# Construct Validity of the MCDI-I Receptive Vocabulary Scale Can Be Improved: Differential Item Functioning Between Toddlers With Autism Spectrum Disorders and Typically Developing Infants

**Cornelia Bruckner**
**Paul Yoder**
**Wendy Stone**
**Megan Saylor**
Vanderbilt University, Nashville, Tennessee

**Purpose:** To evaluate whether the validity of the Receptive Vocabulary scale of the MacArthur Communicative Development Inventory for Infants (MCDI-I; L. Fenson et al., 1991), a parent-report measure of early vocabulary, could be improved for children with autism spectrum disorders (ASD) by removing items that are biased.
**Method:** Logistic regression was used to identify biased items. Items are considered biased if characteristics other than those being measured by the instrument change the probability that a person will get an item correct. Participants in the current study included 272 typically developing infants younger than 18 months of age and 209 toddlers with ASD older than 18 months of age. The age difference between the 2 groups is a result of matching on total size of the receptive vocabulary.
**Results:** Twenty-five items were identified as showing large bias.
**Conclusion:** Deletion of these items from the test should increase the degree to which the authors are measuring the size of the respondent's mental lexicon with the total score from the MCDI-I.

**KEY WORDS: vocabulary, autism, modern test theory**

The MacArthur Communicative Development Inventory for Infants (MCDI-I; Fenson et al., 1991) is an important and frequently used measure of receptive vocabulary in research on children with autism spectrum disorders (ASD). Accurate measurement of the vocabulary of children with ASD is critical for evaluating progress toward the stated goal (U.S. Department of Health and Human Services, 2004) that 90% of individuals with ASD will develop speech by elementary school. In studies of young children with ASD, the MCDI-I (Fenson et al., 1991) is used commonly to match children (Coonrod & Stone, 2004; Leekam, Libby, Wing, Gould, & Taylor, 2002; Stone, Ousley, Yoder, Hogan, & Hepburn, 1997), measure change, and explain variance in other constructs (Aldred, Green, & Adams, 2004; Drew et al., 2002). The MCDI-I, a parent report of early language skills designed for use with typically developing 8- to 16-month-old-infants, includes a 396-item vocabulary checklist from which a score for receptive vocabulary can be derived. Besides being an important

research tool, the MCDI-I is also clinically important because delayed language development and unrepresentative performance during testing make scores from standardized language testing difficult to interpret (Charman, 2004).

Despite the fact that it has already achieved considerable status as a research and clinical tool, there is reason to be optimistic that we can improve the usefulness of the MCDI-I as a measure of receptive vocabulary in children with ASD and other diagnostic groups with similar characteristics, such as language delay. The method used to improve the MCDI-I in this study is an application of modern test theory (Lord & Novick, 1968): *differential item functioning* (DIF; Swaminathan & Rogers, 1990). *DIF* is the term used by the Educational Testing Service (ETS) to describe items that are biased across subgroups of test-takers (Donoghue, Holland, & Thayer, 1993).

## The Mental Lexicon

The *mental lexicon* is a dictionary of all the words in the receptive or expressive vocabulary of a child (Hoff, 2001). The MCDI-I vocabulary checklist was developed to estimate individual differences in the size of the mental lexicon in typically developing infants. Ideally, the items on the MCDI-I are equally representative of the lexicons of all children who are assessed with MCDI-I. However, even after controlling for total raw score on the MCDI-I, some of the words that are understood by most typically developing infants are different from words that are understood by most toddlers with ASD. When this occurs, reports that children understand particular words are thought to be influenced by characteristics other than overall receptive vocabulary size, making it difficult to estimate the size of their vocabulary.

## Characteristics of Group Membership

Characteristics of group membership that may reduce reported overlap between the mental lexicons of toddlers with ASD and typically developing infants are orienting deficits, social communication deficits, restricted object use, and chronological age differences. All four characteristics have been shown to distinguish children with ASD from typically developing children matched on developmental level (Mundy, Sigman, Ungerer, & Sherman, 1986; Sigman & Ruskin, 1999; Wetherby et al., 2004). In addition, there is reason to believe that each characteristic could affect the content of the mental lexicon or parents' ability to correctly interpret that content (see Table 1).

Because of the pervasiveness of differences in chronological age between groups of children with language disorders and typically developing children matched on receptive language, we will elaborate on the ways that a characteristic such as differences in chronological age can reduce overlap between matched groups in the content of the mental lexicon. The MCDI was developed for infants at the beginning stage of lexical development. Toddlers with ASD have delayed onset of verbal communication (American Psychiatric Association, 2000; Kanner, 1943); therefore, toddlers with ASD at the beginning stage of lexical development are likely to be older than the typically developing infants at similar language levels. Language delay in children with ASD makes it difficult to infer their understanding of words related to the

**Table 1.** Characteristics that could lead to reduced overlap between the mental lexicons of toddlers with autism spectrum disorders (ASD) and typically developing infants.

| Characteristic | Definition | Ramifications for the content of the mental lexicon | Ramification for parents' interpretation of that content |
|---|---|---|---|
| Orienting deficits | A deficit in the child's ability to redirect his or her attention to important events in the environment (Courchesne et al., 1994; Dawson et al., 2004; Mundy et al., 1986; Turner & Stone, 2005). | Reduces opportunities for symbol-infused joint engagement (Adamson et al., 2004). | Inconsistent orienting makes interpretation of comprehension of those referent labels difficult. |
| Social communication deficits | A deficit in communication for purely social purposes, such as greetings (Toth et al., 2006). | Larger proportion of the receptive vocabulary words used for requesting and protesting (Stone et al., 1997). | Children with ASD respond less frequently to social routines, and determining their receptive understanding of the words represented in these routines may be difficult (Sigman & Ruskin, 1999). |
| Restricted object use | Restricted range of action schemas or toy preferences (Bruckner & Yoder, 2007). | Not able to predict. | It is difficult to determine the understanding of labels for toys with which children do not play (Tomasello & Mervis, 1994). |
| Age differences | Toddlers with ASD are older than typically developing infants when matched on the size of their receptive vocabulary (Kanner, 1943). | Toddlers with ASD may be more likely to know words for objects and events more prevalent in the lives of toddlers. | It is difficult to judge a child's knowledge of words and events that are not prevalent in the environment (Tomasello & Mervis, 1994). |

experience of infancy. For example, it would be difficult to determine whether a child understands a word such as *bottle* if the child did not use a bottle. There simply would be fewer opportunities to talk about a bottle in the child's presence. Because of differences in the experiences of children in different groups at different chronological ages, we would expect a reduction in the overlap of the children's mental lexicons.

## Application of Item-Level Analysis to the MCDI-I

To determine which words are not representative of the common mental lexicon of toddlers with ASD and typically developing infants, we have to evaluate each word on the vocabulary checklist. To do so, we need to estimate the following: (a) individual differences in the size of the mental lexicon, (b) group membership, and (c) the probability of getting a word correct. To estimate the size of the mental lexicon, we use the total score on the Receptive Vocabulary scale of the MCDI-I. Group membership is categorized as typically developing or ASD and is determined by clinical diagnosis. To estimate the probability of getting a particular word correct, we solve a logistic regression prediction equation that includes the size of the mental lexicon, group membership, and the interaction between group membership and the size of the mental lexicon (Swaminathan & Rogers, 1990). If only the group effect is statistically significant and has a large effect size, then the item is defined as showing uniform DIF. If the interaction effect (Group × Size of the Mental Lexicon) is statistically significant and shows a large effect size, then the item is defined as showing nonuniform DIF.

An important feature of a DIF analysis is that it detects items that are functioning differently between groups controlling for ability. In the case of the MCDI-I, we can detect words that are functioning differently between groups controlling for the size of the mental lexicon. In this way, we do not identify words that are functioning differently due to vocabulary size or a potential language delay (the construct that we are trying to measure) but identify words that are functioning differently because of other characteristics of children with ASD that are not intended to be measured with the MCDI-I.

*Why identify items that show DIF is important?* Characteristics other than the mental lexicon that affect test performance cause *differential test functioning*. Differential test functioning occurs when the total score for a measure functions differently across groups (Raju, van der Linden, & Fleer, 1995). An item functions differently across groups if the probability of correct response differs not only as a function of the ability for which the test was designed to measure but also as a function of the group to which a person belongs (DIF). A test functions differently across groups if the amount of ability represented by the same total score differs across groups. One method for removing this type of contamination from the total score is to delete items that function differently between two subgroups. In this study, the subgroups are children that belong to different diagnostic categories.

It is important to delete only items that function differently between groups if they cause differential test functioning. Items that show only statistically significant DIF may not affect the total score. To determine whether items with statistically significant DIF also have an effect on the total score (differential test functioning), we compared the effect size for the DIF to a criterion effect size. In this study, the two criteria for item deletion included (a) statistically significant DIF and (b) a large DIF effect size. When items are identified with DIF, it is the practice of many test developers, such as the ETS (Donoghue et al., 1993), to remove that item from the test. In the next section, we explain how items with DIF may threaten the validity of a test.

*Why deleting items with DIF is important*. One ramification of not deleting items identified with DIF on the MCDI-I is that researchers and clinicians who use this scale to group and understand their patients will not be matching on receptive vocabulary when they match groups on the total number of words reported as understood. Characteristics of group membership, such as differences in the ability to orient to stimuli, can contaminate the error components of the total scores by replacing random variation with a systematic covariation with characteristics of group membership such as orienting. When the characteristics of group membership are uncorrelated with other variables in the experimental design, they will increase the risk of Type II errors. When characteristics are correlated with other variables in the experimental design, they will increase the risk of Type I errors (Ackerman, 1992).

## Specific Aim of This Study

The aim of this study was to determine whether there are biased items on the MCDI-I. Although the MCDI-I is a frequently used measure of expressive and receptive vocabulary in children with ASD, an extensive literature review revealed that no item-level analyses using modern test theory had been conducted in this population.

# Method

## Participants

The participants for this study were ascertained using de-identified records. (When records are de-identified, all of the child's information is removed from the protocol. For this study, all names were blocked out before the

records were identified in the database.) This protected the personal information of the participants and increased the number of records that could be collected. However, because no information was collected on the participants other than their responses to the items on the MCDI-I and their diagnostic category, we are not able to give a detailed description of the participants. This lack of specific descriptor information does not threaten the internal validity of the results because all that is needed for a valid test of DIF are item responses, total score, and group membership. However, the lack of participant descriptor information does limit our ability to identify the population to which the results may generalize.

*Typically developing participants*. The typically developing participants were 272 infants younger than 18 months of age. These infants were participating in research investigating language and cognitive development. Children were included if their parents reported that their development was normal.

*Participants with ASD*. The participants with ASD were 209 children older than 18 months. These children were drawn from a sample participating in research on the diagnosis of ASD. Children were included in the sample with ASD if they met criteria for ASD on the Autism Diagnostic Schedule (Lord et al., 2000) and on the basis of clinical diagnosis.

## Measure

Receptive vocabulary level was measured using the MCDI-I. The total score for receptive vocabulary was the sum of words understood and words understood and said. MCDI-I scores have shown sufficient reliability and validity in toddlers with ASD and typically developing infants (Charman, 2004; Fenson et al., 1994). The MCDI-I was mailed to the parents before a laboratory appointment or completed during a laboratory appointment for all children in the study. Only one test administration was entered for each child. To minimize data entry error, parents' responses to each item on the vocabulary checklist were double-entered into the database. Any disagreement between the two data entries were flagged and subsequently corrected. The scores on the Receptive Vocabulary subscale were judged to be primarily generated by a single trait (i.e., be unidimensional) based on eigenvalues computed using principal components analysis (Tabachnick & Fidell, 2001).

## Measuring Vocabulary Ability

The reliability and validity of MCDI-I scores, as defined by classical test theory, meet general standards of acceptability. However, a better measure of vocabulary ability is one in which items that show DIF are eliminated (Donoghue et al., 1993). To exploit the benefits of item-level analysis in the estimation of DIF, we removed items that showed DIF from the estimate of receptive vocabulary using a three-step process.

In the first step, we compared item response probabilities between groups using the sum of all correct items, the unadjusted total score for receptive vocabulary, as an estimate of receptive vocabulary ability. Items for which the probability of correct response for typically developing infants divided by the probability of correct response for toddlers with ASD was greater than 1 were identified from this first pass and removed from the total score to create an adjusted total score. This criterion is defined by the ETS as a "large" effect size (Longford, Holland, & Thayer, 1993).

For the second pass, we compared item response probabilities between groups using the adjusted total score as the ability estimate. Any additional items detected with the ETS's large effect size were removed from the estimate of ability after this second step to create the twice-adjusted total scores.

For the third pass, we compared item response probabilities using the twice-adjusted total scores as the estimate of ability. This third pass was the logistic regression procedure, and the parameters estimated in this third pass were used as our test of DIF.

## Testing for DIF

*Parameter estimation*. We used the standard logistic regression model to predict the parents' response to an individual word from the size of the mental lexicon, group membership, and the Group × Total Score interaction term (see Equations 1 and 2; Swaminathan & Rogers, 1990).

$$\Pr(u = 1) = \frac{e^z}{1 + e^z} \quad (1)$$

$$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g) \quad (2)$$

$\Pr(u = 1)$ is the probability that the parent indicates that the child understands the word, $\tau_0$ is the intercept of the regression equation, $\tau_1\theta$ is the effect of total score on the dependent variable, and $\tau_2 g$ is the effect of group on the dependent variable. If this parameter is statistically significant, it indicates uniform DIF. $\tau_3(\theta g)$ is the effect of the Group × Total Score interaction on the dependent variable. If this parameter is statistically significant, it indicates nonuniform DIF. We used EZDIF software (Waller, 1998) to calculate the logistic regression parameters for both groups.

*Testing statistical significance*. We tested the statistical significance of the logistic regression model using a chi-square test with 2 degrees of freedom. This chi-square compares Model A, in which there are no group

differences ($\tau_2 = 0$ and $\tau_3 = 0$) to the nested Model B, in which there are group differences ($\tau_2 \neq 0$ and/or $\tau_3 \neq 0$). The chi-square is an overall test that compares the fit of the two models. According to Model A, a person's response can be predicted by ability alone. According to Model B, a person's response can be predicted by ability, group membership, and the Group Membership × Ability interaction. A nested chi-square with 2 degrees of freedom is used to test the hypothesis that adding the group and interaction parameters significantly improves the fit of the model to the data. If the addition of the group and interaction parameters does not improve the fit of the model to the data, the hypotheses of a statistically significant group and interaction term are rejected simultaneously. If freeing up the two additional parameters does improve the fit of the model, the statistical significance of the group and interaction parameters are tested individually (Swaminathan & Rogers, 1990).

*Effect size measures*. Two metrics for effect size were used in this study. The first metric was used to adjust the total score before the logistic regression procedure was implemented. The second effect size metric was used for coefficients calculated during the logistic regression procedure. Both metrics quantify the magnitude of the group differences in item response probabilities.

The ETS effect size measure delta metric DIF (D-DIF) is a continuous measure; however, ETS (and the EZDIF software) trichotomizes the continuous scale into small, medium, and large effects. A large effect is defined as an absolute value of the D-DIF statistic of at least 1.5 and significantly greater than 1 ($\alpha = .05$; Longford et al., 1993). D-DIF is a ratio of the odds of correct response for the typically developing infants and the toddlers with ASD. We calculated D-DIF using EZDIF (Waller, 1998). There is evidence that applying this criterion is useful for identifying DIF that is true in the population (Donoghue et al., 1993).

The measure of effect size used for the logistic regression procedure is based on item response theory. The logistic regression functions are plotted separately for each group, and the squared weighted difference between the two regression functions is calculated. The value of the noncompensatory DIF index (NCDIF; Raju et al., 1995) statistic can be interpreted as the ratio of the difference between the two functions to all the difference that is available. The difference between the two groups' functions is weighted by the number of participants, at that ability level, in the group of toddlers with ASD. In this way, the effect size is weighted by the total scores where most of the ASD group resides. A NCDIF effect size of greater than .10 was considered sufficient to indicate a large effect size (Waller, Compas, Hollon, & Beckjord, 2005). This effect size has been shown to be predictive of DIF that is true in the population in real and simulated data (Raju, 1999).

*Criteria for classifying an item as needing deletion*. Items were removed that had a significant chi-square, a significant group difference coefficient, and NCDIF greater than .10. Simulations have shown that combining a statistically significant chi-square and a large effect size reduced Type I errors using the logistic regression procedure (Jodoin & Gierl, 2001).

# Results

## Ability Estimate

After the first and second passes, 42 items were removed from the total score. The third pass used the 352 remaining items to calculate the total score.

## Items Meeting Both Criteria for Item Deletion

The logistic regression identified 136 items that were statistically significantly different between groups, as indicated by a chi-square value greater than 3.84. Of these 136 items, 25 also had an NCDIF score greater than .10, indicating a large effect size. Table 2 contains a list of all items meeting these criteria. Uniform DIF is a statistically significant group parameter without a statistically significant Group × Ability parameter ($\alpha = .05$). Nonuniform DIF is a statistically significant Group × Ability parameter.

## Identification of Words That Were Easiest for One Group Over the Other

*Uniform DIF*. The following items showed uniform DIF favoring typically developing children controlling for the size of their receptive vocabulary: *Cheerio*, *bottle,* the child's own name, *phone, kitty, baby, peek-a-boo, banana, puppy, quack, stroller, woof, keys, hello,* and *moo*. Toddlers with ASD were more likely to understand the following words controlling for the size of their receptive vocabulary: *outside, bed, bubble, stop, more,* and *bite*.

*Nonuniform DIF*. Typically developing infants were more likely than toddlers with ASD who had comparable total scores to be reported as understanding the following words if their total score was less than the numbers in the parentheses: *daddy* (130), *mommy* (110), and *dog* (140). Toddlers with ASD were more likely to be reported as understanding *shoe* if their total score was less than 90. In all cases, the two groups had an equal probability of parents reporting that the children understood the word if the total score was above the value in parentheses.

# Discussion

Twenty-five items on the Receptive Vocabulary scale were identified with DIF defined as statistical significance

**Table 2.** A list of receptive items that met all criteria for deletion, in descending order of effect size.

| Item | Group favored | $\chi^2(2, N = 481)$ | $p$ | Group parameter 95% confidence interval (CI) | Group × Total Score Parameter 95% CI | NCDIF |
|---|---|---|---|---|---|---|
| Daddy | TD | 15.09 | .000 | 1.51, 4.22 | −0.03, −0.01 | .491 |
| Cheerio | TD | 37.04 | .000 | 1.60, 3.36 | 0.00, 0.02 | .394 |
| Mommy | TD | 5.48 | .019 | 0.28, 3.12 | −0.03, −0.01 | .358 |
| Bottle | TD | 28.41 | .000 | 0.94, 2.50 | −0.01, 0.00 | .316 |
| Dog | TD | 8.34 | .004 | 0.74, 2.98 | −0.03, +0.00 | .287 |
| Own name | TD | 5.32 | .021 | 0.15, 1.97 | −0.02, 0.00 | .259 |
| Phone | TD | 15.68 | .000 | 0.38, 2.50 | −0.02, 0.01 | .178 |
| Outside | ASD | 5.31 | .021 | −1.80, −0.03 | 0.00, 0.01 | .175 |
| Bed | ASD | 11.59 | .001 | −2.23, −0.31 | 0.00, 0.01 | .168 |
| Kitty | TD | 14.17 | .000 | 0.26, 1.98 | 0.00, 0.01 | .164 |
| Baby | TD | 14.89 | .000 | 0.37, 2.47 | −0.01, 0.00 | .163 |
| Peek-a-boo | TD | 8.65 | .003 | 0.01, 1.38 | 0.00, 0.01 | .157 |
| Banana | TD | 31.07 | .000 | 0.12, 2.20 | −0.01, 0.00 | .156 |
| Bubble | ASD | 31.07 | .000 | −1.91, −0.10 | −0.01, 0.00 | .154 |
| Stop | ASD | 9.40 | .002 | −2.02, −0.18 | 0.00, 0.01 | .138 |
| Puppy | TD | 14.79 | .000 | 0.54, 2.58 | −0.01, 0.00 | .130 |
| More | ASD | 3.98 | .046 | −1.57, −0.03 | 0.00, 0.01 | .127 |
| Shoe | ASD | 4.50 | .034 | 0.52, 3.38 | −0.10, −0.04 | .124 |
| Quack | TD | 10.09 | .001 | 0.12, 2.00 | −0.01, 0.00 | .120 |
| Stroller | TD | 20.34 | .000 | 0.14, 1.90 | 0.00, 0.01 | .118 |
| Woof | TD | 28.59 | .000 | 0.08, 2.04 | 0.00, 0.01 | .117 |
| Bite | ASD | 4.15 | .042 | −1.70, −0.05 | 0.00, 0.01 | .110 |
| Keys | TD | 13.09 | .000 | 0.16, 2.24 | −0.01, 0.00 | .110 |
| Hello | TD | 17.20 | .000 | 0.05, 1.89 | 0.00, 0.01 | .108 |
| Moo | TD | 4.57 | .030 | 0.04, 2.04 | −0.01, 0.01 | .100 |

*Note.* TD = typically developing; NCDIF = noncompensatory differential item function.

and a large effect size. Items with DIF should be deleted from the MCDI-I to create a set of items that represents the mental lexicon of typically developing infants and toddlers with ASD.

The use of the effect size criteria of NCDIF should increase the probability that the items identified with DIF cause differential test functioning. NCDIF quantifies the impact of an item identified with DIF on test functioning by weighting the difference between the groups by the number of participants of the focal group (i.e., toddlers with ASD) at each total score interval. In this way, NCDIF identifies items that show the largest differences in the ability range (total score range) where most of the children with ASD scored. Therefore, although the number of items identified for deletion is small relative to the total number of items on the test, these items should have a large impact on the validity and unidimensionality of the total score.

In this article, we have demonstrated a method that can be used to identify items that may function differently depending on the age of children who are measured using those items. Chronological age differences often exist when children with disabilities are matched to typically developing children on the number of words that they understand. We would expect some difference in the content of the mental lexicon between children of different ages due to different experiences (e.g., time spent at school vs. home). When a difference in the content of the mental lexicon between groups affects item functioning, we can delete the poorly functioning items to remove error due to chronological age from the total score.

## Other Applications of DIF Analysis in Clinical Populations

The presence of DIF is not unique to comparisons between typically developing infants and toddlers with ASD. Measurement of many clinical groups assumes that the instrument validity in typically developing groups holds true for clinical populations. As was demonstrated in this study, characteristics of clinical populations may differ from those of typical populations in important ways that affect item functioning and validity. Some recent studies that have tested for and found items with DIF include an analysis of a health-related quality of life questionnaire that compared children with attention-deficit/hyperactivity disorder to children with attention-deficit/hyperactivity disorder with comorbid conditions (Klassen, Miller, & Fine, 2004) and an analysis of a depression inventory

that compared women diagnosed with depression to women diagnosed with depression and breast cancer (Waller et al., 2005). DIF analyses identify items that can be removed from an instrument to improve validity across qualitatively different groups.

## Limitations and Future Research

A limitation to this study design is that it did not allow direct investigation of the characteristics of children with ASD that may lead to differences in their mental lexicon and parents' ability to estimate that lexicon. Potential methods for expanding our understanding of these characteristics in children with ASD include repeating the analysis with other clinical groups, using expert judges to confirm the relevance of items with DIF, and using classical test theory to test the validity of the revised instrument.

Repeating the analysis in non-ASD clinical groups that share characteristics believed to cause DIF with toddlers with ASD would allow us to replicate DIF for certain characteristics across clinical groups. For example, a group of children with language delay as their only diagnosis may show bias on items related to chronological age. Examining words identified with DIF when comparing typically developing children to diagnostic groups that share some characteristics with children with ASD (e.g., language delay) but do not share other characteristics (e.g., social communication deficits) could help determine which characteristics affect which items.

Another method would be to train expert judges to use decision rules based on characteristics of group membership expected to produce DIF, such as orienting, social communication, restricted object interest, and chronological age. These decision rules would be used by judges to classify items showing empirical DIF into two groups: (a) those that are predicted by the decision rules and (b) those that are not. If decision rules could be used to reliably identify most items with DIF, they would be supported as characteristics causing DIF. Classical test theory could be used to compare the validity of the original MCDI-I to the version with the items identified with DIF deleted. The analysis strategy would be a correlational design comparing the construct validity of the two versions if the MCDI-I using a nomological network. The comparison would be between the size and direction of the correlations between scores on each of the two versions of the Receptive Vocabulary scale and other constructs that should be related to the size of the mental lexicon (Cronbach & Meehl, 1955).

## Summary

This study identified items that weaken the validity of the MCDI-I as a measure of the size of the mental lexicon. Deletion of these items from the test should increase the degree to which we are measuring the size of the mental lexicon with the total score from the MCDI-I. The analysis procedure, logistic regression, is available through most statistical software packages (e.g., SPSS). In this article, we also suggested four possible characteristics that could explain why the two groups (i.e., typically developing children and children with ASD) have differences in the content of their receptive vocabulary when total receptive score is controlled. Potential methods for expanding our understanding of the mental lexicon in children with ASD include repeating the analysis with other clinical groups, using expert judges to confirm the relevance of items with DIF, and using classical test theory to test the validity of the revised instrument.


## Acknowledgments

Financial support for this work was provided by National Institutes of Health National Research Scientist Award NICHD T32HD07226, Behavioral Research Training in Developmental Disability, awarded to Tedra Walden, Vanderbilt University. We thank Donald Compton, Robin McWilliam, James Stieger, and Niels Waller for their assistance in the development and implementation of this study.


## References


**Ackerman, T. A.** (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

**Adamson, L. B., Bakeman, R., & Deckner, D. F.** (2004). The development of symbol-infused joint engagement. *Child Development, 75,* 1171–1187.

**Aldred, C., Green, J., & Adams, C.** (2004). A new social communication intervention for children with autism: Pilot randomized controlled treatment study suggesting effectiveness. *Journal of Child Psychology and Psychiatry, 45,* 1420–1430.

**American Psychiatric Association.** (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

**Bruckner, C., & Yoder, P.** (2007). Restricted object use in young children with autism: Definition and construct validity. *Autism, 11,* 161–171.

**Charman, T.** (2004). Matching preschool children with ASD spectrum disorders and comparison children for language ability: Methodological challenges. *Journal of Autism and Developmental Disorders, 34,* 59–64.

**Coonrod, E. E., & Stone, W. L.** (2004). Early concerns of parents of children with autistic and nonautistic disorders. *Infants and Young Children, 17,* 258–268.

**Courchesne, E., Townsend, J., Akshoomoff, N. A., Saiotah, O., Yeung-Courchesne, R., Lincoln, A. J., et al.** (1994). Impairment in shifting in autistic and cerebellar patients. *Behavioral Neuroscience, 108,* 848–865.

**Cronbach, L. J., & Meehl, P. E.** (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

**Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., & Liaw, J.** (2004). Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Developmental Psychology, 40,* 271–283.

**Donoghue, J. R., Holland, P. W., & Thayer, D. T.** (1993). A Monte Carlo study of factors that affect the Mantel–Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item function* (pp. 137–166). Hillsdale, NJ: Erlbaum.

**Drew, A., Baird, G., Baron-Cohen, S., Cox, A., Slonims, V., Wheelwright, S., et al.** (2002). A pilot randomized control trial of a parent training intervention for pre-school children with autism: Preliminary findings and methodological challenges. *European Child and Adolescent Psychiatry, 11,* 266–272.

**Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J.** (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development, 59,* 1–173.

**Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., et al.** (1991). *Technical manual for the MacArthur Communicative Development Inventories*. San Diego, CA: San Diego State University Press.

**Hoff, E.** (2001). *Language development*. Wadsworth, CA: Thompson Learning.

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

**Kanner, L.** (1943). Autistic disturbances of affective contact. *Nervous Child, 2,* 217–250.

**Klassen, A. F., Miller, A., & Fine, S.** (2004). Health related quality of life in children and adolescents who have a diagnosis of attention deficit/hyperactivity disorder. *Pediatrics, 114,* 541–547.

**Leekam, S. R., Libby, S. J., Wing, L., Gould, J., & Taylor, C.** (2002). The diagnostic interview for social and communication disorders: Algorithms for ICD-10 childhood autism and Wing and Gould autistic spectrum disorder. *Journal of Child Psychology and Psychiatry, 43,* 327–342.

**Longford, N. T., Holland, P. W., & Thayer, D. T.** (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item function* (pp. 137–166). Hillsdale, NJ: Erlbaum.

**Lord, F. M., & Novick, M. R.** (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

**Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., et al.** (2000). The Autism Diagnostic Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30,* 205–223.

**Mundy, P., Sigman, M., Ungerer, J., & Sherman, T.** (1986). Defining the social deficits of autism: The contribution of non-verbal communication measures. *Journal of Child Psychology and Psychiatry, 27,* 657–669.

**Raju, N. S.** (1999). *DFIT4P: A computer program for analyzing differential item and test functioning*. Unpublished computer software, Illinois Institute of Technology.

**Raju, N. S., van der Linden, W. J., & Fleer, P. F.** (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353–368.

**Sigman, M., & Ruskin, E.** (1999). Continuity and change in the social competence of children with autism, Down syndrome, and developmental delays. *Monographs of the Society for Research in Child Development, 64,* 1–114.

**Stone, W. L., Ousley, O. Y., Yoder, P. J., Hogan, K. L., & Hepburn, S. L.** (1997). Nonverbal communication in two- and three-year-old children with ASD. *Journal of Autism and Developmental Disabilities, 27,* 677–696.

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

**Tabachnick, B. G., & Fidell, L. S.** (2001). *Using multivariate statistics*. Boston: Allyn & Bacon.

**Tomasello, M., & Mervis, C. B.** (1994). The instrument is great, but measuring comprehension is still a problem. *Monographs of the Society for Research in Child Development, 59,* 174–179.

**Toth, K., Munson, J., Meltzoff, A. N., & Dawson, G.** (2006). Early predictors of communication development in young children with autism spectrum disorder: Joint attention, imitation, and toy play. *Journal of Autism and Developmental Disorders, 36,* 993–1005.

**Turner, L., & Stone, W. L.** (2005, April). *Social and nonsocial orienting in 2- and 3-year-old children with autism*. Poster presented at the biennial meeting of the Society for Research in Child Development, Atlanta, GA.

**U.S. Department of Health and Human Services.** (2004). *Congressional Appropriations Committee report on the state of autism research*. Retrieved April 14, 2006, from http://www.nimh.nih.gov/research-funding/scientific-meetings/recurring-meetings/iacc/congressional-appropriations-committee-report.pdf.

**Waller, N. G.** (1998). EZDIF: A program for the analysis of uniform and nonuniform differential item functioning. *Applied Psychological Measurement, 22,* 391.

**Waller, N. G., Compas, B. E., Hollon, S. D., & Beckjord, E.** (2005). Measurement of depressive symptoms in women with breast cancer and women with clinical depression: A differential item functioning analysis. *Journal of Clinical Psychology in Medical Settings, 12,* 127–141.

**Wetherby, A. M., Woods, J., Allen, L., Cleary, J., Dickinson, H., & Lord, C.** (2004). Early indicators of ASD spectrum disorders in the second year of life. *Journal of Autism and Developmental Disorders, 34,* 473–493.

Received September 6, 2006

Revision received January 12, 2007

Accepted March 3, 2007

DOI: 10.1044/1092-4388(2007/110)

Contact author: Cornelia Bruckner, who is now with the Napa County Office of Education, 311 Professional Center Drive, Ronnent Park, CA 94928.
E-mail: cornelia.bruckner@gmail.com.

**Construct Validity of the MCDI-I Receptive Vocabulary Scale Can Be Improved: Differential Item Functioning Between Toddlers With Autism Spectrum Disorders and Typically Developing Infants**

Cornelia Bruckner, Paul Yoder, Wendy Stone, and Megan Saylor
*J Speech Lang Hear Res 2007;50;1631-1638*
DOI: 10.1044/1092-4388(2007/110)