% vs. 39%), and the candle (60% vs. 43%), although both groups struggled to identify the second image as marbles and the third image as a candle. About half of the children identified the candle as juice/drink. Most of the girls ranked the orange juice as their favorite (52%) and the boys ranked the orange juice and marbles as their favorite image (30%). In the informed review, one expert stated:

> I feel that it is due to the picture choices given. It seemed like girls were choosing an option that was the most different out of three choices. It was

> a simple process of elimination in which the candle and the juice look very similar due to bright color choices and the fact that it is hard to tell that is actually a candle and not something of a liquid substance.

Reviewers suggested to chang the color of the candle to a neutral, more recognizable one, and to present objects in their more natural state/containers. This item also follows the comment made for Item 46 about the difference in how male and female children process images with detail, so perhaps an option that is solid but not juice would reduce this effect.

Items with divergent DDF. One item had divergent DDF effects and provided a good illustration of how examining distractors for items without DIF is important. In item 74, which asks the child to select the object that can make light, both boys and girls were attracted to incorrect options. When not selecting the correct option of the flashlight, female children were significantly more likely to select the crayon, but males were more likely to select either the fan or the game die. Most of the females ranked the crayon, as well as the fan, as their favorite, while most of the males ranked the die as their favorite, and fan as second favorite. Overall, children were able to correctly identify images about equally, except slightly more females (92%) than males (74%) correctly identified the crayon. In addition, only 23% of boys and 31% of girls were able to correctly indicate if each object made light or not. When comparing these classifications to their original answer: only 21% of boys and 25% of girls were completely consistent for all images, and only 15% of boys and 19% of girls were consistent between their original item response and stating whether the flashlight made light. For the game die, 43% boys indicated it was for boys.

Item 74 is also the only item that had divergent DDF effects and to have support for bias from knowledge, preferences, and perspectives. It was also detected as significant in the analysis that considered score by total correct multiple-choice items. When not accounting for ability in other item types, both genders were likely to choose incorrect options. Both genders struggled to identify the images of the objects and which ones made light. Boys seemed to be choosing the game die and fan because they liked them and because they were seen as boyish. The girls on the other hand are slightly better at indentifying the objects and which ones make light, but often chose their favorite image of crayon, which many girls classified as *for girls* (32%). Based on the interviews, there were some construct relevant differences indicating that item should be modified to contain less gender preferred objects. However, only one expert considered this item to have bias in the reviews, "…since all of the items are red, this MIGHT actually help boys take the time to see all of them and look to what makes light…I would see if objects of multiple colors would reach the same conclusion," suggesting that perhaps the slight tone of pink on the red crayon paper provided a visual distraction for the girls. The incorrect answers may have been due to preferences for the objects used as distractors.

Discussion

In general, the information and patterns that emerged from the qualitative review supported the findings of the statistical analyses and the proposed theoretical relation between DIF and DDF effects in multiple-choice items. Information gathered in the content review for the items that had DIF effects with relatively constant DDF effects mostly supported the theory that problems with those items would be found in the prompt, stem or correct answer. For example, in items 3, 4, and 35, the trends described reveal why children of different genders were selecting the incorrect answers. On the

other hand, for items 4 and 44 preferences toward the images and construct regarding animals in the life sciences were potential sources of bias, where as in item 35 the confusing prompt or lack of knowledge regarding the construct may have contributed to the differences. Item 46 however, had minimal evidence from interviews to support boys choosing the correct response more often, experts expressed color and processing difference in males and females. For items without DIF but with non-consistent or crossing DDF, those results were supported by the content review. For item 21, children may have chosen according to preferences and gender categorization perspectives, while in Items 31, 68, and 77 they may have chosen according to knowledge and preferences. For Item 74, the one item with divergent DDF, children's knowledge, preferences, and perspectives contributed to the effect. For some items, reviewers' comments provided insight into reasons for DDF, especially for items where interview data was not helpful (item 46). On the other hand, the interview responses were generally more descriptive of the mechanisms contributing to potential item bias.

While it is possible to focus a content review on the parts of the item indicated by DIF and DDF results, item components relate together and a problem in one area may manifest as an effect in another. Interview data for all images in all items were collected in a non-leading fashion because the interviewer did not have knowledge of the specific effects for the studied items. This was also important as the method for gaining insight into the responses of preschoolers and was an exploratory process. Panelists were also given an opportunity to comment on items and their images before given the statistical outcomes and some of the interview results.

The Atlasti software and coding procedures for the open ended responses were helpful for in interpreting DIF and DDF in items. In particular, items with DIF and constant DDF, where no one image was a strong distractor or showed unique patterns in the closed-ended questions (ranking, sorting), the closed ended did answers did not provide as much explanation for these items (e.g., 4, 44) as for other items (e.g., 35, 74). Patterns that emerged in the code families based on the quotes from the children's statements helped revealed potential source of bias as described below.

Suggestions Based on Content Review

This study was able to contribute to considerations regarding item development in the preschool science assessment, as well as other picture-based measures for preschool age children. First, animals, more than inanimate objects, seem to produce stronger patterns of gender association by children of both sexes. For instance, in Item 4, cats were perceived 'as girls,' 'for girls,' and 'liked by girls' while dogs were perceived 'as boys,' 'for boys,' and 'liked by boys.' This was especially present in the cat and dog (image 1) and two cats (image 2), whereas in image 3 of the cat trying get the fish, some children attributed male gender to the cat, possibly because the cat was displaying more aggressive behavior in trying to get the fish. An overwhelming attribution of male gender (his/he/him) to the elephant was apparent especially among male children perhaps because elephants are large, rough and strong (item 44), which are characteristics associated with masculinity. Yet unlike many of the inanimate objects, there were no differences in children's knowledge and ability to label the animals. Overall interviews on these items revealed a need for item developers to choose animals carefully. Common animals are more likely associated with feminine and masculine labels. Unless it is

construct relevant to use such animals (e.g., cats, dogs), the use of less encountered, less gender-associated, but familiar animals would be preferred (e.g., horses, turtles). While many images were easily identified by all children, large proportions of children were not able to identify and label several images such as the cat tail (Item 46), the wooden logs (Item 21), and the worm (Item 35), which was often labeled as a snake. This may be due to unknown or unclear images, but additional follow up would be required to make this firm conclusion. In addition, while most images were simple or isolated, some were complex and possibly difficult for children to process quickly (option 1 of the coral reef and option 3 of the forest in Item 46).

For many items, the rates of correctly labeling images were low and differed by gender. In addition, for several items, rates of consistency were strikingly low between original response and follow-up expansion questions (e.g., Is it living? Could it eat a worm? Is it hard?). Specifically, this brings into question the accuracy for which the items are measuring the construct versus an unintended or irrelevant construct, such as guessing, image preference, or novel image. The determining factor in whether to retain or change an item is the construct relevance of the reasons. For example, in Item 68, neither group chose the image selected as being their favorite. Girls who did not know which image was hard, or understood *hard* as *rough*, chose paper towels because they were more likely to know what they were. In the author's review of the item 77, an additional factor might be the use of marbles as decoration in vases with flowers. This may be something female children have noticed and were confusing 'used with water' and being 'mixed with water. These issues could be considered construct relevant as they relate to knowledge of the item content and the respective choices. Items such as 68 had

construct relevant differences (possibility related to ability) and that item should be retained. In items 4 and 74, some children may have chosen an answer based on their gender preferences for the images, which is construct irrelevant and could suggest a need to omit or revise those items.

These analyses were intended to guide interpretation of the results from quantitative DIF and DDF studies; however, the process more generally contributed to knowledge of how to gather information regarding answer choices from preschool children in order to support quantitative results. The process brings into question the exclusive use of multiple-choice items with preschool-age children, especially if upon further investigation into the child's knowledge of the construct and the relationship among the images, their interview responses do not match their original answer choices. It is possible, that more reliable and valid scores are attainable with, or in combination with, open-ended and other answer formats such as verbal, sorting, and matching. In fact, the larger test (full 80 items) was more reliable ($\alpha = .92$) than the set of 56 multiple-choice items ($\alpha = .86$). If for efficiency multiple-choice items must be used, the experts' opinions suggest that pictures should be familiar (at least to most participants), clear and easy to process. If children have previous knowledge about an animal or object, they should be able to identify (and label) those images so that the item can tap only the intended construct and not require additional processing. Ideally, images would also be selected so that children are not required to inhibit their preferences. As suggested by a reviewer, images within one item should also have a similar level of detail and processing, because it has been found that females process details and males process general themes and schemas (Meyers-Levy & Maheswaran, 1991) and those gender

differences in such visual processing skills may be developed in early childhood (Wolin, 2003).

The content review aided in disentangling statistical results such as where the problem occurs in the item, why the problem manifests as difference in the rate of response and how to revise the item to reduce gender differences in some items.

CHAPTER 4: SUMMARY AND CONCLUSIONS

This study examined the use of DIF and DDF with MH-LOR and a content review in the context of screening for gender differences in a measure of preschool science knowledge. The preschool science assessment, was chosen because content validity had been established (Greenfield et al., 2009), the measure was under development, and the multiple-choice, picture-based format had potential to illustrate DIF and DDF. By examining the items for potential item bias, this study provided evidence for the validity of most items, and identified items that need further examination.

In Study I, the statistical analysis in this applied example addressed the primary research questions regarding the utility of adding the new method of DDF to traditional DIF analyses in screening for item bias in multiple-choice items. In Study II, the content review supplied details that could support the statistical findings. Thus, this study showed that a DDF analysis was helpful in explaining DIF items, as well as detecting items that did not have DIF.

Additional Findings

Ability level estimation. When studying the 56 multiple-choice items controlling for ability with the total correct on the full 80-items scale ("first set"), significant DDF effects were detected effects for eight items. When matching ability using the total correct out of 56 multiple-choice items ("second set"), an additional item with DDF was detected (74). Given the slight discrepancies in these findings, questions remain regarding the best method to control for ability when examining a subset of multiple-choice items from a larger scale. If results had been exactly the same, the relevance of this decision would be reduced; however, the way to estimate ability is worth further

consideration because the results differed slightly (i.e., the only item with divergent DDF was detected in the second analysis).

One consideration is the reliability of the scales and subscales. For example, reliability for this sample was slightly higher for the full 80-item scale (Cronbach's α = 0.92) than for the multiple-choice only scale (Cronbach's α = 0.89). This higher reliability would indicate a more robust score for estimating ability level. Conditioning on total test score resulted in more items detected overall (eight) than in the more conservative analysis with the multiple-choice scale score (four). Although reliability increased with the inclusion of the non-multiple-choice items, this may be an indication of the multidimensionality of the test, or could simply be due to the larger number of items. Further research outside the scope of this study is needed, such as exploratory factor analyses, to determine if two or more sets of skills are being measured with multiple-choice and non-multiple choice items.

A second consideration is the criteria used to evaluate the effects. The three items significant in the second set of analyses, were already detected in the first set of analyses (Items 4, 35, and 46). For the additional item in the second set (Item 74), distractors had a moderate, but not significant, DDF effect sizes in the first set. Evaluating DDF effects with an effect size only system (as used with the DDF effects in significant DIF items in Study I) would have flagged Item 74 in the first analysis. The synthesis of quantitative and qualitative data suggests the total score is less sensitive when using stricter criteria, because items just below the threshold for a moderate effect had DDF effects that were well-supported by content review data.

Content areas. The nine studied items covered all three subcontent areas, but were not evenly distributed. Five items were life sciences, three were physical and energy sciences, and one was Earth and space sciences. The use of various animals and plants and distractors for items about living things tended to create more DIF and DDF effects than other topic areas. As discussed, children may assume that certain animals are of a certain gender. Regardless of whether this is natural or socialized, it is especially important to carefully choose pictures when creating science items. It is also important for other areas such as physical materials.

Limitations

Resources for examining sub-threshold effects. While the studies addressed the research questions, the studies had limitations such as limited time and resources. Among the remaining items without statistically significant DIF effects according to the ETS standards, another 10 items in the first DDF analysis and eight items in the second set had one distractor that met the less stringent DDF effect criteria. Two of the items, one flagged for DDF only on the first set (Item 38) and one flagged for DIF (Item 40) on the second set, were piloted for the interview. Due to the images being very similar, the children's responses about each image were similar, and the additional information was minimal. Because time and resources for interviewing the children were limited, these 18 items were not included in this content review.

Power for DDF. A well established advantage to MH-LOR in dichotomous items is that it can detect uniform DIF with moderate sample sizes. A disadvantage to MH-LOR in dichotomous items is that it cannot detect nonuniform and crossing DIF (e.g., Camilli & Shepard, 1994; Greenberg, Penfield, & Greenfield, 2007; Hidalgo &

Lopez-Pina, 2004). Typically, other methods such as LR are required to detect nonuniform DIF. In the case of multiple-choice questions, however, using MH-LOR method for DDF as a supplementary analysis can provide an indication, but not a measure of nonuniform DIF and crossing DIF. When item level response data are available, this can replace the need to conduct LR or other methods of DIF that detect nonuniform DIF, and provide information about potential source of bias in an item. For this study it was estimated that 250 participants per subgroup would be sufficient for stable effect size estimates in the DDF analysis as well. For many items the sample size was sufficient, however, several of the sub-threshold effects described above may have been significant with a slightly larger sample size. On the other hand, this study revealed some items that had DDF effects trends, such as large effect sizes, but also had high standard errors. This was often because some incorrect options were selected by too few children for the power to detect effects. For example, in Item 46, option 3 of the forest was selected by only 5 males and 15 females out of the 507 Study I participants.

Increase Diversity. While the students in the sample came from low-income at-risk populations, and included African-American, Hispanic and some other ethnicities, lack of more diversity was still a limitation to generalizing findings across all preschool populations. Due to sampling limitations, participants in the interviews were homogenous (100% African American). The sample was also limited to low-income children in one metropolitan area, not representative of the U.S. population. Furthermore, although PPVT was used criteria for sufficient English for inclusion in the current study, children with the lowest passing scores could have had difficulty understanding the more complex prompts on the science assessment, either because they

were dual language learners or had a developmental language delay that was not detected. Due to study design and resources, the additional 210 children recruited were not administered the PPVT. In future studies this should be consistent across the entire sample. Low English ability may have caused more guessing at random than among native English speakers. When working with diverse populations, it is important to use a language measure well validated with dual language learners (Qi & Marley, 2009).

Future Directions

The MH-LOR method of evaluating DIF and DDF in multiple-choice items should be replicated with larger samples. Items had high effect sizes, but also high standard errors because some incorrect options were selected by a small subset of children. This may indicate that sample size requirements adequate for DIF with MH-LOR are not large enough to have the power to detect certain patterns of DDF effects. Additional research with larger sample sizes (1000) to ensure adequate subsample sizes (approximately 500) may be necessary to further explore the relation between DIF and DDF using the MH-LOR with sufficient power.

On average the interviews took 30-40 minutes, with some sessions lasting up to one hour. The length of the interviews and limited time to complete them did not allow for a sufficient sample to compare by age as well. One anecdotal pattern by age that needs further examination was the task to categorize images as *for boys* and *for girls*. This 'genderizing' selection seemed to occur more often by older children (48 months by September 1) than by younger children who more often choose, "for boys and girls" for almost all items. In addition, due to the limited time and resources for interviewing the children, items that had only large or moderate effects but were not flagged as significant,

were excluded from this content review. In future DIF and DDF analysis with this and other assessments, more time and resources should be allocated for a content review of all items that match the less stringent criteria, or if more items are flagged as significant. This is important in order to obtain feedback that will allow revisions to increase the fairness of the test. Future studies should also include a larger sample of interviewees, not only for power, but also to examine a secondary demographic variable (e.g., age, ethnicity). Tests developed for national use, especially for widely-used accountability purposes, should gather data from nationally representative samples.

In addition, some items flagged for DIF and DDF effects were also items that generally performed more poorly (e.g., low point-biserial). Research is needed to develop a method to distinguish whether poorer item quality is in part due to DIF/DDF (Item 77) or if DIF/DDF occurs due to poor item quality (e.g., structure and wording of prompt in Item 35). Items that had DIF and/or DDF and a low point-biserial, also had more suggestions for improvements from expert panelists (e.g., Item 35).

Based on the information and suggestions gathered in this study, modifications will be made to current items and concepts will be applied to new items as they are created. When items are retested in future stages of item development, item property information will be available about the increased performance of these items overall, and any reduction in DIF or DDF effects, and any reduction in gender bias.

This study presents only one way to conduct a review. Other ways of exploring children's cognitive processes in answering multiple-choice picture items have been explored with older children (e.g., Ercikan, Arim, Law, Domene, Gagnon, & Lacroix, in press; Morell & Tan, 2009). These techniques included post-test, think-aloud interviews

where middle-school test takers described their thought processes and decision-making processes aloud as they answered the questions. This less directed interview style could be adapted for use with preschool children; however, young children may not independently express such ideas without specific prompting. The current study attempted to gather information about whether children had previously been exposed to the images in the test; however, this information would be more reliable with input from parents and teachers of the extent to which they believe the children are in fact familiar with those images. Another method to help gather children's knowledge and perspectives about the images could include an informal, small-group conversation format where the images from the test are embedded in a book and discussed with children in a play-like and divergent-response atmosphere. While this method was not feasible in this study, it could be incorporated in future DIF and DDF content reviews.

Recommendations

Given that (a) for five of the nine flagged items, only DDF effects were detected, and (b) for four items with DIF effects, three DDF effects were constant and one had nonconstant DDF, it is recommended that both DIF and DDF analysis be used when screening for item bias for multiple-choice items (Banks, 2009; Kato et al., 2009; Penfield, 2010). Each analysis provided unique information regarding the item response patterns. As Banks (2009) and Kato et al. (2009) suggest, for high-stakes assessments, it may be beneficial to conduct multiple methods of DIF/DDF analysis to counter balance the advantages and disadvantages of each method. Further research on other samples of

applied data would also be beneficial to further demonstrate the utility of using DIF and DDF when studying multiple-choice test items in assessments across various research disciplines.

REFERENCES

Abedi, J., Leon, S., & Kao, J. (2007). *Examining differential distractor functioning in reading assessments for students with disabilities.* Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: APA.

Atlas.ti Scientific Software. (2004). Atlas.ti: The Knowledge Workbench (Version 5.0). Berlin: Atlas.ti Scientific Software. http://www.atlasti.com.

Banks, K. (2009). Using DDF in a post hoc analysis to understand sources of DIF. *Educational Assessment, 14*(2), 103-118.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: SAGE Publications.

Cech, S. J. (2008). Fostering a 'science generation' seen as U.S. imperative. *Education Week, 27*(33).

Chen, J., & McNamee, G. D. (2007). *Bridging: Assessment for teaching and learning in early childhood classrooms, PreK-3*. Thousand Oaks, CA: SAGE Publications.

Chittenden, E., & Jones, J. (1999). Science assessment in early childhood programs. *Dialogue on early childhood science, mathematics, and technology education*. Washington, DC: American Association for the Advancement of Science.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. An NCME instructional module. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. New Jersey: Educational Testing Service.

Dorans, N. J., & Kulick, E. M. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach*. (ETS Research Report RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*(4), 309-319.

Douglas, A. R, Atkins-Burnett, S., and Vogel, C. (2008, June). Psychometric and measurement equivalence analyses of teacher rating scales and child outcome measures with a large culturally and linguistically diverse sample of preschool-aged children. Poster presented at the Head Start Ninth National Research Conference, Washington, DC.

Dunn, L. M., & Dunn, L. M. (1997). Examiner's manual for the Peabody Picture Vocabulary Test, Third Edition. Circle Pines, MN: American Guidance Service.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (in press). Application of think-aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement Issues and Practice.*

French, L. (2004). Science as the center of a coherent, integrated early childhood curriculum. *Early Childhood Research Quarterly, 19*(1), 138-149.

Gelin, M. N., Carleton, B. C., Smith, M. A., & Zumbo, B. D. (2004). The dimensionality and gender differential item functioning of the mini asthma quality of life questionnaire. *Social Indicators Research, 68*(1), 91.

Gelman, R., & Brenneman, K. (2004). Science learning pathways for young children. *Early Childhood Research Quarterly, 19*(1), 150-158.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26*(2), 147-160.

Greenberg, A. C., Penfield, R. D., & Greenfield, D. B. (2007, April). Comparing logistic regression and Mantel-Haenszel approaches in detecting uniform DIF in a large scale assessment. Poster presented at the National Conference on Measurement in Education, Chicago, IL.

Greenberg, A. C., Penfield, R. D., & Greenfield, D. B. (2008, June). Differential item functioning in a large-scale preschool science assessment. Poster presented at Institute of Education Sciences, Washington, DC.

Greenfield, D. B., Dominguez, X. Greenberg, A. C., Fuccillo, J. M., & Maier, M. F., 2010. Lens on Science: Development and initial validation of an item response theory-based assessment of preschool science knowledge and process skills. Manuscript in preparation.

Greenfield, D. B., Jirout, J., Dominguez, X., Greenberg, A. C., Maier, M. F., & Fuccillo, J. M. (2009). Science in the preschool classroom: A programmatic research agenda to improve science readiness. Manuscript accepted for publication.

Grisham-Brown, J., Hallam, R., & Brookshire, R. (2006). Using authentic assessment to evidence children's progress toward early learning standards. *Early Childhood Education Journal, 43*(1), 45-51.

Gullo, D. F. (1994). *Understanding assessment and evaluation in early childhood education*. New York: Teachers College Press.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications.

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Holland, P. W., & Thayer, D. T. (Eds.). (1988). *Differential item performance and the Mantel-Haenszel procedure.* Hillsdale, NJ: Erlbaum.

Jodoin, M. G., & Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Kato, K., Moen, R. E., & Thurlow, M. L. (2009). Differentials of a state reading assessment: Item functioning, distractor functioning, and omission frequency for disability categories. *Educational Measurement: Issues and Practice, 28*(2), 28-40.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19(3),* 5-15.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of Mathematics Items Associated With Gender DIF. *International Journal of Testing, 4*(2), 115-136.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute, 22*, 710-748.

Meyers-Levy, J., & Maheswaran, D. (1991). Exploring differences in males' and females' processing strategies. *Journal of Consumer Research, 18*(1), 63-70.

Morell, L., & Tan, R. J. B. (2009). Validating for use and interpretation: a mixed methods contribution illustrated. *Journal of Mixed Methods Research, 3*(3), 242-264.

National Research Council. (2008). Judging the quality and utility of assessments. In C. Snow & V. Hemel (Eds.), *Early Childhood Assessment: Why, What and How*. (pp. 181-231). Washington, DC: The National Academies Press.

Osterlind, S. J. (2006). *Modern measurement*. Upper Saddle River, NJ: Pearson Education, Inc.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement*, *29*, 150-151.

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement, 45(3), 247-269.*

Penfield, R. D. (2010). Modeling DIF effects using distractor level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement, 34*(3), 151-165.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay, & C. R. Rao (Eds.), *Handbook of statistics, Volume 26: Psychometrics* (pp. 125-167). New York: Elsevier.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper and pencil mathematics testing in a large scale state assessment program. *Journal of Technology, Learning, and Assessment, 3*(6). Retrieved 6/25/2007, from http://escholarship.bc.edu/jtla/vol3/6/.

Popham, M. J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

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. Early Childhood Special Education Online. Retrieved 10/19/ 2009.

Restrepo, M. A., Schwanenflugel, P. J., Blake, J., Neuharth-Pritchett, S., Cramer, S. E., & Ruston, H. P. (2006). Performance on the PPVT-III and the EVT: Applicability of the measures with African American and European American preschool children. *Language, Speech, and Hearing Services in Schools, 37*(1), 17-27.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Ryan, K.E. (2008). Fairness Issues in Educational Accountability. In (Eds.), K. E. Ryan, & L. A. Shepard (Eds.), *The future of test-based educational accountability*, (pp. 191-195). New York, NY: Routledge.

Schmitt, A. P., & Bleistein, C. A. (1987). *Factors affecting differential item functioning for black examinees on Scholastic Aptitude Test analogy items. Research report no. 87-23.* Educational Testing Service, Princeton, NJ.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27*(1), 67-81.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Education Measurement, 27*(4), 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Veale, J. R., & Foreman, D. I. (1983). Assessing cultural bias using foil response data: Cultural variation. *Journal of Educational Measurement, 20*(3), 249-258.

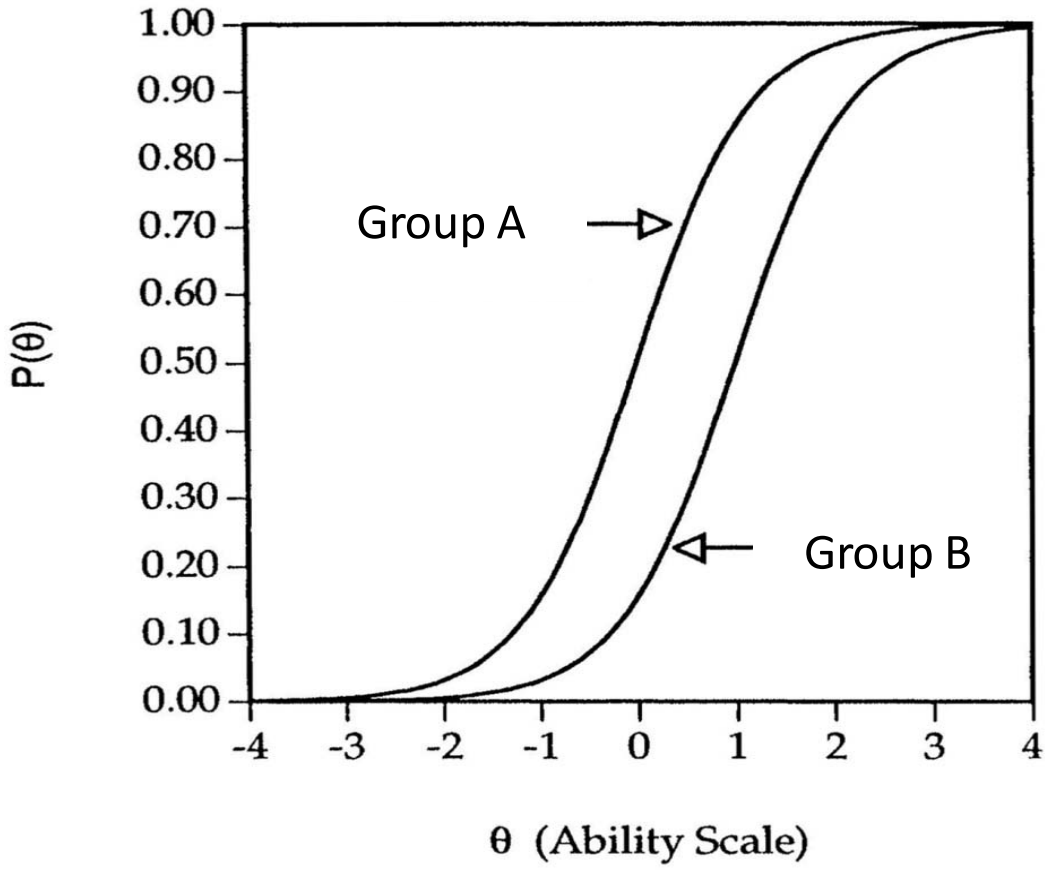
Webb, M. L., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test-III. *Educational and Psychological Measurement, 68*(2), 335-351.

Willingham, W. W., & Cole, N. S. (1997). Gender and fair assessment. Mahwah, NJ: Lawrence Erlbaum Associates.

Wolin, L. D. (2003). Gender issues in advertising: An oversight synthesis of research. *Journal of Advertising Research, 43*(1), 111-129.

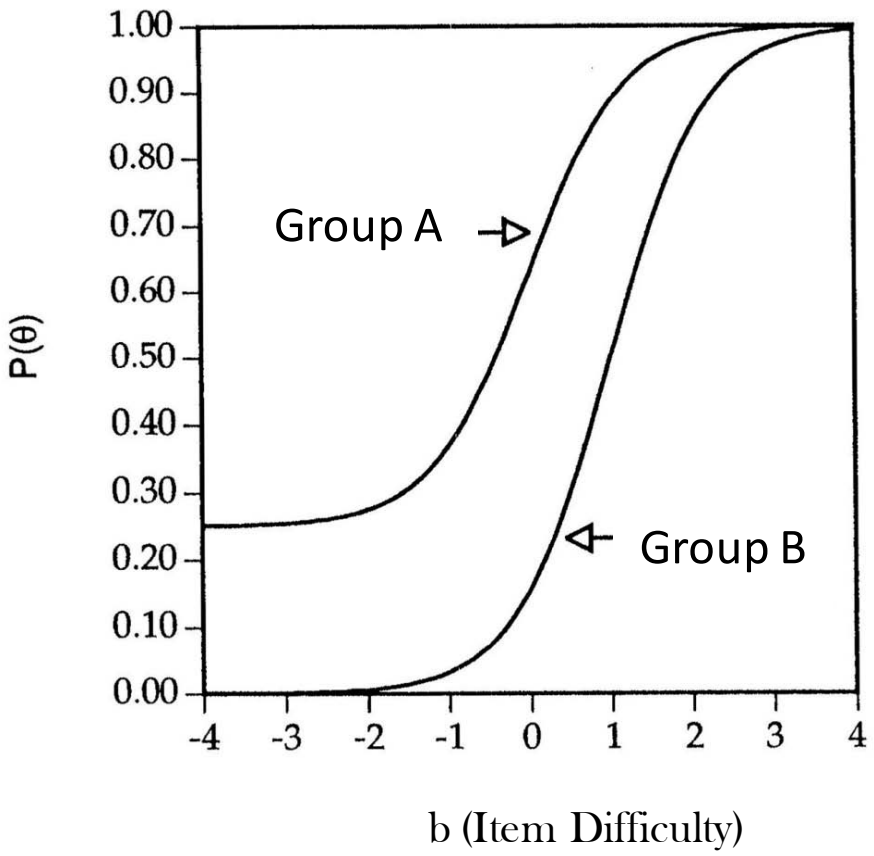
FIGURES

Figure 1. Uniform DIF occurs when there is constant DDF



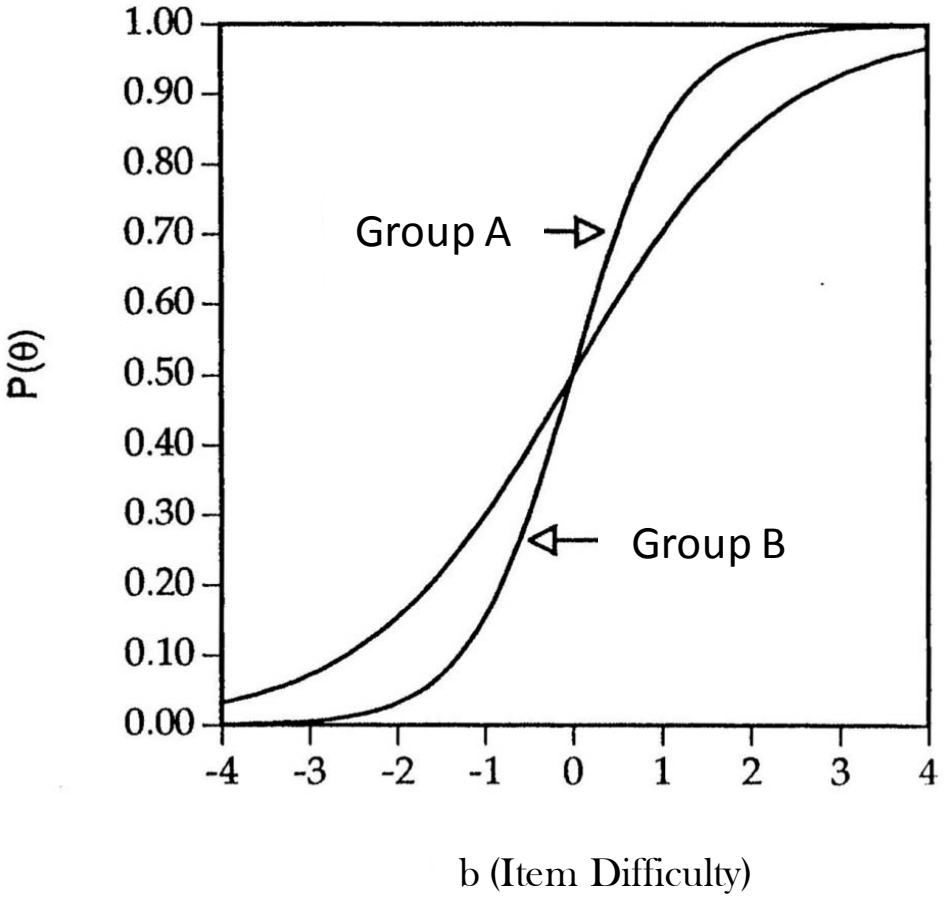
(modified from Camilli & Shepard, 1994)

Figure 2. Nonuniform DIF can occur when there is nonconstant DDF



(modified from Camilli & Shepard, 1994)

Figure 3. Nonuniform DIF, which can sometimes be crossing DIF can occur when when the nonconstant DDF is divergent DDF



(modified from Camilli & Shepard, 1994)

TABLES

Table 1.

Outcomes for comparison of DIF and DDF results for each multiple-choice item					
DDF	***DIF***	Group difference in	*Favors Males*	*Favors Females*	*Total*
Constant	**Yes**	probability of correct response	2	1	3
Nonconstant	**No**	choosing a certain distractor	3	2	5
Divergent	**No**	probability of choosing distractors (equal and opposite directions)	1	0	1
TOTAL			6	3	**9**

Table 2.

Summary of significant moderate DIF and DDF effects for all multiple-choice items

		DIF Effects					*DDF Effects*					
Ability score	*ITEM*	*MH LOR*	*SE*	*MH CHI*	*LOR Z*	*ETS*	*Distractor*	*MH LOR$_j$*	*SE*	*MH CHI*	*LOR Z*	*ETS*
Total scale	4	-0.64**	0.20	9.84	-3.17	B	1	-0.76**	0.25	9.13	-3.02	B
							3	-0.59**	0.27	4.15	-2.15	B
(80 items)	21	-0.40	0.22	2.89	-1.80	A	1	-0.49*	0.25	3.44	-1.94	A
							2	-0.35	0.34	0.79	-1.03	A
							T	-1.20**	0.41	7.55	-2.90	B
	31	-0.44	0.22	3.76	-2.00	A	1	-0.52**	0.25	4.04	-2.10	B
							2	-0.27	0.29	0.62	-0.92	A
	35	0.49**	0.22	4.54	2.21	B	2	0.67**	0.27	5.53	2.46	B
							3	0.13	0.30	0.09	0.44	A
							T	0.21	0.36	0.18	0.59	A
	44	0.60**	0.27	4.60	2.24	B	2	0.37	0.43	0.45	0.85	A
							3	0.61	0.50	0.96	1.22	A
							T	0.50*	0.35	1.53	1.41	A
	46	0.54**	0.25	4.20	2.15	B	2	0.21	0.43	0.08	0.49	A
							3	1.05	0.61	2.27	1.73	A
							T	0.47*	0.32	1.78	1.48	A
	68	0.43	0.22	3.60	1.95	A	1	0.05	0.30	0.00	0.17	A
							2	0.58**	0.27	4.47	2.14	B
	77	0.33	0.19	2.66	1.68	A	2	0.55**	0.22	5.72	2.46	B
							3	-0.02	0.26	0.00	-0.07	A
MC only score	4	-0.65**	0.20	10.13	-3.23	B	1	-0.84**	0.26	10.64	-3.29	B
							2	-0.62**	0.27	4.66	-2.27	B
(56 items)	35	0.60**	0.22	6.63	2.71	B	1	0.85**	0.28	8.11	3.01	B
							2	0.19	0.28	0.29	0.67	A
							3	0.38	0.36	0.74	1.07	A
	46	0.55**	0.26	4.06	2.16	B	1	0.32	0.43	0.26	0.74	A
							2	0.97	0.56	2.54	1.72	A
							3	0.52*	0.33	1.97	1.58	A
	74	0.14	0.20	0.33	0.67	A	2	-0.56	0.40	1.46	-1.39	A
							3	-0.60	0.37	2.21	-1.61	A
							4	-0.58**	0.25	5.11	2.28	B

* *SE* < .35. ** p < .05.

Note. A = ETS small effect, B = ETS moderate effect.

[a]Distractor is listed as image number from item, T = top picture

Table 3.

Percentage of males and females with knowledge of item images, labeling and describing; consistency with original response

		item and images correct			*construct correct*		*full consistent*		*partial consistent*	
ITEM	*Distracto*	*Males*	*Females*	*Total?*	*Male*	*Female*	*Male*	*Female*	*Male*	*Female*
4		87	88	88	23	27	21	27	13	6
	1	96	88	92						
	2	100	96	98						
	3	100	92	96						
21		22	24	23						
	T	100	84	92						
	1	65	68	67						
	2	70	48	58						
	3	30	24	27						
31		26	56	42	2	4	4	8	21	27
	1	83	92	88						
	2	100	96	98						
	3	96	96	96						
35		35	48	42	19	19	8	19	13	15
	T	61	84	73						
	1	83	100	92						
	2	78	64	71						
	3	96	92	94						
44		96	92	94						
	T	91	92	92						
	1	74	76	75						
	2	48	48	48						
	3	17	28	23						
46		87	80	83	13	15	8	15	2	2
	T	48	72	60						
	1	30	48	40						
	2	13	32	23						
	3	48	80	65						
68		91	84	88	15	25	15	25	8	6
	1	70	80	75						
	2	52	92	73						
	3	70	80	75						
74		87	84	85	23	31	21	25	15	19
	1	83	88	85						
	2	35	48	42						
	3	65	88	77						
	4	74	92	83						
77		65	56	60	10	6	15	17	8	8
	1	78	96	88						
	2	39	56	48						
	3	43	60	52						

Note. Construct correct = Child answered correctly when asked about image (e.g., Is it living?), Consistent = original response and construct decision are aligned, Partial consistent = original response and construct decision aligned for correct response

Table 4

Ranking of preference for images by male and female children

	Favorite	
Item	*Males*	*Females*
4	1	2
	2	1/3
	3	
21	T	1
	1	2
	3	T
	2	3
31	**1**	1
	3	2
	2	3
35	T	3
	3	1
	1	T
	2	2
44	T	3
	2	1/2
	1	T
	3	
46	T/2	1
	1	T
	3	3
		2
68	3	1
	2/1	2
		3
74	2	3/4
	3	1
	4	2
	1	
77	1/2	1
	3	2
		3

Note. Numbers are option numbers in items, T = top image

Table 5

Percentage of Males and Females who selected each image as their favorite

		Favorite picture	
Item	*Distractor*	*Males*	*Females*
4	1	43	28
	2	30	48
	3	26	20
21	T	39	20
	1	30	44
	2	13	24
	3	17	12
31	1	57	48
	2	9	28
	3	30	24
35	T	43	8
	1	26	48
	2	13	40
	3	9	4
44	T	39	16
	1	13	40
	2	22	20
	3	17	20
46	T	26	28
	1	26	8
	2	22	36
	3	13	24
68	1	48	20
	2	17	52
	3	17	24
74	1	30	16
	2	22	28
	3	17	28
	4	13	24
77	1	30	52
	2	30	28
	3	17	16

Note. Distractor column shows option numbers in items, T = top image

Table 6.

Percentage of Male and Female children's who categorized of images as "for boys", "for girls" or "for boys and girls"

		For Boys		*For Girls*		*For Boys and Girls*	
Item	*Distracto*	*Males*	*Females*	*Males*	*Females*	*Males*	*Females*
4	1	35	12	22	16	43	72
	2	39	24	26	24	35	48
	3	57	20	13	20	26	56
		0	0	0	0	0	0
21	T	26	20	22	24	52	56
	1	43	24	22	24	35	52
	2	61	48	9	8	30	44
	3	48	24	4	20	43	44
		0	0	0	0	0	0
31	1	65	16	4	32	26	52
	2	17	12	30	32	48	56
	3	13	12	48	56	35	32
35	T	39	48	0	8	43	44
	1	22	4	30	32	35	64
	2	30	16	22	28	39	52
	3	30	0	26	48	30	52
44	T	30	16	4	16	52	64
	1	30	12	9	28	52	56
	2	26	20	17	28	43	44
	3	30	36	17	16	43	44
46	T	22	8	22	28	43	60
	1	26	36	13	8	48	52
	2	35	12	0	24	48	60
	3	22	20	13	20	52	56
68	1	17	8	26	32	39	56
	2	22	20	13	12	48	64
	3	26	8	9	32	48	56
74	1	26	16	13	32	43	48
	2	43	32	4	24	35	40
	3	22	8	9	36	52	52
	4	17	4	22	32	43	60
77	1	26	4	9	24	43	68
	2	26	8	0	24	52	64
	3	9	12	26	32	43	52

APPENDICES

APPENDIX A

MH-LOR for DIF

Computation of MH-LOR using 2x2 contingency tables

For each item, and each total test score, *s*, a contingency table can be formed to illustrate the number of reference group members who responded correctly to that item (A_s), the number of reference group members who responded incorrectly (B_s); the number of focal group members who responded correctly (C_s), and the number of focal group members who responded incorrectly (D_s) as noted in Table B1. Using the notation of Table A1, first, the proportion of individuals with a correct response or incorrect response in each group can be computed as in Table A2.

Table A1.

Frequencies of correct/incorrect response at ability level s

Group	Correct (1)	Incorrect (0)	Total
Reference	A_s	B_s	$T_{AB\,rs}$
Focal	C_s	D_s	$T_{CD\,rs}$
			T_s

Table A2.

Proportion of correct/incorrect at ability level s

Group	Proportion Correct	Proportion Incorrect	Total
Reference	$A_s / T_{AB\,rs}$	$B_s / T_{AB\,rs}$	*1*
Focal	$C_s / T_{CD\,fs}$	$D_s / T_{CD\,fs}$	*1*

At each level of *s*, the odds of getting the item correct for the reference group is calculated using

$$O_{sr} = \frac{A_s / T_{rs}}{B_s / T_{\mathrm{rs}}}$$

and for the focal group is calculated using

$$O_{sf} = \frac{C_s / T_{fs}}{D_s / T_{fs}}.$$

In both equations, the *T* in the denominator will cancel out and the ratios reduce to

$$\frac{A_s}{B_s}$$

and

$$\frac{C_s}{D_s},$$

respectively. The next step is to calculate the ratio of the odds of someone in the reference group correctly responding to the item over the odds of a focal group member correctly responding to the item at each level *s*,

$$OR_s = \frac{A_s / B_s}{C_s / D_s}$$

$$= \frac{A_s D_s}{B_s C_s}.$$

The common odds ratio for the studied item is a weighted average of the odds ratios across *s*, given by

$$\hat{\alpha} = \frac{\sum w_s OR_s}{\sum w_s}.$$

Mantel & Haenszel (1959), proposed the optimal weight (w_s) of

$$w_s = \frac{B_s C_s}{T_s},$$

which yields the following form of the common odds ratio estimator

$$\hat{\alpha}_{MH} = \sum_{j=1}^{S} \left[\frac{A_s D_s \big/ T_s}{B_s C_s \big/ T_s} \right].$$

This calculation produces estimates on a non-symmetrical scale of 0 to ∞, with the value of 1 indicating no DIF. To make interpretation of DIF more meaningful and useful, the natural log transformation of $\hat{\alpha}_{MH}$ produces a value on the scale of -∞ to ∞, where zero is no DIF (Camilli & Shepard, 1994)

$$\hat{\beta}_{MH} = \log[\hat{\alpha}_{MH}] .$$

APPENDIX B

MH-LOR for DDF

Computation of MH-LOR$_j$ using 2x2 contingency tables

For each item, and each total test score, *s*, a contingency table can be formed to illustrate the number of reference group members who responded correctly to that item (A_s), the number of reference group members who selected distractor *J* (B_{sj}); the number of focal group members who responded correctly (C_s), and the number of focal group members who selected distracor *J* (D_{sj}) as noted in Table B1. The total number of test takers in each comparison J is note in the tables as $T_{XYj.}$. Using the notation of Table B1, first, the proportion of individuals with a correct response or incorrect response in each group can be computed as in Table B2.

Table B1.

Frequencies of responses correct/distractor J at ability level s

Group	Correct (1)	Distractor*J* (0)	Total
Reference	A_s	B_{sj}	$T_{ABj\,rs}$
Focal	C_s	D_{sj}	$T_{CDj\,rs}$
			T_{sj}

Table B2.

Proportion of correct/distractof J at ability level s

Group	Proportion Correct	Proportion Distractor J	Total
Reference	$A_s / T_{ABj\,rs}$	$B_{sj} / T_{ABj\,rs}$	*1*
Focal	$C_s / T_{CDj\,fs}$	$D_{sj} / T_{CDj\,fs}$	*1*

At each level of *s*, the odds of selecting the correct option for the reference group is calculated using

$$O_{sr_j} = \frac{A_s / T_{ABjrs}}{B_{s_j} / T_{ABjrs}}.$$

and for the focal group is calculated using

$$O_{sf_j} = \frac{C_s / T_{ABjfs}}{D_{s_j} / T_{ABjfs}}.$$

In both equations, the *T* in the denominator will cancel out and the ratios reduce to

$$\frac{A_s}{B_{s_j}}.$$

and

$$\frac{C_s}{D_{s_j}},$$

respectively. The next step is to calculate the ratio of the odds a member of the reference group correctly responding to the item over the odds of a focal group member correctly responding to the item at each level *s*,

$$OR_{s_j} = \frac{A_s \big/ B_{s_j}}{C_s \big/ D_{s_j}}$$

$$= \frac{A_s \, D_{s_j}}{B_{s_j} C_s} \, .$$

The common odds ratio for the studied item is a weighted average of the odds ratios across *s*, given by

$$\hat{\alpha}_j = \frac{\sum w_s OR_s}{\sum w_s} \, .$$

Mantel & Haenszel (1959), proposed the optimal weight (w_s) of

$$w_s = \frac{B_{s_j} C_s}{T_s} \, ,$$

which yields the following form of the common odds ratio estimator

$$\hat{\alpha}_{MH\,j} = \sum_{j}^{S} \left[\frac{A_s D_{s_j} \big/ T_{js}}{B_s C_{s_j} \big/ T_{js}} \right] .$$

This calculation produces estimates on a non-symmetrical scale of 0 to ∞, with the value of 1 indicating no DIF. To make interpretation of DIF more meaningful and useful, the natural log transformation of $\hat{\alpha}_{MH_j}$ produces a value on the scale of -∞ to ∞, where zero is no DIF (Camilli & Shepard, 1994)

$$\hat{\beta}_{MH_j} = \log[\hat{\alpha}_{MH_j}] .$$

Appendix C

Relating DIF and DDF using MH-LOR

Where DIF is defined as the group difference in the probability of a *correct response* given by

$$P(Y = j|\theta, G = 1) \neq P(Y = j|\theta, G = 0), \quad (1)$$

and a DDF effect as the probability of selecting each *j*th *distractor* is modeled as

$$P(Y = j|\theta) = \frac{\exp(-c_j - a_j\theta - G\omega_j)}{1+\sum_{j=1}^{J}\exp(-c_j - a_j\theta - G\omega_j)}, \quad (2)$$

then equation 2 can be substituted and reduced to the following equation where the weighted aggregate can be modeled as

$$\ln\left[\frac{\sum_{j=1}^{J}\exp(-c_j)\exp(-a_j\theta)}{\sum_{j=1}^{J}\exp(-c_j)\exp(-a_j\theta)\exp(-\omega_j)}\right]. \quad (3)$$

When *a* is constant, there is uniform DIF. When *a* is nonconstant, there is nonuniform DIF. A more extensive derivation is found in Penfield (2010).

APPENDIX D

Examples of Items

Sample Item
PE.10.Properties.4.Comp.a .V1
Prompt: Point to the one that is an animal.
Answer: Child points to 2

1

2

3

Item 4
LS.Species.1.Comp.d
One of these pictures shows two animals from the same species. Point to the picture that shows two animals that are the same.
Child points to 2

Item 21
PE.Properties.3.Refl.a.V2
Look at this box. (point to top) **Point to what it is made of.** (run finger through the bottom pictures)
Child points to 3

Top

1

2

3

Item 31
LS.03.Living.1.Comp.c.V2
Point to the living thing. *Child points to 3*

1

2

3

Item 35
LS.28.Food.2.Pred.a
Here is a worm. (point to the worm)
Point to what might eat a worm. (run your finger through bottom pictures).
Child points to 1

Top

1

2

3

Item 44
LS.20.Habitat.2.Obs.b.V1
This elephant is missing a part. Point to the part that is missing.
Child points to 1

Top

1

2

3

Item 46
LS.56.Systems.4.Refl.a.V1
Here is a fish bowl. (point to bowl) **The fish bowl is a model of which habitat?** (run your finger through bottom pictures) *Child points to 1*


Top


1


2


3

Item 68
PE.03.Properties.1.Comp.c (property: texture) .V1
Point to the picture of something that is hard.
Child points to 3


1


2


3

Item 74
ES.Sunlight.5.Comp.b
Point to the picture of something that makes light.
Child points to 1

1

2

3

4

Item 77
PE.Water.2.Know.V1
One of the things in these glasses can mix with water. Point to what can mix with water.
Child points to 1

1

2

3

Sample Verbal item
ES.20.DayNight.1.Desc.b
Prompt: How can you tell that it is nighttime in this picture?
Answer: Child says "because there is no sun", or "because it's dark," or "because there are stars/moon"

APPENDIX E

INTERVIEW PROTOCOL

For all non-studied items:
1) Administer item as usual and record child's response

Only after each studied DIF/DDF item
1) *Administer item and record child's response* _____
2) "You picked this picture. Tell me why you picked this picture."

After all items are administered, proceed with studied items

APPENDIX E (cont'd)

Set up cards in same format as each item. (Animal pairs)

1. What animals are in this picture? (point to first picture)

- ➢ **Are they the same?**
- If say "same," ➢**Tell me how they are the same.**
  - If unknown, **What do you think they are? What could it be?**
  - If incorrect, **That is a good guess; it's a dog and cat/two cats/cat and fish.**

  (repeat for next 2 pictures)

2. I'm going to read my question again. One of these pictures shows two animals from the same species. Point to the picture that shows two animals that are the same.
What does *species* mean? (reprompt for animal)

3. Do not point

- **Look at all the pictures again. Which of these animals have you seen at home or outside? You can pick more than one. Point to the pictures of things you have seen at home before.**

- **Look at all the pictures again. Which of these animals have you seen at school? You can pick more than one. Point to the pictures of things you have seen at school before.**

4. Look at all the pictures (circle images with finger)

- **Which picture do you like the most? Pick one that is your favorite.** (remove chosen picture)
- **Why is that your favorite?**
- **Now look at these pictures** (circle remaining). **Which do you like the most?**
- **Why do you like that one the most?**

You're doing such a good job listening.
5. Now want to know if you think each picture is for boys, for girls or for boys and girls.
Hold up each card in the following order.

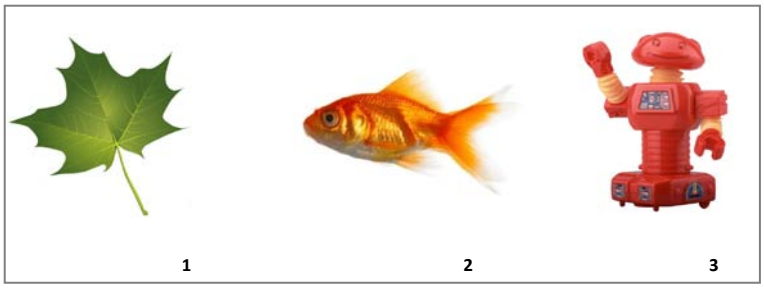
2 1 3

Do you think this for boys, for girls, or for boys and girls? (if NR, repeat once)

(If the child has trouble sorting) Take each card and ask, **Here are three boxes, one for a boy, one for a girl and one for a boy and a girl. Put each picture in the box where you think it belongs.**
(If the child still has trouble sorting) **Where does this picture go?** or **Which box does this picture go in?**
(if still not sorting) **Does this go in the boy, girl, or boy and girl box?**

APPENDIX F

Please mark your choice with an X. You may mark more than one choice. Please elaborate if you select options a-d for any items or images.



1

LS.Living.3.Compare

Point to the one that is an animal.

Child points to 2

A. Do you think this item is favorable or unfavorable to boys or girls?
YES_____ No_____

If YES, please continue:

B. Do you think any of the images are particularly favorable to or biased against boys or girls?

	Overall	Image 1	Image 2	Image 3
a. Favorable to boys	___	___	___	___
b. Favorable to girls	___	___	___	___
c. Bias against boys	___	___	___	___
d. Bias against girls	___	___	___	___
e. Fair to both genders	___	___	___	___

C. If you indicated any options a-d, please explain your reasoning, including support for whether the reason is construct relevant or construct irrelevant:

Overall

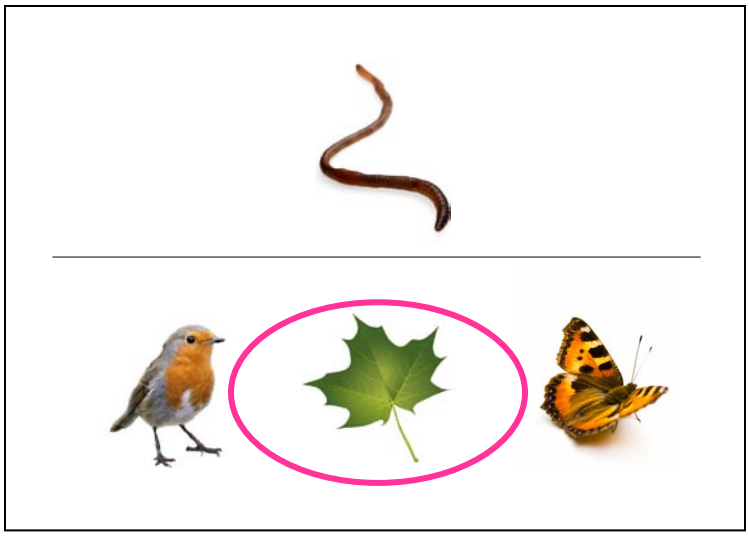
1:
2:
3:

D. If you found this item or its images to be unfair in anyway, how would you suggest revising the item to improve its fairness for both boys and girls?

Overall: 1 (4)

1:
2:
3:

APPENDIX G



6

LS.Food.2.Predict

Here is a worm. (point to the worm)

Point to what might eat a worm. (run your finger through bottom pictures).

Child points to 1 (bird)

Statistical outcomes:

- **Overall, boys were significantly more likely than girls to answer this question correctly.**
- **When answering incorrectly, girls were significantly more likely than boys to select the leaf.**
- **Girls answering incorrectly were significantly more likely to select the leaf than the correct options.**

Follow up outcomes:

- **Slightly more girls (100%) than boys (83%) identified the bird, and more girls (72%) identified the fishbowl than boys (48%); 28% of girls incorrectly identified the leaf as flower.**
- **When asked whether each option could eat a worm, 19% of boys and girls answered correctly, while girls (19%) were more consistent (8%) than boys with their original response.**
- **The most girls ranked the butterfly as their favorite and the most boys ranked the worm as their favorite image.**

A. Do you think that the item is actually bias against girls due to this difference?

YES____ NO____

B. If yes, please explain, including support for whether the reason is construct relevant or construct irrelevant:

C. If yes, how would you suggest revising the item to improve its fairness for both boys and girls?