

University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

5-2008

Improving screening for externalizing behavior problems in very young children : applications of item response theory to evaluate instruments in pediatric primary care.

Christina Ruth Studts 1971-
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

Recommended Citation

Studts, Christina Ruth 1971-, "Improving screening for externalizing behavior problems in very young children : applications of item response theory to evaluate instruments in pediatric primary care." (2008). *Electronic Theses and Dissertations*. Paper 1398.
<https://doi.org/10.18297/etd/1398>

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

**IMPROVING SCREENING FOR EXTERNALIZING BEHAVIOR PROBLEMS
IN VERY YOUNG CHILDREN: APPLICATIONS OF ITEM RESPONSE
THEORY TO EVALUATE INSTRUMENTS IN PEDIATRIC PRIMARY CARE**

by

Christina Ruth Studts
B.A., University of Notre Dame, 1993
M.S.W., University of Kentucky, 1997

A Dissertation
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Kent School of Social Work
University of Louisville
Louisville, Kentucky

May 2008

Copyright 2008 by Christina R. Studts

All rights reserved

**IMPROVING SCREENING FOR EXTERNALIZING BEHAVIOR PROBLEMS
IN VERY YOUNG CHILDREN: APPLICATIONS OF ITEM RESPONSE
THEORY TO EVALUATE INSTRUMENTS IN PEDIATRIC PRIMARY CARE**

by

Christina Ruth Studts
B.A., University of Notre Dame, 1993
M.S.W., University of Kentucky, 1997

A Dissertation Approved on

April 1, 2008

By the following Dissertation Committee:

Michael A. van Zyl, Ph.D., chair

Gerard M. Barber, Ph.D., M.P.H.

Andrew J. Frey, Ph.D.

V. Faye Jones, M.D., Ph.D.

DEDICATION

To Jamie and Shannon Kathleen...

*...mo mhíle grá,
tá mo chroí istigh ionat.*

ACKNOWLEDGEMENTS

Many people have influenced my work on this project, and I would like to offer each of them my sincere appreciation. First, I would like to recognize Riaan van Zyl, my dissertation chair and source of incredible support and inspiration throughout my graduate career at the University of Louisville. Riaan taught the first doctoral class I took at the Kent School, and it is no exaggeration when I say that I knew I wanted to convince him to be my mentor after the very first evening of that class. I have been so fortunate to benefit from his wisdom, guidance, and encouragement. He has truly prepared me to pursue excellence in this project and in my future endeavors.

I would also like to thank each of my dissertation committee members—Rod Barber, Jim Clark, Andy Frey, and Faye Jones—who have helped me see problems from many perspectives. They have consistently offered their support and expertise to enrich my understanding of this topic and of research in a broader sense. I could not have asked for more in a committee.

My family has also been a constant source of encouragement and strength. My parents, Bob and Kathy Clark, instilled in me a love of learning, a passion to achieve, and a focus on concern for others. My sister and fellow lifelong student, Jeannie Clark-Mabey, has shared in the joys of returning to school in our 30s. My husband Jamie and daughter Shannon have endured my constant attachment to the computer and endless monologues about data analyses, all the while offering unconditional love and support.

Special thanks to Jamie for inspiring me as a dedicated and accomplished scientist-practitioner, and to Shannon for being such a good napper. My family has been with me every step of the way—and for Shannon, I mean that literally.

My gratitude also goes to Ruth Huber and Norma Melton—boss lady and Queen of the doctoral program, respectively—who have kept me moving forward, reminded me of deadlines, and generously provided assistance and nurturing along the way. Ruth and Norma provide a remarkable foundation for the Kent School doctoral students with their unwavering support.

Dana Sullivan and Janet Carpenter were my “unofficial” committee members, offering encouragement, advice, and cheerleading from qualifying exams through the defense. I thank them for their humor, wisdom, and listening skills, and for counseling sessions and camaraderie over good meals, by phone, and even in the gyms of strange cities.

Thanks also to Barbara Donadio and the physicians, staff, and patients of Duke Children’s Primary Care in Durham, North Carolina. My experiences working as a pediatric clinical social worker with such an energetic, knowledgeable, and committed team of health care professionals were the catalyst for my interest in the issues addressed in this dissertation.

Several people enabled the work herein to happen through their collaboration and cooperation. Jo Ann Wood first contacted potential clinics on my behalf. Faye Jones, Josh Honaker, Melissa Hancock, and Judith Theriot generously allowed me to become a fixture in their clinic waiting rooms. The physicians and staff of UCHS, UCHS-South, Children & Youth Project, and Oldham County Pediatrics were consistently supportive

and helpful. Judith Friedrich, Demeka Campbell, and Cynthia Bowman-Stroud assisted in data collection efforts, offering their time, organizational skills, and personable dispositions to help me attain a daunting sample size. Finally, and most importantly, my thanks goes to the 900 parents of preschool-aged children who were willing to take the time to answer my questions.

ABSTRACT

IMPROVING SCREENING FOR EXTERNALIZING BEHAVIOR PROBLEMS IN VERY YOUNG CHILDREN: APPLICATIONS OF ITEM RESPONSE THEORY TO EVALUATE INSTRUMENTS IN PEDIATRIC PRIMARY CARE

Christina R. Studts

May 10, 2008

Externalizing behavior problems in very young children are associated with an array of negative and costly long-term outcomes. Pediatric primary care is a promising venue for implementing screening practices to improve early identification of this social and public health problem. In this setting, screening requires a brief, easily scored instrument which can detect sub-clinical to clinical levels of the latent construct within the context of early childhood development. Further, items used should perform consistently with children of all sociodemographic backgrounds. This study applied item response theory analyses to investigate the precision, utility, and differential item functioning (DIF) of items measuring externalizing behavior problems in two caregiver-report questionnaires: the PSC-17 (Gardner et al., 1999) and the BPI (Peterson & Zill, 1986; Zill, 1990). Caregivers ($N = 900$) of children ages 3 to 5 responded to both instruments and a sociodemographic questionnaire in the waiting rooms of four pediatric primary care clinics. Sociodemographic characteristics of the children were diverse: 47% were female, 50% were of minority race, and 43% were of low socioeconomic status

(SES). Eighteen items comprising the instruments' combined externalizing subscales were evaluated for (a) levels of externalizing behavior problems best measured, and (b) DIF exhibited by child sex, race, and SES. Samejima's (1969) graded response model was fit to the data, and two methods of DIF-detection were employed. Estimation of item parameters allowed consideration of the levels of externalizing behavior problems at which each item was most informative. Five items were found to measure only low to average levels of externalizing problems in the target population, while the remaining 13 were informative at sub-clinical to clinical levels. Significant DIF was detected in 8 of 18 items by child sex, race, or SES. A set of 4 items was identified which (a) provided the most information at sub-clinical to clinical levels of externalizing behavior problems, and (b) exhibited the least amount of DIF by child sex, race, and SES. These items may constitute a promising tool for screening purposes with preschool-aged children in the primary care setting, potentially improving early identification of very young children with externalizing behavior problems.

TABLE OF CONTENTS

	PAGE
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	vii
LIST OF TABLES	xvi
LIST OF FIGURES	xviii
 CHAPTER	
I. PROBLEM STATEMENT AND STUDY OVERVIEW	1
Screening in Pediatric Primary Care.....	3
The Promise of Item Response Theory.....	6
Purpose and Methodology	8
Clarification of the Scope of the Study.....	9
Problem Definition.....	9
Study Parameters	10
Terminology Note.....	10
Significance of the Study.....	11
II. EXTERNALIZING BEHAVIOR PROBLEMS IN VERY YOUNG CHILDREN.....	13
Problem Definition and History.....	14
Prevalence	17
Causes and Consequences.....	19
Risk Factors	20

Protective Factors.....	22
Consequences.....	23
Approaches to the Problem.....	24
Reactive Approaches	24
Proactive Approaches	25
Barriers to Early Identification and Early Intervention	26
Complexity of Screening Very Young Children	27
Problematic Social Attitudes.....	28
Fragmentation across Systems.....	29
The educational system.....	29
The health care system.....	32
The Potential of Pediatric Primary Care	34
The “Medicalization” of Externalizing Behavior Problems in Children ...	34
Identification in Pediatric Primary Care: Problems and Promise	36
Standardized Screening Tools for Externalizing Behavior Problems.....	40
Problems with Identified Instruments.....	43
Screening of very young children	44
Disparities in identification.....	46
Shortcomings of Classical Test Theory	48
Summary and Research Questions.....	49
Research Questions.....	54
III. ITEM RESPONSE THEORY: APPLICABILITY TO MEASUREMENT OF EXTERNALIZING BEHAVIOR PROBLEMS IN VERY YOUNG CHILDREN	56

Limitations of Classical Test Theory	58
Development of Item Response Theory	60
IRT: Modern Model-Based Measurement	61
Models for Dichotomous Items	62
Models for Polytomous Items	65
Assumptions and Limitations of IRT	67
Unidimensionality	68
Local Independence	68
Trace Line Functions	69
Practical Limitations	70
Theoretical Advantages of IRT	71
The Graded Response Model	74
Information and Precision	78
Differential Item Functioning	83
Summary and Hypotheses	86
Study Hypotheses	89
IV. METHOD	91
Participants	91
Recruitment Sites	91
UCHS and UCHS-S	92
C&Y	92
OCP	92
Inclusion and Exclusion Criteria	93

Procedure	94
Data Collection	94
Gift Card Drawing	95
Measures	96
Pediatric Symptom Checklist-17 (PSC-17)	96
Behavior Problems Index (BPI).....	97
Sociodemographic Questionnaire	99
Caregiver characteristics.....	99
Child characteristics.....	99
Other relevant factors.....	100
Data Analysis	101
Descriptive Analyses	101
CTT-Based Analyses	101
IRT-Based Analyses	102
Evaluation of IRT model assumptions.....	103
Fitting the IRT model.....	104
Detection of DIF	107
Power and Sample Size Considerations.....	110
CTT-based analyses.....	110
IRT-based analyses	110
Integration of Findings.....	111
V. RESULTS	112
Sample Characteristics.....	112

Caregivers	112
Children.....	114
Classical Test Theory Psychometric Analyses	117
Distributional Properties	118
Reliability.....	118
Validity	119
Concurrent validity	120
Known groups validity.....	120
Group Differences by Child Sex, Race, and Socioeconomic Status.....	122
Differences by sex.....	123
Differences by race	123
Differences by SES	123
Psychometric properties and sex, race, and SES	126
Item Response Theory Analyses.....	126
Evaluation of IRT Model Assumptions	127
Unidimensionality.....	127
Local independence	128
Specific trace line functions.....	128
Research Question 1: Precision and Utility of Measurement.....	131
Model fit.....	131
Item parameter estimates	133
Test information.....	140
Item information	140

Research Question 2: Item-level Measurement Bias	144
IRT-LR.....	144
OLR.....	154
Comparisons of DIF findings.....	158
Extent of DIF effects.....	163
VI. DISCUSSION.....	171
Scale Performance in Context: Classical Test Theory Analyses	172
Research Question 1: Precision and Utility of Measurement	174
Precision of Measurement along the Continuum.....	175
Selecting Among Equally Informative Items	176
Research Question 1 Summary.....	177
Research Question 2: Item-level Measurement Bias	178
Detection of Significant DIF.....	179
Magnitude and Direction of DIF Effects	180
DIF by child sex.....	181
DIF by child race	181
DIF by child SES	182
Unadjusted versus DIF-adjusted IRT scores.....	183
Research Question 2 Summary.....	184
Item Content: Relevance to Screening of Very Young Children	185
Item Content and Item Information	186
Easy items	186
Difficult items	187

Summary: Item content along the continuum.....	188
Item Content and Differential Item Functioning	188
Item content and child sex	189
Item content and child race.....	190
Item content and child SES.....	191
Item content and DIF-free items	192
Summary: Item content and DIF.....	193
Integration of Results: Identification of “Best” Items	194
Implications.....	198
Limitations	203
Directions for Future Research	207
Summary and Conclusions	210
REFERENCES	213
APPENDIX A: Diagnostic Criteria.....	234
APPENDIX B: Preamble Consent	237
APPENDIX C: Eligibility Checklist and Script to Invite Participation	240
APPENDIX D: PSC-17 (Gardner et al., 1999).....	241
APPENDIX E: Scoring Instructions for PSC-17 (Gardner et al., 1999).....	243
APPENDIX F: BPI (Peterson & Zill, 1986; Zill, 1990).....	244
APPENDIX G: Scoring Instructions for BPI (Peterson & Zill, 1986; Zill, 1990)	247
APPENDIX H: Sociodemographic Questionnaire	249
CURRICULUM VITAE.....	252

LIST OF TABLES

TABLE	PAGE
1. Caregiver Characteristics ($N = 900$)	113
2. Child Characteristics ($N = 900$)	115
3. Caregiver-Reported Child Behavioral Health History ($N = 900$)	117
4. Descriptive Statistics for PSC-17, BPI, and Selected Subscales	119
5. Internal Consistency of PSC-17, BPI, and Selected Subscales	120
6. Known Groups Validity: Parent Belief that Child has Behavior Problems	121
7. Known Groups Validity: Child History of Contact with Mental Health Professional (MHP)	122
8. Differences in Mean Scores by Child Sex	124
9. Differences in Mean Scores by Child Race	124
10. Differences in Mean Scores by Child Socioeconomic Status	125
11. Summary of Exploratory Factor Analysis Results for Combined Externalizing Subscale ($N = 861$).....	129
12. Goodness of Fit: Frequencies and Means of Chi Square to Degrees of Freedom Ratios.....	133
13. Item Descriptives and Graded Response Model Parameter Estimates for Total Sample ($N = 900$).....	135
14. Maximum Item Information Estimates and Locations	142
15. Differential Item Functioning of Combined Externalizing Subscale Items by Child Sex	147
16. Differential Item Functioning of Combined	

	Externalizing Subscale Items by Child Race	149
17.	Differential Item Functioning of Combined Externalizing Subscale Items by Child Socioeconomic Status.....	151
18.	Comparison of Results of DIF Detection by Two Methods	159
19.	Graded Response Model Item Parameter Estimates Adjusted for Items Displaying DIF by Child Sex.....	165
20.	Graded Response Model Item Parameter Estimates Adjusted for Items Displaying DIF by Child Race.....	165
21.	Graded Response Model Item Parameter Estimates Adjusted for Items Displaying DIF by Child Socioeconomic Status	166
22.	Differences in Unadjusted and DIF-Adjusted Theta Score Estimates within Sociodemographic Groups.....	170

LIST OF FIGURES

FIGURE	PAGE
1. Item characteristic curve (ICC) for a hypothetical item in the two-parameter logistic model (2PL).....	64
2. Three hypothetical item characteristic curves (ICCs) with differing item parameters	62
3. Graded response model option characteristic curves (OCCs) for a hypothetical item with three response options.....	77
4. Item information functions for three hypothetical items	81
5. Test information function for a set of three hypothetical items.....	82
6. Non-parametric trace line plot for item PSC-17 4 (“Refuses to share”).....	130
7. Sample fit plots for the graded response model option characteristic curves (OCCs) of two items	132
8. Plots of graded response model option characteristic curves (OCCs) for all items in the combined externalizing subscale.....	137
9. Test information function plot for all items in the combined externalizing subscale.....	141
10. Relative levels of item information provided in the sub-clinical to clinical range of externalizing behavior problems	143
11. Plots of graded response model option characteristic curves (OCCs) by group for items exhibiting differential item functioning by child sex.....	167
12. Plots of graded response model option characteristic curves (OCCs) by group for items exhibiting differential item functioning by child race.....	168
13. Plots of graded response model option characteristic curves (OCCs) by group for items exhibiting differential item functioning by child SES	169

14. Relative levels of DIF-free item information provided by items
in the sub-clinical to clinical range of externalizing behavior problems196

CHAPTER I

PROBLEM STATEMENT AND STUDY OVERVIEW

Violence, aggression, rule-breaking, defiance, and cruelty: These and other externalizing behavior problems manifest not only in adolescents and older children, but also in very young children. Preschool-aged children who are *early starters* with respect to such behaviors are at high risk of a continuing developmental pathway of antisocial behaviors (Hann & Borek, 2001). An array of serious and costly long-term consequences of negative behavioral patterns in early childhood has been identified, including school failure, substance abuse, adult criminal activity, and higher hospitalization and mortality rates (Moffitt, 1994; Patterson, DeBaryshe, & Ramsey, 1989). In addition, lower health-related quality of life (Sawyer et al., 2002), increased rates of health care utilization (Zuckerman, Moore, & Gleib, 1996), increased rates of suicidality (Shaffer, Fisher, & Dulcan, 1996), and adult diagnoses of Antisocial Personality Disorder (U.S. Department of Health and Human Services [U.S. DHHS], 1999) are known health-related outcomes associated with early externalizing behaviors.

Based on epidemiological studies of older children, conservative estimates of the prevalence in the United States (U.S.) of externalizing behavior problems in children between the ages of 3 and 5 suggest that from 1% to 6% may meet diagnostic criteria for Oppositional Defiant Disorder and Conduct Disorder (American Psychiatric Association [APA], 2000; Shaffer et al., 1996; U.S. DHHS, 1999). However, it is likely that more

than one in five children exhibit sub-threshold psychosocial symptoms (E. J. Costello & Shugart, 1992; U.S. DHHS, 1999), increasing risk for development of later problems. The overwhelming majority of U.S. children exhibiting these problems do not receive specialized services (Kataoka, Zhang, & Wells, 2002).

Primary and secondary prevention efforts, such as early identification and early intervention, have been lauded as essential strategies for alleviating this social and public health problem (Forness, Kavale, MacMillan, Asarnow, & Duncan, 1996; Hoagwood & Johnson, 2003). However, significant barriers to these proactive approaches exist, due in part to attitudes underlying service philosophies of social institutions typically in contact with very young children (Kauffman, 1999; U.S. Department of Education, 2003; U.S. DHHS, 1999; U.S. General Accounting Office [GAO], 2003). Due to fragmented service systems and approaches among these institutions (e.g., the educational system and the health care system), when parents are concerned about their child's behavior, the decision regarding which system to contact for assistance can have major repercussions (Hoagwood & Johnson, 2003).

Unfortunately, what all systems have in common is the tendency to under-identify early signs of externalizing behavior problems (Forness & Knitzer, 1992; Hoagwood & Erwin, 1997; Redden, Forness, Ramey, Ramey, & Brezausek, 2003). One key reason for under-identification is the complexity of screening for behavioral problems within the developmental context of this age group: A behavior deemed pathological for one very young child in a given situation may be developmentally appropriate for another. Further, the influences of varying combinations of biologic, familial, and social-environmental characteristics and histories complicate assessment efforts (Kagan, 1997).

Screening in Pediatric Primary Care

Pediatric primary care is an ideal setting for screening and early identification efforts (Agency for Healthcare Research and Quality [AHRQ], 2002), offering additional resources beyond those offered by the educational system to expand primary and secondary prevention practices. While the significance of psychosocial issues in primary care settings has been recognized, primary care physicians—the de facto mental health service providers (Regier, Goldberg, & Taube, 1978) in the U.S.—have struggled with persistent under-identification of children in need of services (E. J. Costello, 1986; E. J. Costello & Edelbrock, 1985; E. J. Costello et al., 1988; Lavigne et al., 1993). As gatekeepers to specialized behavioral services provided by social workers and other mental health professionals, physicians fill a crucial role in early identification efforts. However, assessment methods favored by most pediatric health providers are typically informal (American Academy of Pediatrics, 2000) and have low sensitivity: Pediatric primary care providers identify only 20% of children with mental health issues identified by psychologists using standardized assessment instruments (E. J. Costello et al., 1988; Lavigne et al., 1993). Importantly, when pediatric primary care providers do refer preschool-aged children with clinically significant behavioral problems for specialized services, the odds that a child accesses such services increase significantly, compared to similar children without physician referrals (Lavigne, Arend, Rosenbaum, Binns, Christoffel, Burns et al., 1998).

To improve rates of identification in pediatric primary care, standardized screening approaches using reliable and valid instruments may be helpful (Halfon, Regalado, McLearn, Kuo, & Wright, 2003; L. G. Hill, Coie, Lochman, & Greenberg,

2004). While many instruments have been developed, most are inappropriate for screening purposes in primary care settings, due to (a) excessive length for administration, scoring, and interpretation; (b) prohibitive costs; and (c) development with non-representative norming samples. In contrast, brief, easily scored, freely available instruments such as the Pediatric Symptom Checklist-17 (PSC-17; Gardner et al., 1999) and the Behavior Problems Index (BPI; Peterson & Zill, 1986; Zill, 1990) may be valuable tools for pediatric primary care. Each of these instruments includes subscales intended to measure externalizing behavior problems.

While the PSC-17 and the BPI have been used in research and clinical settings, concerns have been raised regarding their reliability and validity with very young children, minority children, and children of low socioeconomic status (SES). Though both scales were initially designed for use with children ages 4 and above, psychometric analyses have reported problems with the full-length PSC (Jellinek, Murphy, & Burns, 1986) with children under age 6, and have not attended to differential effects of age with the BPI (Parcel & Menaghan, 1988; Zill, 1985, 1990). No published studies have investigated the potential utility of these readily available instruments with children under age 4, though targeting children in the preschool age range for screening is imperative for prevention efforts. In addition, some studies have suggested disparities in screening results derived from these instruments by sex (Jellinek et al., 1999; Parcel & Menaghan, 1988), race (Jutte, Burgos, Mendoza, Ford, & Huffman, 2003; Simonian & Tarnowski, 2001; Simonian, Tarnowski, Stancin, Friman, & Atkins, 1991; Spencer, Fitch, Grogan-Kaylor, & McBeath, 2005), and SES (Jellinek, Little, Murphy, & Pagano, 1995; Jellinek et al., 1999). While variability in symptom expression and perception across population

subgroups is known to exist (U.S. DHHS, 2001), bias in screening instruments can result in both over-identification and under-identification of children in certain groups, stymieing equitable and appropriately targeted primary and secondary prevention efforts (Spencer et al., 2005) and perpetuating social injustices and health disparities.

All published psychometric evaluations of the PSC-17 and the BPI have relied upon traditional analyses based on Classical Test Theory (CTT). Unfortunately, CTT-based analyses are limited in their capacity to assess measurement performance independent of the particular samples included in investigations (Nunnally & Bernstein, 1994). Thus, reliability and validity estimates reported for the PSC-17 and the BPI are dependent on the characteristics of the specific samples used, and application of these instruments with children not represented by these samples may result in changes in psychometric properties (Lord & Novick, 1968). Other shortcomings inherent in CTT-based methods of scale development and evaluation include (a) the untenable assumption that the standard error of measurement (SEM) is constant across all levels of the measured construct (Hambleton & Swaminathan, 1985; Nugent, 2005); (b) floor and ceiling effects (Hambleton, Swaminathan, & Rogers, 1991; Ware, 2003); (c) excessive length (Hambleton et al., 1991; Ware, 2003); and (d) the inability to extricate item-level bias from true group differences in levels of the measured construct (Hambleton & Swaminathan, 1985).

These limitations may explain the variability in estimates of reliability and validity of the PSC-17 and BPI when used with groups of children differing by sex, race, and SES (Jellinek et al., 1995; Jellinek et al., 1999; Jutte et al., 2003; Navon, Nelson, Pagano, & Murphy, 2001; Parcel & Menaghan, 1988; Spencer et al., 2005). Since

existing psychometric analyses have relied solely on CTT-based methods, the following important questions remain regarding the quality of measurement provided by these instruments with the population of interest:

1. How precise is the measurement offered by PSC-17 and BPI items at various levels of externalizing behavior problems?
2. What is the range of externalizing behavior problems adequately measured by these scales? In particular, items capable of detecting sub-clinical levels of behavior problems reliably are needed for effective primary and secondary prevention efforts (E. J. Costello & Shugart, 1992).
3. Finally, do items in these scales exhibit biases in performance between groups (e.g., differing by sex, race, and SES) when controlling for level of externalizing behavior problems?

These vital questions of measurement quality cannot be answered using traditional CTT-based methods of scale evaluation. Alternative, advanced methods are available and are described in the next section.

The Promise of Item Response Theory

Item response theory (IRT) is an exciting, modern statistical approach which could improve measurement in both practice and research applications. This measurement theory is distinct from CTT, offering applications and information which are unattainable with traditional psychometric methods. IRT-based methods involve the fitting of joint probability mathematical models, predicting the probability of item endorsement as a function of the level of the underlying construct being measured (Hambleton & Swaminathan, 1985). The core theoretical advantage of IRT is its concept

of *parameter invariance*, enabling “test-free” and “sample-free” measurement (Hambleton & Swaminathan, 1985). Stable parameters describing item characteristics allow measurement properties analogous to the physical measurements of weight and height, in which attributes of the sample or measurement tool used are independent of the invariance of the underlying metric (Lord, 1980). While random samples are not required for either CTT or IRT analyses, the novel data offered by IRT regarding item- and scale-level measurement performance can be generalized from one sample to another, unlike the traditional psychometric indices obtained via CTT methods. Thus, the use of convenience samples for IRT analyses is entirely appropriate and does not limit generalizability.

This model-based approach to measurement allows investigation of several issues impossible to address with traditional CTT-based methods. For example, IRT model-fitting provides a basis for comparing the relative merit of items in terms of the amount of information they provide for measuring specific levels of the underlying construct of interest, such as externalizing behavior problems (Hambleton & Swaminathan, 1985). Similarly, the degree of precision of measurement of an item at various levels of the underlying construct can be determined. In addition, the application of IRT methodology enables the identification of items exhibiting *differential item functioning* (DIF), or item bias, in which responses to an item are affected not only by the level of the underlying construct, but also by extraneous characteristics, such as sex, race, or SES (Teresi, 2001).

The use of IRT-based methods to evaluate externalizing subscales of the PSC-17 and the BPI could greatly enhance understanding of the applicability of these scales to early identification efforts in the primary care setting. Items could be identified which

provide the most information and the most precise measurement of sub-clinical and clinical levels of externalizing behaviors among children ages 3 to 5. By investigating possible DIF exhibited by items in these scales, concerns regarding health disparities and under- and over-identification of minority and low-SES children with current assessment strategies could be addressed. Brief sets of items could be recommended which provide the most informative, most precise, and least biased measurement at desired levels of externalizing behavior problems for the target population.

Purpose and Methodology

The purpose of this study was to evaluate the quality (i.e., precision and utility) of measurement provided by externalizing subscale items in the PSC-17 and BPI with preschool-aged children seen in pediatric primary care practices. In addition, items were investigated for DIF between groups differing by child sex, race, and SES. Results were reviewed in order to identify a set of items most appropriate for use in screening very young children for externalizing behavior problems in diverse pediatric primary care settings.

To achieve these goals, a cross-sectional survey design was employed. Consistent with the requirements of IRT-based analyses, a large sample ($N = 900$) was selected from four pediatric primary care practices serving sociodemographically diverse populations of children. Nonrandom sampling procedures were used, in which a convenience sample of potential participants was recruited in the waiting rooms of the pediatric primary care practices. Due to unique properties of IRT, this strategy did not limit generalizability of results. Primary caregivers of children ages 3 to 5 were invited to participate in the study, which involved completion of a set of three measures: the PSC-17 (Gardner et al., 1999);

the BPI (Peterson & Zill, 1986; Zill, 1990); and a sociodemographic questionnaire developed by the author. Descriptive and CTT-based analyses were conducted to characterize the study sample and traditional psychometric properties of the PSC-17 and BPI, for comparison with previous studies.

The crux of this investigation, however, lay in the IRT-based analyses of item responses. Samejima's (1969) graded response model, an IRT model developed for items with polytomous ordered response options, was fit, and the resulting item parameter estimates were compared. The amount of information and precision provided by each item along the continuum of externalizing behavior problems was assessed, and each item was examined for DIF between groups differing by sex, race, and SES. Using the results of these analyses, items were identified which appeared to (a) measure sub-clinical to clinical levels of externalizing behavior problems in preschool-aged children most precisely, and (b) exhibit the least amount of DIF between groups of interest. The most informative, precise, and unbiased items were proposed as a set suitable for improved measurement of externalizing behavior problems among very young children in the pediatric primary care setting.

Clarification of the Scope of the Study

To clarify the scope of this study, the following definition, parameters, note regarding terminology, and summary of study significance are provided:

Problem Definition

The social and public health problem of interest in this study is that of externalizing behavior problems in very young children. For the purposes of this study, externalizing behavior problems include those characterized by diagnoses of

Oppositional Defiant Disorder and Conduct Disorder (APA, 2000). Sub-clinical behaviors, such as those associated with Disruptive Behavior Disorder Not Otherwise Specified (NOS), are also relevant to this definition (APA, 2000). Externalizing behaviors below clinical thresholds are important to identify for the purposes of primary and secondary prevention (E. J. Costello & Shugart, 1992). However, behaviors typically associated with Attention Deficit Hyperactivity Disorder (ADHD) are excluded from the current definition (e.g., impulsivity, restlessness, difficulty sustaining attention, and so on; APA, 2000). An extensive literature exists regarding ADHD and its causes, consequences, identification, and treatment, all of which is beyond the scope of this study.

Study Parameters

The current study focuses solely on the population of very young children (i.e., ages 3 to 5) followed in pediatric primary care settings. While externalizing behavior problems manifest in children and youth of all ages, the preschool-aged target population is of special interest due to its relevance to primary and secondary prevention efforts. In addition, though not all children are followed by pediatric primary care providers, the focus of this study is on improving screening efforts in this venue; therefore, differences between children who are and are not seen in primary care are not addressed.

Terminology Note

In describing the process, results, and implications of evaluating screening instruments for externalizing behavior problems, certain terminology are employed based upon classical and modern measurement theory. In particular, several ways of referring to the problem of interest are utilized, depending on the context of the discussion. In IRT,

the underlying trait, attribute, or behavior being measured is generally denoted by the Greek letter θ . This notation is used throughout Chapter III, as statistical formulas and equations constitute an important portion of that chapter. In other areas of this text (e.g., Chapters IV and V), *theta* is employed in place of the Greek letter, for ease of reading. In discussions of interpretation and implications rather than in the context of statistical formulas, the terms *latent construct* or simply *externalizing behavior problems* are used. All of these terms— θ , theta, latent construct, and externalizing behavior problems—are interchangeable when used to describe the problem of interest in this study.

Significance of the Study

This study highlights the importance of early identification of very young children with externalizing behavior problems, with a special focus on the pediatric primary care setting. Shortcomings of current methods of early identification are delineated. These include limitations inherent in the pediatric primary care setting, as well as those related to traditional psychometric development of screening instruments. Application of IRT is shown to be a valuable approach to improving measurement of this social and public health problem in the target population of preschool-aged children. Improvements in screening technologies are offered, potentially leading to the reduction of social injustices perpetuated by the use of items biased against particular sociodemographic groups. Findings of this study, while directed primarily at the pediatric primary care setting, may be equally applicable to other settings, including preschools, early childcare, mental health, and the child welfare system.

Implications of the study include several important considerations for the social work field with regard to research, education, and practice. The social work profession is

uniquely positioned to continue research in this vein, including both qualitative and quantitative follow-up studies as well as continued efforts in the application of IRT methods. Social work education should support the development of increased familiarity with both traditional and advanced psychometric methods among students at all levels: Informed use of screening instruments among social work practitioners, as well as continued development of improved screening technologies among social work researchers, are only possible with attention to measurement theory in social work education. Social workers function in increasingly interdisciplinary settings—both in research and practice—and should understand the appropriateness, or lack thereof, of measurement instruments used within their realm of influence. Indeed, the development and use of screening technologies which could enhance early identification and facilitate the elimination of existing disparities is in harmony with the mandates of the National Association of Social Workers Code of Ethics (NASW, 2000).

CHAPTER II

EXTERNALIZING BEHAVIOR PROBLEMS IN VERY YOUNG CHILDREN

Externalizing behavior problems among very young children in the U.S. are a growing social and public health concern. This chapter provides a review of the literature addressing externalizing behavior problems in children between the ages of 3 and 5, offering a context for the proposed investigation of the quality of screening instruments used for early identification of such problems in the primary care setting.

First, the definition and history of externalizing behavior problems in very young children are reviewed. Prevalence estimates and problems with such estimates are described. Next, research on the causes and consequences of this social problem is summarized, with special attention to studies exploring risk factors, protective factors, and long-term consequences of early emergence of externalizing behavior problems. The importance of a proactive approach (i.e., via primary and secondary prevention efforts) is highlighted as it relates to early identification of externalizing behavior problems in very young children.

Barriers to prevention efforts are also described, including complexities in the assessment of very young children, problematic social attitudes, and fragmentation of services and approaches adopted by involved social systems and institutions.

Identification of the primary care setting as an ideal venue for efforts toward early identification of externalizing behavior problems among very young children is

supported. Research exploring problems with screening for mental health issues in primary care is presented, and the availability and incumbent shortcomings of standardized instruments used in screening efforts are reviewed. Disparities in rates of identification are emphasized, particularly those associated with child sex, race, and SES.

Finally, specific research questions regarding the utility and performance of two screening instruments are posed. These questions lead directly to a discussion in Chapter III of a promising modern measurement approach that could improve screening for externalizing behavior problems among very young children in the primary care setting.

Problem Definition and History

According to the National Institute of Mental Health (NIMH; Hann & Borek, 2001), the term *externalizing behavior problems* refers to a range of conduct problems and rule-breaking behaviors which are more frequent and severe than the typical range of expected behaviors in children of the same developmental stage. Other terms often used to describe this problem are *antisocial*, *challenging*, and *disruptive* behaviors in children. Behaviors of concern include physical and verbal aggression, defiance, lying, stealing, truancy, delinquency, physical cruelty, and criminal acts. In addition to the negative impact these behaviors have on children and those in their social environments, when they (a) are present in persistent patterns, (b) are observed across settings (e.g., at home and at daycare or preschool), and (c) lead to clinically significant impairment in functioning, they can fulfill the requirements for one of two mental health disorder diagnoses: Oppositional Defiant Disorder (ODD) or Conduct Disorder (CD; see Appendix A; APA, 2000). Sub-clinical externalizing behaviors which do not meet the diagnostic criteria for ODD or CD may be categorized as Disruptive Behavior Disorder

NOS (see Appendix A; APA, 2000). Externalizing behavior problems, if unchecked, appear to be relatively stable in children: Longitudinal studies have shown a strong correlation between aggressive behaviors and attributes in 3 year old children and measurements of the same constructs in the same children 8 and 10 years later (Lavigne, Arend, Rosenbaum, Binns, Christoffel, & Gibbons, 1998; Raine, Reynolds, Venables, Mednick, & Farrington, 1998).

Recognition of emotional and behavioral disorders in children is a relatively recent phenomenon. The concept of mental illness in children did not arise until the late 19th century, and child mental illness was not differentiated from adult mental illness until the early 20th century (National Advisory Mental Health Council Workgroup [NAMHCW], 2001). The first child guidance clinic in the United States was established in 1909, and the first English-language text on child psychiatry was published in 1935 (Sanua, 1990; Snodgrass, 1984). Not until the 1970s, during a World Health Organization meeting on classification of mental health disorders for the International Classification of Diseases (ICD), was the idea of separately coding clinical diagnoses for child psychiatry first introduced (NAMHCW, 2001). Several years later, the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III) finally assigned child and adolescent disorders a separate and distinct section (APA, 1980). Today, the state of the research on child mental health issues still reflects relatively early stages of understanding.

In response to the recognition and categorization of childhood mental health disorders as distinct from those ascribed to adults, researchers have explored the validity of childhood mental health issues, including externalizing behavior problems. Only two

decades ago, Achenbach and Edelbrock (1981) conducted one of the first factor analysis studies on child mental health issues. In distinguishing between so-called *externalizing* and *internalizing* (i.e., anxiety and mood-related) problems, they provided the basis for development of many broadband scales designed to measure and distinguish between these types of mental health issues. In assessing the utility and appropriateness of DSM diagnoses for preschool-aged children, Keenan and colleagues (1997) and Keenan and Wakschlag (2000) conducted a series of studies with very young children in clinic settings. They concluded that the problem behaviors exhibited by clinic-referred children were “more than the terrible twos” (p. 33), suggesting that DSM categorization of problems was appropriate even for very young children. However, while externalizing behavior problems have been identified and classified as a group, caution has been urged in assigning heterogeneous children to homogeneous categories, as similar-appearing symptoms may actually obfuscate important differences among those assigned the same diagnosis (Kagan, Snidman, McManis, Woodward, & Hardway, 2002).

In a landmark 1999 report on mental illness, the U.S. Surgeon General defined mentally healthy children as characterized by a positive quality of life, good functioning across settings, and freedom from disabling symptoms of psychopathology (U.S. DHHS, 1999). Most children between the ages of 3 and 5 engage in rule-breaking and defiance as typical developmental phenomena, but learn to replace noncompliance and aggression with prosocial strategies as they develop cognitive, language, and social skills. Children in this age range who do not learn or use more prosocial strategies but instead continue and increase their externalizing behaviors are sometimes referred to as *early starters* (Hann & Borek, 2001). These early starters may be more likely than children without

early onset behavior problems to exhibit lifecourse-persistent antisocial behaviors, which continue through middle and late childhood, adolescence, and adulthood (Hann & Borek, 2001). Compared to children who develop externalizing behavior problems at later ages, early starters have been shown to exhibit more intransigent problems in later childhood and adolescence, with increased severity of a developmental pathway of antisocial behaviors (Ge, Donnellan, & Wenk, 2003; Moffitt, 1994; Patterson et al., 1989). Studies have suggested that very young children with externalizing behavior problems are at high risk for escalating and intensified behaviors including bullying, physical aggression, cruelty to animals, vandalism, and violent criminal acts (Hann & Borek, 2001). Such findings, combined with heightening concerns about school violence over the past decade, have led to increased public awareness of this social and public health problem, as well as to a burgeoning research agenda.

Prevalence

To date, no epidemiological studies have been completed which focus on the mental health issues of very young children in the United States. The closest current prevalence estimates hail from the Methodology for Epidemiology of Mental Disorders in Children and Adolescents (MECA) study, which aimed to describe the prevalence of all mental disorders in children between the ages of 9 and 17 (Shaffer et al., 1996). It is thought that prevalence of mental disorders is similar for children below the age of 9 (U.S. DHHS, 1999).

Annual prevalence of all mental disorders in children ages 9 through 17 is estimated at 20.9% of the nearly 36 million children in this age range, a proportion similar to the prevalence of adult mental disorders found in the Epidemiological

Catchment Area studies in the 1980s (U.S. DHHS, 1999). For disruptive behavior disorders alone, the prevalence drops to 10.3% (Shaffer et al., 1996); however, the disruptive behavior disorders category in MECA included ADHD, which is excluded from the present definition of externalizing behavior problems. When assessment criteria for all mental disorders are restricted to symptoms meeting DSM-III criteria plus significant functional impairment—defined as a Child Global Assessment Scale (CGAS) rating of below 60 (Shaffer et al., 1983)—the overall annual prevalence of mental health disorders in the population falls to 11% (totaling approximately 4 million children in the 9 to 17 years age range); for extreme impairment (CGAS below 50), the estimated prevalence falls to 5% (1.8 million children in the 9 to 17 years age range).

Annual prevalence estimates for diagnoses of ODD and CD among children ages 9 through 17 range from 1% to 6%, depending on the level of impairment specified (Shaffer et al., 1996). Applying these prevalence estimates to the 10 million U.S. children between the ages of 3 and 5 years, it is likely that 100,000 to 600,000 preschool-aged children could meet diagnostic criteria for these disorders (assuming that age-appropriate diagnostic criteria are identified across this age continuum).

Several studies have focused on proportions of children accessing mental health services, in order to gauge levels of unmet need. Using data from the 1997 National Survey of American's Families, Kataoka and colleagues (2002) found that among children whose parents reported clinically elevated levels of mental health problems (mostly behavioral in nature), 79% had not had any contact with a mental health provider or service during the prior year; these researchers concluded that nearly 8 million children and adolescents may need but not receive mental health services. Only a small

fraction of children and adolescents in need receive mental health services, leading the U.S. Surgeon General to declare this situation a public health crisis (U.S. DHHS, 1999).

It is important to note that the apparent ambiguity in prevalence estimates reported is reflective of the need for greater consensus in regard to level of functional impairment, how such impairment is measured, and its role in defining “caseness” for epidemiological purposes. Several issues hinder the formulation of prevalence estimates for child mental health problems in general and for externalizing behavior problems in particular. For instance, reliance on DSM criteria can be controversial as well as confusing, due to unclear thresholds and boundaries between diagnostic classifications (NAMHCW, 2001), as well as differences in presentation of symptoms among various age groups of children. Further, issues of stigma, health disparities, and barriers to access may bias such estimates (U.S. DHHS, 2001). Under-representation of minorities in many studies, combined with lack of minority researchers and mental health professionals, have also been highlighted as factors contributing to underestimates of the prevalence of these problems (U.S. DHHS, 2001). Finally, lack of universally accepted, reliable, valid, and brief screening instruments poses a challenge for broad-scale epidemiological research (U.S. DHHS, 1999).

Causes and Consequences

Many studies have contributed to understanding the risk factors, protective factors, and consequences of externalizing behavior problems in young children. Progress is being made toward considering combinations of measurable factors for incorporation in testable models, rather than focusing on studies of individual factors in isolation (U.S. DHHS, 1999). The importance of such recognition of the complexity of different

developmental pathways to similar behavioral patterns has been argued by Kagan (e.g., 1997; Kagan, Snidman, & Arcus, 1998). In this section, recent research is summarized identifying factors associated with the etiology and outcomes of externalizing behavior problems in very young children.

Risk Factors

Research identifying risk factors for externalizing problems in young children has focused primarily on four broad domains: child characteristics, family factors, peer influences, and the broader social environment. Each domain encompasses a broad range of risk factors.

Hann and Borek (2001) provided an extensive review of child characteristics which have been identified as risk factors for externalizing behavior problems. These include low empathy (Cohen & Strayer, 1996; Eisenberg et al., 1996; Miller & Eisenberg, 1988); innate temperament (Bates, Pettit, Dodge, & Ridge, 1998; Guerin, Gottfried, & Thomas, 1997; Kagan, 1992; Kagan et al., 1998); daring and impulsivity (Farrington & Hawkins, 1991); weaknesses in executive functioning and inhibitory control processes (Oosterlaan, Logan, & Sergeant, 1998); biased social processing, such as a tendency to interpret others' intentions as hostile (Dodge, Pettit, Bates, & Valente, 1995; Hudley & Graham, 1993); deficits in moral reasoning and social problem solving (Rubin, Moller, & Emptage, 1987); lowered heart rate and dampened heart rate variability (Mezzacappa et al., 1997); low birth weight (U.S. DHHS, 1999); prenatal exposure to alcohol, drugs, or cigarette smoke (Brennan, Grekin, & Mednick, 1999; Brown et al., 1991; Coles et al., 1991); and possible genetic influences suggested by twin studies (Cyphers, Phillips, Fulker, & Mrazek, 1990; Edelbrock, Rende, Plomin, &

Thompson, 1995). Child-level factors alone are insufficient indicators of risk, however. Throughout a research program targeting the evaluation of temperament and its role in development, Kagan (e.g., 1992, 1997; Kagan et al., 1998; Kagan et al., 2002) has emphasized the importance of considering combinations of child-level factors with family, peer, and social-environmental characteristics in the developmental pathway to later behavioral profiles.

Family factors considered to heighten risk include poor parental responsiveness and engagement (Shaw, Keenan, & Vondra, 1994; van den Boom, 1994); young maternal age (Fergusson & Lynskey, 1993); poor maternal attachment in infancy (Erickson, Sroufe, & Egeland, 1985; Shaw, Owens, Vondra, & Winslow, 1996); hostile or rejecting parent behavior (Belsky, Hsieh, & Crnic, 1998; Shaw et al., 1998); harsh and inconsistent discipline (Campbell, 1994; Campbell, Pierce, & Moore, 1996); parental or sibling history of delinquency or criminality (Farrington & Hawkins, 1991); and marital or family conflict (Brody, Stoneman, & Flor, 1996; Schwartz, Dodge, Pettit, & Bates, 1997).

Peer influences are generally viewed as more significant with school-aged children than with preschool-aged children. However, several studies have linked peer rejection (Lochman, Coie, Underwood, & Terry, 1993; Lochman & Wayland, 1994) and friendships with aggressive peers (Farver, 1996; Kupersmidt, Burchinal, & Patterson, 1995) with externalizing behaviors even in very young children.

Risk factors linked to the broader social community are difficult to explicate, as they are frequently confounded with environmental characteristics associated with family and peer characteristics. Qualities of the social community identified as risk factors in the

literature include exposure to violent crimes in neighborhoods (Gorman-Smith & Tolan, 1998; Griffin, Scheier, Botvin, Diaz, & Miller, 1999); interaction of low SES and poor parenting (Conger et al., 1992); frequent moves (Tucker, Marx, & Long, 1998); low-ability school tracking (Farmer, 1993; Gamoran, Nystrand, Berends, & LePore, 1995); being in a classroom or daycare environment with disruptive peers (Kellam, Ling, Merisca, Brown, & Ialongo, 1998); and negative interactions and feedback from teachers (Van Acker, Grant, & Henry, 1996; Wehby, Dodge, & Valente, 1993). Many of these risk factors tend to occur in clusters, as they can be related to characteristics of communities, especially in areas of low SES (Hann & Borek, 2001). Despite difficulties in untangling the effects of social-environmental factors in the developmental pathway to behavioral problems, such efforts are crucial in developing understanding of the variation in the types, qualities, and consequences of externalizing behavior problems observed in children who differ in sociodemographic characteristics.

Protective Factors

Werner (1984) identified a range of protective factors that also represent child, family, peer, and social environmental characteristics. Studies investigating resilience have reported many protective factors, including childhood displays of empathic and prosocial behavior (Tremblay, Pihl, Vitaro, & Dobkin, 1994); healthy parent-infant attachment (Olds et al., 1998); parental expressions of validation and warmth (Feldman & Weinberger, 1994; Scaramella, Conger, & Simons, 1999); parental exploration of child's emotional experiences (Hooven, Gottman, & Katz, 1995); high degrees of community social control (Sampson, 1997; Sampson, Raudenbush, & Earls, 1997); and presence of positive behavioral supports in the school or daycare setting (Lewis, Sugai, & Colvin,

1998). Early intervention efforts often focus on developing and strengthening the protective factors thought to reduce the risk of negative outcomes.

Consequences

The long-term consequences of externalizing behavior problems in early childhood are negative and daunting. Many authors have reviewed research identifying these sequelae (e.g., Loeber, 1990; Moffitt, 1994; Patterson et al., 1989; Walker, Colvin, & Ramsey, 1995). The emergence of these behaviors at a young age, designating a child as an early starter, has been associated in numerous studies with outcomes (reviewed in Cicchetti & Cohen, 1995) such as school failure; dropping out; rejection by teachers, peers, and caregivers; delinquency in adolescence; substance abuse; adult criminal activity; lifelong dependence on social services; and higher hospitalization and mortality rates. According to Reinke and Herman (2002), early onset of behavior problems is a powerful predictor of the frequency and severity of behavior problems in adolescence.

Other negative consequences of early externalizing behavior problems have been recognized as well. The health-related quality of life of children with such problems has been demonstrated to be lower than that of children with no psychosocial issues, and as even lower in several domains than that of children with physical health problems (Sawyer et al., 2002). Lavigne, Arend, Rosenbaum, Binns, Christoffel, Burns, and colleagues (1998) and Zuckerman and colleagues (1996) also found significant positive relationships between preschool-aged children's levels of behavioral problems and rates of health care utilization. In addition, among children whose behaviors reach the diagnostic criteria of CD, rates of depression, suicidal ideation, and suicide attempts have been found to be increased (Shaffer et al., 1996), and 25-50% of children with CD are

expected to meet the diagnostic criteria for Antisocial Personality Disorder as adults (APA, 2000; U.S. DHHS, 1999). Considering the range of negative outcomes linked to early emergence of externalizing behavior problems, it is clear that the social and economic costs of this concern are substantial.

Approaches to the Problem

Primary and secondary prevention, incorporating early identification and early intervention, may be the most promising approaches for reducing the negative long-range consequences of externalizing behavior problems in young children (Hoagwood & Johnson, 2003). General approaches to the problem of externalizing behavior problems in very young children can be broadly classified as either *reactive* or *proactive*; distinctions between these are discussed in this section.

Reactive Approaches

Reactive approaches to child behavior problems generally involve intervening with the problem once it is already established. Such approaches correspond to *tertiary* prevention (Pransky, 1991), or attempting to prevent a child's already significant behavior problems from becoming worse. Tertiary prevention is the type of intervention most often offered in the mental health system, and arguably in most school systems as well. For example, Duncan and colleagues (1995) and Forness and colleagues (1998) reported that school services for behavioral difficulties often were not implemented until late elementary school, despite parental recognition of the problems as early as preschool. In addition, Forness and Kavale (2001) specifically described the special education system's efforts with behavior issues as primarily addressing already entrenched

problems. Such reactive approaches are generally found to be more expensive with less dramatic improvements achieved (Pransky, 1991).

Proactive Approaches

In contrast, proactive approaches involve *primary* and *secondary* prevention (Pransky, 1991) and are often associated with a public health perspective. These approaches focus on early identification and early intervention, in which screening is a crucial component. In response to externalizing behavior problems in very young children, the goals of primary and secondary prevention efforts are either to prevent problems from developing by reducing their risk factors, or to prevent fledgling problems from developing into clinical disorders. Hoagwood and Johnson (2003) have made compelling arguments for a preventive, public health orientation in addressing child behavior problems. Pransky (1991) argued that many social problems share overlapping and common underlying risk factors, and that by fostering collaboration and preventive efforts among service sectors, these risk factors could be addressed more effectively. Thus, effective prevention efforts should involve collaboration across systems (Reid, 1993), with unbiased and accurate screening methods to identify very young children in need of further assessment.

Several authors have described successful efforts with early identification and early intervention in preventing later problems among children who received services. For example, Minkovitz and colleagues (2003) evaluated the collaborative Healthy Start program, identifying benefits such as increased satisfaction of parents with health services and increased compliance with preventive health measures. Hawkins and colleagues (1999) followed children from early elementary school grades through age 18,

reporting that behaviorally at-risk children who received intervention services in early elementary school grades demonstrated reduced rates of school failure, teen pregnancy, having multiple sex partners, and delinquent behavior, compared to those receiving services in later grades. In the medical literature, Olds and colleagues (1998) described positive outcomes following a primary prevention nurse home-visiting program aimed at building secure attachments between parents and infants. Other successful projects have similarly targeted early identification and intervention as effective proactive strategies; a selection of these are reviewed by Simpson and colleagues (2001).

Many authors have argued that primary and secondary prevention programs, implemented across settings, with the goal of changing the trajectory of potentially negative behaviors, are needed to address the problem of externalizing child behavior problems (e.g., Boyce, Hoagwood, Lopez, & Tarullo, 2000; Cicchetti & Cohen, 1995; Coie et al., 1993; Forness et al., 1996; Greenspan, 1992; Kazdin & Wassell, 1999; Loeber & Farrington, 1998; Patterson et al., 1989; Patterson, Reid, & Dishion, 1992). Screening is a crucial component of early identification efforts. The challenge in implementing a proactive approach is to determine where the barriers to prevention lie and how to overcome them.

Barriers to Early Identification and Early Intervention

While consensus among researchers is apparent regarding the need for early identification and early intervention with externalizing behavior problems in children, a myriad of barriers hinder the implementation of prevention strategies. Examples of such barriers include issues regarding (a) the complexity of screening within a developmental and socio-environmental context; (b) social attitudes undermining a prevention approach;

and (c) fragmentation among systems—in particular, the educational and health care systems—charged with addressing the social and public health problem of externalizing behavior problems. Each of these topics is addressed briefly in this section.

Complexity of Screening Very Young Children

Whether in research, educational, or health care arenas, difficulties in screening very young children for externalizing behavior problems pose barriers to the implementation of primary and secondary preventive practices. The complexity of determining whether a child's externalizing behaviors constitute a problem or disorder rather than a typical stage of development in a mentally healthy child presents challenges in assessment (Merritt, Thompson, Keith, & Johndrow, 1993; Reijneveld, Brugman, Verhulst, & Verloove-Vanhorick, 2004; Task Force on Research Diagnostic Criteria, 2003; Thomasgard & Metz, 2004). This is especially true among children with varying biologic, familial, and social-environmental characteristics and histories (Kagan, 1997). Attention to developmental stage, level of functioning, and social environment is crucial, because consideration of the mere presence of diagnostic symptoms may lead to errors in assessment—a behavior which is problematic for one child in one situation may be developmentally appropriate for another. Reliance on symptoms listed in the DSM as the sole indicators of a behavioral disorder disregards the fact that most diagnostic categories for young children have not been validated through research, but rather have been derived from those created for adults (NAMHCW, 2001; U.S. DHHS, 1999). Further, the boundaries, thresholds, and degrees of overlap for disorders in children are the subjects of much debate (NAMHCW, 2001). Therefore, consideration of the child's functioning in the context of development and social environment is necessary.

Other complexities in screening further impede reliable and valid identification of externalizing behavior disorders in children. Aside from the DSM categorization system, no universally accepted language or measurement approach exists (U.S. DHHS, 1999). Hesitancies to “label” a child, correctly or incorrectly, pose philosophical barriers to assessment (Hinshaw, 2005; Kauffman, 1999). Issues such as the stigma attached to mental health disorders and health disparities in accessing services (Hann & Borek, 2001; U.S. DHHS, 2001), as well as gaps in relevant research (e.g., the limited number of studies focusing on racial minorities and female children), further impede accurate identification of affected children. Underlying these issues is an array of social attitudes undermining the prevention approach.

Problematic Social Attitudes

In a review targeted to special educators, Kauffman (1999) provided compelling examples of problematic social attitudes posing barriers to implementation of preventive practices, despite empirical support for a proactive approach. Several of the attitudes described are relevant beyond the field of education, pervading service philosophies in the health care field as well.

Societal objections to early identification efforts such as screening include those based on (a) concerns that children will be labeled and stigmatized; (b) distaste for a “medical model”; (c) characterization of intervention services as failure-driven; (d) preference for false negatives over false positives; and (e) claims of diversity (Kauffman, 1999). According to Kauffman’s argument, each of these attitudes undermines attempts to implement early identification measures, which are often characterized as potentially damaging. From this perspective, screening leads to harmful practices such as labeling

children (correctly or erroneously), focusing on pathology, and failing to accept cultural and other differences. Specific examples of the barriers to prevention posed by such attitudes can be observed in the fragmentation of approaches across involved systems, at the level of policies as well as practice.

Fragmentation across Systems

Approaches to identification and treatment of children with externalizing behavior problems are determined by the systems, policies, institutions, and agencies involved. Unfortunately, a lack of coordination and differences in philosophy across systems has resulted in a fragmentation of approaches (New Freedom Commission on Mental Health, 2003). Children between the ages of 3 and 5 still spend most of their time at home, even if they attend preschool or kindergarten. Thus, the three primary institutional systems with the earliest opportunity to identify children with externalizing behavior problems are the family, the educational system, and the health care system. If parents are concerned about their child's behavior, they are very likely to approach either their child's teacher or physician for more information. Which system is contacted may have a significant impact on what action is taken. This issue can be illustrated through a brief overview of salient policy and practice issues within two systems influential in the identification of externalizing behavior problems in children: the educational system and the health care system.

The educational system. The key educational policy related to early identification of externalizing behavior problems is the U.S. Department of Education's Individuals with Disabilities Education Act (IDEA; 1990, 1997, 2004). This federal policy mandates that children with disabilities be identified and receive free and appropriate education.

Children ages 0 to 21 are eligible for support services according to IDEA, and each state is mandated to have a systematic Child Find effort to identify and serve all eligible children. The quality of Child Find efforts, however, varies from state to state, and no agreement exists regarding whether at-risk children (as opposed to children clearly exhibiting emotional or behavioral disorders) are eligible for support services required by IDEA (U.S. Department of Education, 2003), hampering primary and secondary prevention approaches.

Issues regarding the implementation of policy within actual practices in the educational system further illustrate barriers to early identification of externalizing behavior problems. In general, assessment for behavioral disorders in the schools is only initiated after a child's behavior is deemed "uncontrollable" by a regular classroom teacher. At that time, a series of meetings ensues with the child's parent(s), teacher, school guidance counselor, school psychologist, and other school staff. The process of assessment in the schools involves, in effect, a gatekeeper system, in which children who do not meet strict criteria for certain disabilities, as defined in IDEA (1990, 1997, 2004), do not receive support services (U.S. DHHS, 1999). Often, children with genuine externalizing behavior problems are classified into other areas of disability due to attempts to avoid the stigma of an *emotionally disturbed* (ED) classification (U.S. DHHS, 1999). Alternatively, rather than being classified as ED, they may simply be considered discipline problems, resulting in punishment, suspensions, and even expulsion, rather than support services (Merrell & Walker, 2004). Further complicating access to services for behavior problems is the *socially-maladjusted exclusionary clause* in IDEA, which allows for the exclusion of students whose actions are deemed to reflect social

maladjustment rather than emotional or mental health disturbance. According to Merrell and Walker, interpretations of this clause have often assumed that children's behavior problems which appear purposive or goal-oriented are evidence of social maladjustment rather than ED, resulting in ineligibility for ED classification under IDEA. Because intentional misbehavior is one criterion of DSM behavioral disorders, children with diagnoses of ODD, CD, or other disruptive behavior disorders are thus often deemed to be discipline problems rather than made eligible for support services through the educational system (Cheney & Sampson, 1990; Clarizio, 1992; Merrell & Walker, 2004; Skiba & Grizzle, 1992). Forness and Knitzer (1992) argued that such problems with the federal definition are related to under-identification of children in need of behavioral services in school settings.

The inefficient process of assessment by the school system can result in lengthy time lags between a parent's recognition of a problem and an accurate identification by the school (U.S. DHHS, 1999; Yoshikawa & Knitzer, 1997). Further, school assessments are often handled by staff who are insufficiently trained to evaluate behavioral disorders (U.S. DHHS, 1999). Hoagwood and Erwin (1997) reported that fewer than 1% of children in the public school system have been identified for ED services, despite prevalence estimates of need up to 10 times higher. Referrals for behavioral services reportedly peak in late elementary school and middle school grades, despite parental awareness of issues dating back to preschool for many children (U.S. DHHS, 1999). This pattern is incompatible with a prevention perspective.

Specifically regarding prevention efforts with preschool-aged children, Head Start has been identified as a promising arena for early identification due to its population of

at-risk children and its focus on development and school readiness. However, many authors report significant problems with the identification efforts in this setting as well (Pianta & Cox, 1999; Yoshikawa & Knitzer, 1997; Zigler & Styfco, 1994). Despite conservative estimates of need for behaviorally-oriented services ranging from 6% to 10%, only 1% of children in Head Start receive such services (Redden et al., 2003; Sinclair, Del'Homme, & Gonzalez, 1993).

Dedicated researchers and school officials continue to work toward improvement of early identification and intervention services in the school setting (see especially Feil, Severson, & Walker, 1998, for a system designed for preschool-aged children). However, for children ages 3 to 5 who are not reliably identified via Child Find programs, resources in addition to the educational system may be needed to increase the likelihood of early identification of behavior problems.

The health care system. At the federal level, there are many health care policies and agencies relevant to behavioral problems in very young children, and to mental health issues among children in general. Two key areas are briefly highlighted: (a) efforts in mental health care coordinated by the federal Substance Abuse and Mental Health Services Agency (SAMHSA), and (b) screening programs mandated by Medicaid and State Children's Health Insurance Program (SCHIP).

Within the U.S. Department of Health and Human Services, SAMHSA provides funding and support to state and local efforts to administer and implement mental health and substance abuse services. SAMHSA also leads the Systems of Care Initiative, a laudable effort initiated in 1993 to improve collaboration of mental health and substance abuse services across systems and sectors. While SAMHSA maintains a website of model

programs for a variety of mental health and substance abuse issues in several settings, widespread implementation of best practices has yet to occur (U.S. DHHS, 1999).

Additionally, the mental health system is not currently a likely candidate to provide early identification services to very young children. Most children served by this system have already been identified as needing services, and in general, preschool-aged children are rarely in contact with mental health agencies (U.S. DHHS, 1999).

Also within the U.S. Department of Health & Human Services, the Medicaid and SChip programs are intimately linked with the provision of assessment and treatment for externalizing behavior problems in children. These programs, which provide health insurance coverage to low-income children and families, are implemented at the state level, with great variability in quality and coverage (U.S. GAO, 2003). A mandated section of Medicaid is the Early Periodic Screening and Developmental Testing (EPSDT) program, which requires providers receiving Medicaid reimbursements to conduct periodic screenings of children for health, developmental, vision, and dental needs; however, while social and emotional development are clearly related to behavioral disorders in children, behavioral screening is not universally included in EPSDT (U.S. GAO, 2003). Furthermore, children with private health insurance or no insurance are even less likely to receive routine screenings for behavioral problems, as few, if any, systematic psychosocial prevention practices exist in most health care settings.

As highlighted above, the implementation of federal and state policies within the actual practices of health care agencies results in uneven attempts to implement a prevention approach with regard to externalizing behavior problems in children. Despite these limitations, the health care system is a key resource in this area, due in part to its

frequent contact with the majority of children aged 3 to 5 years (U.S. DHHS, 1999). The remainder of this chapter focuses on one particular health care setting—pediatric primary care—as a crucial component in the improvement of early identification efforts with preschool-aged children exhibiting externalizing behavior problems.

The Potential of Pediatric Primary Care

Pediatric primary care, as a system which follows most young children from birth to school age for well child and acute care visits, has been identified as an optimal arena for screening and early intervention efforts (AHRQ, 2002). While the educational system (including the federal Head Start program) is charged with identifying and assisting all children in need of special services, issues regarding timely and effective screening and intervention have plagued educational institutions. Primary care could serve as an additional prong to these efforts, especially for young children with limited contact with schools. The potential of pediatric primary care as a vital contributor to efforts toward early identification and intervention with very young children with externalizing behavior problems is promising. In this section, two related topics are reviewed: the evolution of social acceptance of externalizing behavior problems as a disorder requiring professional treatment, and the growing recognition of the potential role of pediatric primary care clinicians as stakeholders in this arena.

The “Medicalization” of Externalizing Behavior Problems in Children

As the recognition of externalizing behavior problems in children emerged over the past century, social understanding of this issue began to acknowledge the need for professional interventions. Conrad and Schneider (1980), in an important work, discussed significant historical changes in the social construction of deviant behavior in society.

Deviance, according to the authors, has been attributed to moral failings, criminal intention, and sickness, dependent on in which era it presents itself. In modern Western society, Conrad and Schneider argued, the stature of the field of medicine as a source of scientific knowledge and authority has led to a shift in the social definition of deviance from a moral sin or willful criminal act to a state of illness beyond the direct control of the afflicted person. These authors suggested that social judgment shifted from a preference for punishment or moral absolution in response to these behaviors, to conceptualization of the patient suffering with deviant behavior as adopting the “sick role” as described by Parsons (1951): exempt from normal social responsibilities, not responsible for the condition, desiring recovery, and intending to seek out and comply with treatment. In response to this shift in social understanding, the provider of medical treatment became an agent of social change, intervening with the “sickness” of deviance.

The work of Conrad and Schneider (1980) has been referenced in a description of how medical advances in reducing infant and child mortality rates have resulted in expansion of authority in the pediatric field from treating biological diseases to managing child behavior (Pawluch, 1983). Tuchman (1996) also drew upon the theorized shift “from badness to sickness” in suggesting that one reason for incongruence in approaches to this problem between school and medical settings may be the reluctance of the educational system (and other social institutions) to fully accept the so-called “medicalization” of deviance. Tuchman argued that the extent of schools’ acceptance of this paradigm, in particular, is in obtaining physician diagnoses to justify to school boards the provision of expensive special education services and supports to children with behavioral problems. She presented results of an extensive qualitative study suggesting

that once a physician's diagnosis is secured, most school personnel revert to their previous assumptions regarding the home environment and parenting deficits as the sole etiology of a child's disruptive behaviors. Tuchman's research contributes to understanding the problem of child externalizing behavior problems through a social constructionist perspective, emphasizing disparate shared meanings in different settings, and how such clashes in meaning can stymie collaboration.

Identification in Pediatric Primary Care: Problems and Promise

In tandem with the development of a sociological literature on the medicalization of behavior problems, researchers in primary care moved in a similar direction. The literature on primary care treatment of psychosocial problems, including behavioral issues, originated in the 1970s, when Haggerty (1974) referred to such problems as *the new morbidity*, considered to be outside the realm of traditional health care. Several years later, Regier and colleagues (1978) described primary care providers as the de facto mental health service providers, due to the proportion of patients with mental health issues who received care solely from their primary care physicians. Evidence of attention to behavioral issues in pediatric practice is seen in recent increases in prescriptions for psychotropic medications for children, rising from 1.4% to 3.9% between 1987 and 1996 (Olfson, Marcus, Weissman, & Jensen, 2002). Further, approximately 85% of psychotropic medications taken by children are prescribed by pediatricians (Goodwin, Gould, Blanco, & Olfson, 2001). According to the results of the 59th Periodic Survey of members conducted by the American Academy of Pediatrics, the vast majority of responding pediatricians agreed that they should be responsible for identification of

behavioral health issues, including externalizing behavior problems, in their patients (Stein et al., 2008).

However, under-identification of children with psychosocial issues has been a persistent problem in the primary care setting. E. J. Costello and Edelbrock (1985) reported that physicians they surveyed identified an average of 5.7% of their patients as needing assistance with psychosocial issues, reflecting only 17% of those patients identified by psychologists using standardized interviews and instruments. Findings from several studies regarding recognition of mental health issues in pediatric primary care revealed that over 60% of parents of children with significantly elevated levels of psychosocial problems reported that they only received mental health care from their physicians, despite physicians' recognizing only 1 of every 7 children in need of such services (E. J. Costello, 1986). Costello and colleagues (1988) further concluded that in a rush to diagnose, or via misdiagnosis, physicians missed 83% of patients presenting with clinically significant elevations of psychosocial problems, as identified by a psychologist.

Others have reported similar findings. Lavigne and colleagues (1993) found that physicians had a sensitivity rate of 20% and specificity of 93% in identifying children with significant mental health problems, as compared to psychologists' assessments. While Kelleher and colleagues (2000) have suggested that identification of psychosocial problems in pediatric primary care has increased from 7% to 18% in the past 20 years, most authors agree that under-identification and substandard assessment practices are the norm. In a recent Fellows Survey conducted by the American Academy of Pediatrics (2000), findings indicated that most pediatricians prefer to use informal methods to assess for child behavioral or other mental health-related issues, despite the lack of precision

and increase in bias associated with such practices. These findings are of particular concern due to the gatekeeper role filled by physicians with regard to children's access to specialized behavioral and mental health services.

Concerns regarding differences between physicians' and parents' perceptions of what constituted a behavioral problem have been raised (U.S. DHHS, 1999). The Surgeon General's report on child mental health summarized research on a range of issues posing problems for physicians in this capacity, including difficulties making referrals to community resources; a fear of opening a "Pandora's box" via asking about psychosocial issues; and a lack of universally accepted, brief, reliable, and valid screening tools—not to mention time for physicians or other staff to interpret them (U.S. DHHS, 1999). Screening alone is not the only issue, however; in fact, Horwitz and colleagues (1998) found that while 50% of parents in their study disclosed psychosocial concerns to their child's pediatrician, less than 40% of the time did the physician respond with appropriate guidance, reassurance, information, or referral. Schuster and colleagues (2000) concurred that health care professionals rarely offer parents information regarding recommended child-rearing practice. Several characteristics of the health care setting pose issues in this regard, including the short (11-14 minutes) length of the average session (Woodwell, 1999); the lack of systematic training on child mental health issues received by physicians and nurses (Gardner, Kelleher, Pajer, & Campo, 2004; Hawkins-Walsh & Stone, 2004; Horwitz, Leal, Leventhal, Forsyth, & Speechley, 1992); and the primary focus on physical and cognitive health and development in the primary care setting (Borowsky, Mozayeny, & Ireland, 2003).

Despite these problems, the potential impact of physicians on increasing the chances of children receiving needed mental health services has been supported by Lavigne, Arend, Rosenbaum, Binns, Christoffel, Burns, and colleagues (1998). In this study, researchers found that among preschool-aged children with clinically significant levels of behavior problems, once level of severity and age of child were controlled, the only significant predictor of whether a child received services that year was whether they had a physician's referral. Physician referral doubled to quadrupled the odds that a child had seen a mental health specialist, compared to the odds for children without physician referrals.

In short, despite their potential positive effects on access to early intervention services, front-line staff in pediatric primary care settings are often under-prepared and under-supported in screening for externalizing behavior problems within a developmental and social environmental context. Even when the goal is purely to triage and refer for specialty services, the lack of universally accepted valid, reliable, and brief screening instruments, and the restrictions inherent in the pediatric primary care system, may impede accurate identification and referral of these children (U.S. DHHS, 1999). However, to achieve early identification of very young children with externalizing behavior problems, screening in the pediatric primary care setting may be critical. According to principles set by the World Health Organization (Strong, Wald, Miller, & Alwan, 2005), screening should involve brief, reliable, and valid methods, acceptable to consumers, with acceptable cost-benefit ratios, which result in high yield (i.e., high numbers of otherwise unidentified children would, as a result of screening efforts, receive services). Screening instruments which are (a) well-constructed; (b) developmentally and

culturally appropriate; (c) low cost; and (d) quickly administered, scored, and interpreted would be valuable tools for pediatric primary care providers in this regard.

Standardized Screening Tools for Externalizing Behavior Problems

Given the results of the American Academy of Pediatrics (2000) Fellows Survey indicating that physicians prefer informal methods of assessment, incorporating acceptable and valid instruments into a more systematic assessment approach may be important (Halfon et al., 2003). Use of reliable and valid standardized instruments has been shown to improve the accuracy of screening for externalizing behavior problems in children (L. G. Hill et al., 2004). While limitations of parent-completed reports of behavioral symptoms have been identified (e.g., Kagan et al., 2002), use of such measures as screening tools, rather than diagnostic instruments, may be valuable. Such systematic screening could be helpful in improving the early identification of children in need of intervention in primary care, facilitating referrals to behavioral or mental health services provided by social workers and other mental health professionals.

Novel tools and resources for assessment and systematic research have been developed, including DSM criteria adjusted specifically to account for the rapidly changing developmental status of preschool-aged children (the RDC-PA; Task Force on Research Diagnostic Criteria, 2003), as well as the DSM-PC (Wolraich, Felice, & Drotar, 1996), a version of the Diagnostic and Statistical Manual developed specifically for use in primary care settings. The DSM-PC organized content within a developmental context, illustrating the continuum of emotional and behavioral functioning (Drotar, 1999, 2004; Jellinek, 1997; Kelleher & Wolraich, 1996) and making it a promising tool for use in pediatric primary care settings. Apart from integrating new classification systems

targeted toward very young children, however, the use of standardized screening tools appropriate for pediatric primary care practice could provide a simple, low-cost, time-efficient strategy for early identification.

Numerous instruments intended to measure behavior problems among young children exist, including the Preschool Behavior Questionnaire (PBQ; Behar & Strinfield, 1974); the Preschool and Kindergarten Behavior Scales 2nd Edition (PKBS-2; Merrell, 2003); Burks Behavior Rating Scale (BBRS; Burks, 1996); the Behavior Assessment System for Children 2nd Edition (BASC-2; Reynolds & Kamphaus, 1992); the popular Child Behavior Checklist/1.5-5 (CBCL/1.5-5; Achenbach & Rescorla, 2000); and others. Each of these instruments, while useful in other settings, exhibits shortcomings particular to their use in pediatric primary care settings. For example, the intended completer of the PBQ is a preschool teacher, which may not be feasible or efficient in initial screenings performed in the primary care setting. On the PKBS-2, 76 items must be answered and scored; the cost for materials may be excessive for some settings; and the norming samples used are not described in terms of SES, possibly limiting interpretation of scores with disadvantaged populations. Similarly, the BBRS presents several problems for administration and interpretation in the primary care setting, including the need for hand-scoring of 105 items, cost of screening materials, and lack of psychometric information available regarding reliability and validity. While it is a popular and well-supported assessment tool, the BASC-2 preschool rating form consists of 132 items—excessive for use as an initial screening instrument in primary care settings. Further, the SES of children used in norming samples was not reported by its authors. Finally, the CBCL/1.5-5, while arguably the gold standard for assessment of child behavior problems, also poses

challenges for use as an initial screening tool in primary care due to its length of 99 problem items. In addition, costs associated with both the BASC-2 and the CBCL/1.5-5 may present barriers to widespread use in primary care.

Scales exist which address the issues identified above regarding utility for screening in the primary care setting, specifically incorporating shorter length and lower cost. For example, the PSC-17 (Gardner et al., 1999) was developed specifically for use in pediatric primary care. Consisting of only 17 items, the PSC-17 is intended to serve as a general screening tool for various psychosocial concerns in children. This shortened form of the original PSC (Jellinek & Murphy, 1988) includes subscales measuring internalizing, attention, and externalizing problems (Gardner et al., 1999). The externalizing subscale of the PSC-17 targets behaviors commonly associated with ODD and CD. Both the PSC-17 and the original PSC are available at no cost from the authors, who encourage their use in practice and research. A survey of pediatricians who tested the full-length PSC in practice revealed that 96% intended to retain it as part of their normal clinical routine (Bishop, Murphy, Jellinek, & Dusseault, 1991), suggesting its acceptability to many practicing physicians. The format of the PSC-17 is brief, and validity studies have suggested that it distinguishes well between clinic-referred and non-referred children (Gardner et al., 1999), though sensitivity estimates were lower than expected with some populations (Gardner, Lucas, Kolko, & Campo, 2007). Its authors caution that scores should be used only as suggestive of the need for further assessment, consistent with the purposes of a screening instrument.

Another brief, freely available instrument applicable to screening for behavioral problems in the primary care setting is the BPI (Peterson & Zill, 1986; Zill, 1990). The

BPI consists of 28 items for parent report (26 for use with preschool-aged children), and was developed to provide a shorter instrument suitable for screening in surveys, based on earlier work by the authors of the CBCL (Achenbach & Edelbrock, 1981). Subscales of the BPI measure headstrong, antisocial, peer problems, anxious/depressed, hyperactive, and immature/dependent domains of behavioral problems (Zill, 1990). The headstrong, antisocial, and peer problems subscales of the BPI are especially relevant to screening for externalizing behavior problems. The BPI has been used in several national longitudinal studies, including the Child Health Supplements to the National Health Interview Survey (National Center for Health Statistics, 1989; Zill, 1985) and the Child Supplements to the National Longitudinal Survey of Youth (NLSY; Center for Human Resource Research, 2000). While not often reported as an instrument used in clinical practice, its utility in distinguishing children with and without clinically significant psychosocial symptoms has been demonstrated in several studies (e.g., Gortmaker, Walker, Weitzman, & Sobol, 1990). In addition, its similarity to the CBCL (Achenbach & Edelbrock, 1981), combined with its brevity, make the BPI a good candidate for screening in pediatric primary care settings.

Problems with Identified Instruments

Though progress has been made in developing instruments such as the PSC-17 and BPI, further research on the appropriateness of these instruments for screening very young children in primary care is still needed (Borowsky et al., 2003; Jellinek et al., 1999). For research and practice related to externalizing behavior problems of young children, the performance of the relevant subscales of each instrument (i.e., the externalizing subscale of the PSC-17 and the headstrong, antisocial, and peer problems

subscales of the BPI) are particularly important to understand. Possible shortcomings of the full scales have been identified by several authors, including concerns about their reliability and validity with younger children, minority children, and children of low SES. These concerns, as well as issues related to the underlying psychometric theory behind their development and evaluation, are discussed in detail in the following sections.

Screening of very young children. Both the PSC-17 and the BPI are intended for use with children ages 4 through 17, and their utility in screening children below the age of 4 has not been established. Further, both scales are available in only one form, as opposed to age-adjusted instruments such as the CBCL (Achenbach, 1991; Achenbach & Rescorla, 2000) and BASC-2 (Reynolds & Kamphaus, 1992). While the convenience of using single version forms in a pediatric primary care setting is appealing, the utility of scale items for measuring behavioral concerns which may present differently in very young children is unknown. The developmental context of behavioral problems for preschool-aged children may influence the measurement performance of any such tool in important ways (Kagan et al., 2002).

In a report describing the performance of the PSC-17 among children seen in primary care settings, the youngest children included were age 4 (Gardner et al., 1999). These very young children were grouped in analyses with school-aged children up to age 7, and psychometric properties were not described within age groups. Although an investigation explicitly examining the performance of the full-length PSC among 4 and 5 year old pediatric patients concluded adequate reliability and validity (Little, Murphy, Jellinek, & Bishop, 1994), consensus regarding its performance in this age group has not been reached. Assessing the performance of the full-length PSC in screening children

aged 2 through 18 for psychosocial issues, Navon and colleagues (2001) reported lower sensitivity and specificity for children under age 6. In addition, lower prevalence of psychosocial problems among preschool-aged children has been estimated in several studies evaluating the feasibility of widespread use of the full-length PSC in various primary care settings (Jellinek et al., 1999; Pagano & Murphy, 1996), despite suggestions that prevalence should be nearly equivalent across age groups (U.S. DHHS, 1999).

Notably, for the full-length PSC, the cut-scores recommended for use with preschool-aged children are lower than those suggested for school-aged children (Jellinek et al., 1999). However, for the PSC-17, no age-adjustments in cut-scores have yet been suggested (Gardner et al., 1999). A more intensive evaluation of the psychometric properties of the items included in this scale may be warranted, in order to understand the appropriateness of PSC-17 items for measurement of externalizing behavior problems in preschool-aged children.

While the BPI has been used extensively in national longitudinal studies of correlates, predictors, and outcomes of child behavior problems, its psychometric properties have rarely been examined in depth. As with the PSC-17, most studies using the BPI have considered only children ages 4 and older, due to datasets available for secondary analysis; no studies have reported measurement performance with behavior problems of children under age 4. Normed scores, including both percentiles and standard scores, have been calculated for children ages 4 and 5 (Center for Human Resource Research, 2000). In these analyses, raw scores associated with standardized means and clinical cut-offs tended to be slightly higher for very young children than for older children. Previous investigations of the psychometric properties of the BPI among

various age groups did not report differences in indicators of reliability and validity (Parcel & Menaghan, 1988; Zill, 1985, 1990). However, as with the PSC-17, further evaluation of the quality of measurement of externalizing behavior problems in very young children would be helpful in determining the BPI's potential utility in a screening capacity.

Disparities in identification. Variability in symptom expression and perception across population subgroups is an accepted characteristic of mental health problems worldwide (U.S. DHHS, 2001). One result of such variability can be differences in base rate estimates of the prevalence of mental health problems among such subgroups—for example, among groups differing by sex, race, or SES. When screening tools are used to assess for possible externalizing behavior problems among very young children, it is vital that these instruments are not biased against particular groups. Bias in screening instruments can result in both over-identification and under-identification of children in need of further assessment and services, limiting the efficiency and accuracy of primary and secondary prevention efforts (Spencer et al., 2005). Social injustices and health disparities are perpetuated by such biases. Further, inherent differences in children who experience similar clustering of behavioral symptoms may affect the quality of measurement offered by screening instruments (Kagan et al., 2002).

In a study using the full-length PSC with preschool-aged children, mean scores of boys were significantly higher than those of girls, and more boys than girls had scores exceeding the clinical cut-off (Jellinek et al., 1999). Differences by sex were also reported in a sample of Austrian preschool-aged children, again with boys scoring higher than girls (Thun-Hohenstein & Herzog, 2008), as well as in a sample of school-aged

Dutch children (Reijneveld, Vogels, Hoekstra, & Crone, 2006). Similar results have been reported in studies using the BPI as a measure of behavior problems (Parcel & Menaghan, 1988). This finding is common in the child mental health literature (Shaffer et al., 1996; U.S. DHHS, 1999), in which boys with externalizing behavior problems are routinely identified more frequently than girls. While these findings may simply reflect actual prevalence differences between boys and girls, no studies were found which investigated the possibility of bias in individual items comprising these screening tools.

Regarding screening of minority populations, several authors have argued that the full-length PSC is adequately sensitive and specific (Jellinek et al., 1999; Murphy et al., 1992), though consensus on this point is lacking (e.g., Jutte et al., 2003). Simonian and colleagues (1991) and Simonian and Tarnowski (2001) assessed the cultural sensitivity of screening instruments used in primary care settings. These authors argued that not only has insufficient attention been directed toward this concern, but that their data suggest that (a) race is significantly associated with parental responses regarding child behavior, and (b) clinical cut-offs are not identical between racial groups. In assessing the BPI for equivalence across U.S. ethnic groups, Spencer and colleagues (2005) conducted an in-depth confirmatory factor analysis, concluding that the standard BPI subscales are “valid principally for White children” (p. 585), but not necessarily for minority children. Cultural differences in full-length PSC scores have also been revealed in several international studies (e.g., Reijneveld et al., 2006; Thun-Hohenstein & Herzog, 2008), leading some to suggest the need for adjusted cut-scores for particular populations. Given concerns regarding both under- and over-identification of children of minority status for

behavioral services, clarity regarding possible biases in the items comprising these scales is needed.

Similarly, scale performance with children of low SES is of concern. In a national feasibility study of use of the full-length PSC in primary care settings, Jellinek and colleagues (1999) reported that more than twice as many low-income as middle-income children were identified with psychosocial problems—though the low-income group used was arguably more representative of a lower-middle-income group (Simonian et al., 1991). In an earlier study, Jellinek and colleagues (1995) simultaneously found higher rates of psychosocial dysfunction among lower SES children, but also lower sensitivity of the PSC with low-income children (80%) as compared to middle-class children (95%). Possible effects of SES on full-length PSC scores were also reported among Dutch children (Reijneveld et al., 2006), though another international study detected no such differences in Austrian children (Thun-Hohenstein & Herzog, 2008). Regarding the BPI, no studies were located which explicitly tested the quality of measurement of the instrument among groups of different SES. Though the samples comprising the NLSY datasets were weighted heavily toward lower SES groups, it appears that performance of the BPI with these groups has been assumed to be acceptable. Examination of this assumption is important in evaluating the quality of measurement provided by the BPI both for studies analyzing the large surveys in which it has been used, as well as for the potential use of the BPI as a screening instrument in clinical practice.

Shortcomings of Classical Test Theory. An additional area in which many screening instruments, including the PSC-17 and BPI, may exhibit weaknesses is in relation to their development and evaluation using methods based on Classical Test

Theory (CTT). In CTT, items in a scale are generally summed to yield a total score, representing true score plus error (DeVellis, 2003; Nunnally & Bernstein, 1994). Estimates of scale reliability, validity, and standard error of measurement (SEM) in CTT are inextricably linked to the sample of respondents from whom the scale data were collected; thus, interpretations regarding the meaning of a child's score on an externalizing behavior subscale depend on the degree to which the norming sample was similar to the child in question (Crocker & Algina, 1986; Embretson & Reise, 2000; Hambleton et al., 1991). In addition, the assumption of constant SEM across all score levels has been demonstrated to be untenable (Hambleton et al., 1991; see Nugent, 2005, for an example), resulting in lack of certainty regarding the magnitude of measurement error along the continuum of the measured construct. A more detailed explanation of the limitations of CTT for scale development and evaluation is presented in Chapter III.

Information on the quality of measurement of externalizing behavior problems provided by subscales of the PSC-17 and BPI is currently limited to conclusions based on CTT methods and assumptions. It is possible that the application of modern methods of item and scale evaluation could yield valuable information regarding the measurement properties of these instruments, which could guide their usage in primary and secondary prevention efforts in pediatric primary care.

Summary and Research Questions

Externalizing behavior problems in very young children pose a serious social and public health problem in the U.S. Characterized by early emergence of behaviors associated with diagnoses of ODD and CD, preschool-aged children who are early starters are likely to exhibit increased severity of a developmental pathway of antisocial

behaviors (Ge et al., 2003; Hann & Borek, 2001; Moffitt, 1994; Patterson et al., 1989). These children are at risk for escalating behavior problems including bullying, physical aggression, cruelty to animals, vandalism, and violent criminal acts (Hann & Borek, 2001).

Prevalence estimates of externalizing behavior problems in very young children are difficult to obtain due to (a) the lack of epidemiological studies in this age group (U.S. DHHS, 1999), and (b) challenges in assessment related to ambiguous diagnostic thresholds in a developmental context (NAMHCW, 2001). Issues of stigma, health disparities, and barriers to access have also been identified as hampering accurate prevalence estimates (U.S. DHHS, 2001). However, based upon studies assessing older children, between 1% and 6% of preschool-aged children are thought to meet diagnostic criteria for ODD or CD (Shaffer et al., 1996), and up to 20% may exhibit sub-threshold psychosocial symptoms (U.S. DHHS, 1999). The vast majority of these children do not receive specialized services (Kataoka et al., 2002).

Many studies have contributed to the understanding of the risk factors, protective factors, and consequences of externalizing behavior problems in young children. A range of child characteristics, family factors, peer influences, and social environmental characteristics have been identified as risk and protective factors (Hann & Borek, 2001; Werner, 1984). Acknowledgment of the complex sets of interacting risk and protective factors is crucial to improving understanding of behavioral patterns observed in children (Kagan, 1997). Several authors have reviewed research indicating that the consequences of externalizing behavior problems in early childhood can be serious and costly, including, but not limited to, school failure, substance abuse, adult criminal activity, and

higher hospitalization and mortality rates (Cicchetti & Cohen, 1995; Loeber, 1990; Moffitt, 1994; Patterson et al., 1989; Walker et al., 1995). Issues such as decreased health-related quality of life (Sawyer et al., 2002), increased rates of health care utilization (Lavigne, Arend, Rosenbaum, Binns, Christoffel, Burns et al., 1998; Zuckerman et al., 1996), increased rates of suicidality (Shaffer et al., 1996), and adult diagnoses of Antisocial Personality Disorder (U.S. DHHS, 1999) have also been associated with early emergence of externalizing behavior problems.

The literature reflects consensus that primary and secondary prevention approaches, incorporating early identification and early intervention, may be the most promising strategies for addressing this problem (Boyce et al., 2000; Cicchetti & Cohen, 1995; Coie et al., 1993; Forness et al., 1996; Greenspan, 1992; Hoagwood & Johnson, 2003; Kazdin & Wassell, 1999; Loeber & Farrington, 1998; Patterson et al., 1989; Patterson et al., 1992). While tertiary prevention characterizes most services routinely offered in the mental health and educational systems, such reactive approaches are generally found to be expensive and of limited effectiveness (Pransky, 1991). A variety of efficacious and effective primary and secondary programs have been described (Minkovitz et al., 2003; Olds et al., 1998; Simpson et al., 2001), suggesting that early identification leading to early intervention may reduce the risk of long-term negative consequences among children who receive services.

Though a preventive public health approach is called for by many, several barriers to prevention are posed by social and systemic attitudes underlying service philosophies (Kauffman, 1999). These barriers include the fragmentation of approaches between involved systems, as well as complexities in screening within developmental and social

environmental contexts. Across systems, inconsistent quality in the implementation of EPSDT social and emotional screening (U.S. GAO, 2003), in Child Find efforts (U.S. Department of Education, 2003), and in implementation of promising evidence-based practice models (U.S. DHHS, 1999) stymie the widespread application of primary and secondary preventive approaches to care. Similarly, the gatekeeper system in schools (U.S. DHHS, 1999) and varying interpretations of federal eligibility requirements (Forness & Knitzer, 1992) lead to the under-identification of children in need of behavioral services in the school setting, including Head Start (Sinclair et al., 1993). This combination of factors highlights the need for resources beyond the public school system for improved early identification of this social and public health problem.

The primary care setting has been identified as a promising venue for efforts toward early identification of externalizing behavior problems among very young children (AHRQ, 2002). Social acceptance of the concept of externalizing behavior problems as a disorder requiring professional treatment has resulted in increased recognition of the potential role of pediatric primary care clinicians in this arena (see Conrad & Schneider, 1980; Pawluch, 1983; Tuchman, 1996). Attention to behavioral issues in pediatric practice has increased over the past 30 years, but substantial under-identification of children with psychosocial issues has been a persistent problem in the primary care setting (E. J. Costello, 1986; E. J. Costello & Edelbrock, 1985; E. J. Costello et al., 1988; Lavigne et al., 1993). One recent survey found that most pediatricians prefer to use informal methods to screen for psychosocial issues in pediatric patients (AAP, 2000), highlighting shortcomings in assessment approaches used in practice.

In response to these problems with early identification in primary care, the use of reliable and valid standardized instruments has been promoted (Halfon et al., 2003). While parent-report checklists may suffer certain limitations (Kagan et al., 2002), their use as screening, rather than diagnostic, instruments may be valuable. The PSC-17 (Gardner et al., 1999) and the BPI (Peterson & Zill, 1986; Zill, 1990) represent instruments which may be especially appropriate for use in the primary care setting due to their brevity and ease of scoring. However, these tools are not without shortcomings. Questions have been raised regarding reliability and validity with younger children, minority children, and children of lower SES, in particular. Both the PSC-17 and the BPI have primarily been used with children ages 4 and above, and their utility in screening children below the age of 4 has not been established. Lower sensitivity and specificity of the full-length PSC with children under age 6 have been described (Navon et al., 2001), while examinations of the psychometric properties of the BPI have not attended to age as a factor of interest (Parcel & Menaghan, 1988; Zill, 1985, 1990). Reports of differing screening results by sex (Jellinek et al., 1999; Parcel & Menaghan, 1988; Reijneveld et al., 2006; Thun-Hohenstein & Herzog, 2008), race (Jutte et al., 2003; Simonian & Tarnowski, 2001; Simonian et al., 1991; Spencer et al., 2005), and SES (Jellinek et al., 1995; Jellinek et al., 1999; Reijneveld et al., 2006) have not been followed with item-level analyses of possible bias in these instruments. Further, available psychometric evaluations of these instruments have relied solely on CTT-based methods, which are limited in their capacity to assess measurement performance independent of the particular samples included in investigations (Nunnally & Bernstein, 1994).

The use of reliable and valid standardized instruments, such as the PSC-17 and the BPI, for the early identification of very young children with externalizing behavior problems in pediatric primary care settings could improve primary and secondary prevention efforts in this arena. However, what is known regarding the quality of measurement provided by these instruments is limited by the shortcomings of CTT-based methods and assumptions. Modern methods of investigating the quality of measurement provided by these instruments for preschool-aged children of differing sex, race, and SES could yield valuable information regarding their utility in preventive practice efforts.

Two research questions arise directly from this discussion:

Research Question 1: What is the quality (i.e., precision and utility) of measurement provided by items in the PSC-17 and BPI measuring externalizing behavior problems in very young children?

Research Question 2: Do any items measuring externalizing behavior problems in the PSC-17 and BPI exhibit measurement bias with very young children by (a) sex, (b) race, or (c) SES?

Answers to these research questions could guide use of the PSC-17 and BPI in both practice and research, and could also facilitate the selection of a set of items which are most informative and least biased when used with very young children in diverse pediatric primary care populations. Given the limitations inherent in CTT-based scale evaluation, this study provided a more comprehensive and informative assessment of the

quality of these measures using a powerful modern measurement theory: item response theory.

CHAPTER III

ITEM RESPONSE THEORY: APPLICABILITY TO MEASUREMENT OF EXTERNALIZING BEHAVIOR PROBLEMS IN VERY YOUNG CHILDREN

In social work research and practice, CTT is the predominant framework espoused by developers and users of measurement instruments (Nugent & Hankins, 1992). While major advances in measurement theory have been made over the past 50 years in other fields (i.e., education and psychology), most researchers in the health and social sciences are only in the beginning stages of exploring the potential utility of modern psychometrics. One analytical approach which could improve measurement in both practice and research applications is item response theory (IRT), a modern measurement theory developed in the 1950s and 1960s (Hambleton & Swaminathan, 1985). IRT is a revolutionary approach which enables applications and outcomes impossible to achieve using traditional psychometric methods. Its concept of *parameter invariance*, in which findings are independent of the particular sample with which analyses were conducted, sets it apart from CTT methods. In brief, IRT aims to enable “test-free” and “sample-free” measurement, akin to the physical measurements of weight and height in which attributes of the sample or measurement tool used are independent of the invariance of the underlying metric (Lord, 1980).

As discussed in Chapter II, the use of reliable and valid standardized instruments has been suggested for improving screening for externalizing behavior problems in very

young children in the pediatric primary care setting. Two instruments which may be suitable for this use are the PSC-17 (Gardner et al., 1999) and the BPI (Peterson & Zill, 1986; Zill, 1990). However, concerns exist regarding performance of these instruments in measuring behavior problems in very young children, especially among children differing by sex, race, and SES. The measurement qualities of both scales have been evaluated solely by CTT-based methods, limiting conclusions about their properties to situational use with samples similar to those investigated in psychometric studies. Further, each scale was initially developed using CTT-based methods (Gardner et al., 1999; Zill, 1990), known to encompass certain theoretical and practical limitations (Hambleton & Swaminathan, 1985). The application of IRT-based methods to evaluate the quality of measurement provided by these scales with the population of interest promises exciting new possibilities for understanding and improving tools for screening in diverse pediatric primary care settings, improving screening accuracy and reducing unjust disparities.

In this chapter, an overview is presented of the applicability of methods based on IRT to improve the measurement of externalizing behavior problems in very young children. First, the limitations inherent in traditional CTT-based methods are discussed. The development of IRT in response to these limitations is described. A brief overview of the concepts and model-based measurement approaches of IRT is provided, with descriptions of the various models used for items with dichotomous and polytomous response options. The assumptions and limitations associated with IRT methods are summarized, as well as the theoretical advantages offered by IRT over CTT. A detailed discussion is provided of the application of one model particularly salient to items with polytomous ordinal rating scales, such as those constituting the PSC-17 and the BPI.

Useful products of the fitting of IRT models, including item and test information functions, are described as they apply to scale evaluation and to items comprising the PSC-17 and BPI in particular. Similarly, IRT methods designed to detect item-level bias are reviewed, as they apply to concerns raised in Chapter II regarding performance of the PSC-17 and the BPI with specific groups of children. Finally, hypotheses based upon the two research questions concluding Chapter II are posed, related to the application of IRT methods of scale evaluation to items in the PSC-17 and the BPI.

Limitations of Classical Test Theory

To appreciate the advantages offered by IRT, it is important to understand the limitations inherent in CTT. Problems associated with the development, scoring, and evaluation of scales using CTT methods include the sample-dependent and test-dependent nature of all estimates, such as scores and coefficients of reliability and validity (Crocker & Algina, 1986; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). In other words, these attributes of a scale are inextricably related to (a) the particular set of items included in the scale, and (b) the particular sample of respondents with whom the estimates were determined. In practical terms, this implies that any changes to the content or combination of items included in a scale, as well as any use of a scale with a group not represented by the sample with whom the scale was normed, will have unknown effects on the quality of measurement offered by the scale (Crocker & Algina, 1986; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985).

Other limitations associated with scales developed with CTT methods include (a) the likelihood of restriction of range (i.e., ceiling and floor effects), due to the redundancy of items tapping similar levels of the latent construct in order to increase reliability

(Hambleton et al., 1991; Ware, 2003); (b) the untenable assumption that the SEM is constant across all levels of the latent construct (Hambleton & Swaminathan, 1985; Nugent, 2005); (c) the prohibitive length of scales for screening or for use in the fast-paced primary care setting, again due to efforts to increase reliability (Hambleton & Swaminathan, 1985; Ware, 2003); (d) the preponderance of multiple scales developed to measure the same variables, resulting in lack of comparability across studies and applications (Ware, 2003); (e) possible bias introduced by the use of one form repeatedly over time in longitudinal studies or in clinical monitoring, stemming from difficulties encountered in developing truly parallel forms (Crocker & Algina, 1986); and (f) inability to identify item-level bias when confounded with true group differences in levels of the latent construct (Hambleton & Swaminathan, 1985).

Implications of these limitations are troubling when considering what is known about the measurement performance of instruments evaluated only with CTT methods. For example, estimates of reliability and validity for the PSC-17 and BPI are not absolute, but can change depending on the composition of measured samples. Such disparities in estimates can be observed in reports of differing performance of each instrument among various groups of children (e.g., Jellinek et al., 1995; Jutte et al., 2003; Navon et al., 2001; Spencer et al., 2005). Based on available psychometric evaluations of these scales, no information is available regarding (a) the precision of measurement offered at various levels of externalizing behavior problems; (b) the range of externalizing behavior problems adequately measured by these scales—in particular, whether they are capable of detecting sub-clinical levels of behavior problems reliably, as needed for effective primary and secondary prevention efforts (E. J. Costello & Shugart,

1992); nor (c) biases in item performance between groups, when controlling for level of externalizing behavior problems.

Development of Item Response Theory

IRT was developed in response to the limitations of CTT with respect to scale development and evaluation. In particular, the sample- and test-dependent properties of CTT indices of measurement performance prompted attention to the theoretical and practical shortcomings of traditional psychometric theory. Though the foundations of modern measurement theory can be traced to Thurstone's conceptualization of latent traits in the 1920s, the development of IRT is generally attributed to pioneering work by Lord (1953). Throughout the 1950s and 1960s, researchers including Lord, Birnbaum, Rasch, and Wright introduced logistic latent trait models and methods for model parameter estimation, highlighting potential applications of IRT methods in education, industry, and psychology (Bock, 1997; Hambleton & Swaminathan, 1985).

By the 1980s, advances in computer technology and software expanded the accessibility of IRT methods to researchers and practitioners in measurement-oriented fields (Hambleton & Jones, 1993). Expectations for psychometric instruments which could not be assured via CTT methods—such as mandating a stable measurement unit across all levels of the latent construct and expecting that items within a scale should be exchangeable—led IRT developers away from classical measurement assumptions (Hambleton & Swaminathan, 1985). However, since then, application of IRT methods has remained centered in education, industry, and psychology, with other social science fields lagging behind (Hays, Morales, & Reise, 2000; Ware, 2003). Recent demonstrations of IRT methods have been conducted with measures of health-related

outcomes, including symptom severity (Bjorner, Kosinski, & Ware, 2003a; Bjorner, Kosinski, & Ware, 2003b) and health-related quality of life (Hays et al., 2000; Ware, 2003). Only rarely, however, have IRT analyses been applied to measures of child behavior problems (Gumpel, 1998; Lambert et al., 2003; Stevenson, Thompson, & Sonuga-Barke, 1996).

IRT: Modern Model-Based Measurement

While CTT incorporates concepts of test score, true score, and error score into applications that generally focus on test-level functioning of instruments, IRT is a distinct statistical theory specifying and incorporating item-level, test-level, and respondent-level properties into measurement development and evaluation (Hambleton & Jones, 1993). IRT differs significantly from classical methods due to its mathematical modeling framework, which allows linking of item characteristics to respondent level of the underlying unobservable (or *latent*) construct of interest (e.g., externalizing behavior problems in children). At its core, IRT consists of a set of generalized linear models capable of modeling the probability of a particular response to an item based upon (a) the level of the latent construct possessed by the respondent, and (b) certain stable characteristics of the item (Embretson & Reise, 2000). The basic premise is that for a given item measuring a latent construct, the probability of item endorsement should rise as a respondent's level of the latent construct increases. In addition, the stable characteristics of the item are not dependent on the particular sample or other items used in assessing measurement performance, due to the concept of *parameter invariance* (described later in this chapter).

The simplest application of IRT modeling is to *dichotomous* items, characterized in knowledge-based testing as *correct* or *incorrect*, and in trait- or symptom-type testing as *endorsed* or *not endorsed* (Embretson & Reise, 2000). A more complex application is to *polytomous* items, including items with either ordered (e.g., Likert-type) or unordered (e.g., nominal multiple choice) response options (Hambleton & Swaminathan, 1985). For most types of items, the probability of a randomly selected individual's response to an item is generally represented as a nonlinear monotonic function of the level of the latent construct, taking into consideration certain item characteristics. This relationship is graphically represented by the *item characteristic curve* (ICC) for dichotomous items, and by *option characteristic curves* (OCCs) for polytomous items (sometimes referred to as category response curves; Hambleton & Swaminathan, 1985).

Models for Dichotomous Items

An example of a basic logistic IRT model is one frequently applied with dichotomous items: the two-parameter logistic model (2PL), originally proposed by Birnbaum (1968). Mathematical representation of the 2PL is presented to illustrate several common features of IRT models:

$$P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}} \quad (i = 1, 2, \dots, n). \quad (1)$$

Equation 1 provides the 2PL *item characteristic function* for a dichotomously-scored item (i.e., correct/incorrect, true/false, etc.). In the 2PL, $P_i(\theta)$ represents the probability of the endorsement of item i , given a particular level of the latent construct, represented by θ . The mathematical constant e is the base of the natural logarithm, which is

approximately equal to 2.71828. The mathematical constant D represents a scaling factor, generally set to 1.7 in order to minimize differences between the 2PL and a two-parameter model derived from the cumulative normal distribution (a more computationally complex approach to IRT modeling; Hambleton & Swaminathan, 1985). The difficulty of item i is represented by b_i , and refers to the level of the latent construct (θ) at which the probability of item endorsement is equal to .5 (i.e., the level of the latent construct at which 50% of respondents would endorse item i). The discrimination of item i is represented by a_i , a value proportional to the slope of the tangent line to the item characteristic function at its steepest point, which is at its difficulty level (i.e., at b_i). Steeper slope of the curve at this point is associated with greater precision of discrimination between respondents at similar levels of θ ; flatter slopes suggest weaker item capacity to discriminate between respondents.

When the item characteristic function depicted in Equation 1 is graphed for a single item i with particular item parameters b_i and a_i over a range of values of θ , the result is the ICC, illustrated for a hypothetical dichotomous item in Figure 1. Several features of the ICC graph are notable. First, the range of the latent construct θ depicted on the x -axis generally extends from -3.0 to +3.0, where θ is arbitrarily scaled to have a mean of 0 and standard deviation of 1.0. The probability of item endorsement asymptotically approaches 0 at decreasing levels of θ and 1.0 at increasing levels of θ . The monotonically increasing s-shaped curve is characteristic of logistic functions. Note that the difficulty level (b_i) of a given item is located at the level of θ at which the probability of endorsement is .5, and the tangent line with slope proportional to the item's discrimination parameter (a_i) touches the ICC at the point at which θ is equal to b_i . For

the illustrated hypothetical item with difficulty level $b_i = 0.25$ and discrimination level $a_i = 1.0$, the probability of item endorsement for respondents with a latent construct level 1 standard deviation below the mean is approximately .2; for respondents with latent construct levels 2 standard deviations above the mean, the probability of endorsement is approximately .85; and for respondents at the mean latent construct level, the probability of endorsement is approximately .45.

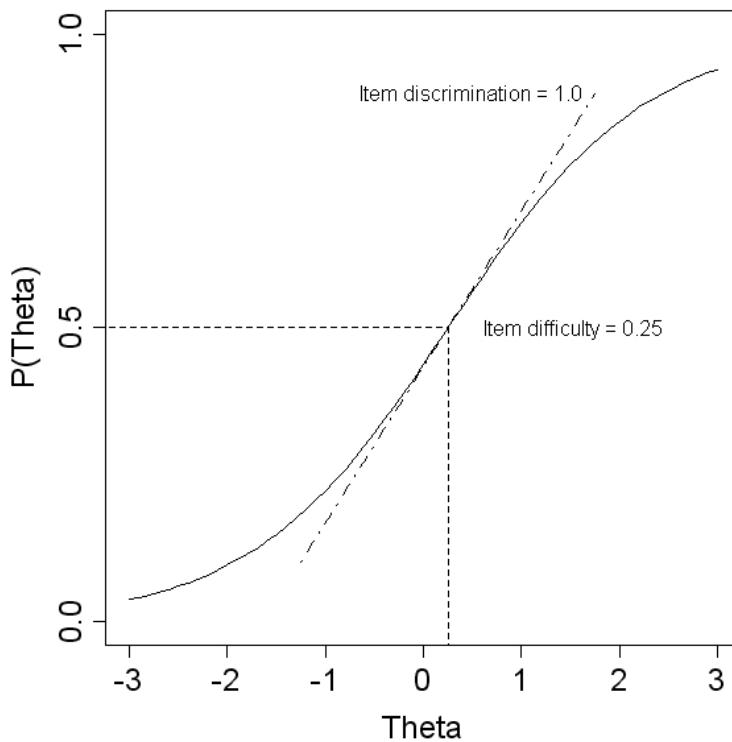


Figure 1. Item characteristic curve (ICC) for a hypothetical item in the two-parameter logistic model (2PL; $b_i = 0.25$, $a_i = 1.0$).

In a two-parameter model such as the 2PL, both item parameters can vary between items. Thus, items can differ in their difficulty levels (i.e., placement along the x-axis), as well as in their discrimination levels (i.e., maximum steepness of slope). One-

parameter models exist which constrain the discrimination levels of all items to be equal (usually at $a = 1.0$), and these models are often referred to as *Rasch* models (for their developer; Hambleton & Swaminathan, 1985). In addition, three-parameter models are possible, which include an additional parameter (c_i) allowing the lower asymptote of the ICC to be greater than 0 (Hambleton & Swaminathan, 1985); however, these models are more applicable to knowledge-testing items, in which the probability of guessing correctly increases the base level of probability of a correct response.

In Figure 2, three hypothetical ICCs in the 2PL are depicted with differing difficulty and discrimination parameters. If one were interested in items which accurately measured respondents with levels of the latent construct between 1 and 2 standard deviations above the mean, of these three items, Item 3 would appear to be most helpful. For Item 1 ($b_1 = -2.0$, $a_1 = 1.2$), all respondents with θ levels above the mean would be nearly equally likely to endorse the item. For Item 2 ($b_2 = 0.0$, $a_2 = 0.5$), the probabilities of item endorsement change very slowly for the θ levels of interest, obscuring distinctions between respondents at similarly, but not identically, high levels of θ . In contrast, Item 3 ($b_3 = 1.5$, $a_3 = 1.8$) can discriminate well between respondents at the desired levels of θ . This example illustrates the applicability of IRT modeling to the identification and selection of items with specific, desired measurement properties.

Models for Polytomous Items

IRT models are not limited to dichotomous items, such as those illustrated above. For polytomous items, multiple functions are possible for each item, each representing the probability of choosing a particular categorical response option given a specific level of the latent construct (Hambleton & Swaminathan, 1985). In a polytomous item, the

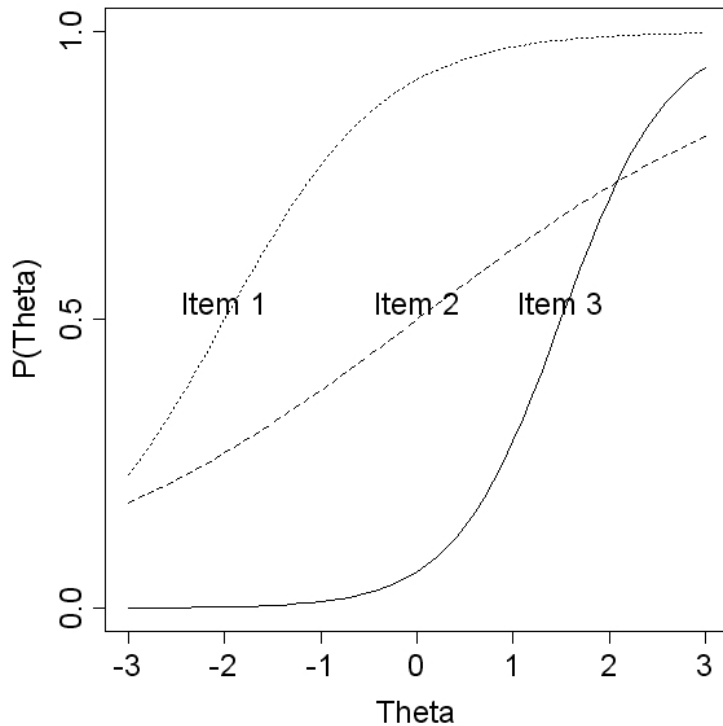


Figure 2. Three hypothetical item characteristic curves (ICCs) with differing item parameters ($b_1 = -2.0$, $a_1 = 1.2$; $b_2 = 0.0$, $a_2 = 0.5$; and $b_3 = 1.5$, $a_3 = 1.8$).

likelihood of choosing a particular response option is a function of the levels of the latent construct; if response options are ordered, respondents with higher levels of θ are more likely to choose higher response options. These *option characteristic functions* can be graphically represented by OCCs, just as dichotomous item characteristic functions are depicted by ICCs. The points of intersection of the OCCs for a polytomous item indicate the levels of θ at which shifts in response options are most likely for that item. Points of intersection of OCCs are referred to as *difficulty thresholds*; there are always one fewer thresholds than response options (Hambleton & Swaminathan, 1985).

Many IRT models have been developed which can be applied to items with multiple nominal response categories (Bock, 1972), as well as to items with Likert-type

polytomous ratings for which response options are ordered. These include the graded response model (Samejima, 1969), the partial credit model (Masters, 1982), the ordinal model (Thissen & Steinberg, 1986), and the generalized partial credit model (Muraki, 1992). Later in this chapter, the graded response model (Samejima, 1969), which is applicable to items with polytomous ordered response options (as found in the PSC-17 and the BPI), is described and illustrated in detail.

This discussion of dichotomous and polytomous IRT models highlights the potential utility of IRT models in evaluating the quality of measurement provided by a given item at specific levels of a latent construct. The process of estimating item parameters using a given set of data capturing response patterns to a set of items is referred to as *item calibration*, and the resulting parameter estimates provide valuable information for item and scale evaluation (Hambleton & Swaminathan, 1985). However, a stringent set of assumptions underlies the application of these models.

Assumptions and Limitations of IRT

While IRT offers powerful item-level analysis techniques, its utility is tempered by its strong underlying assumptions. These assumptions can pose limitations to the practical implementation of IRT methods. First, IRT models assume *unidimensionality* of scales. Second, they assume *local independence*. Finally, each IRT model assumes a particular *trace line function* for an item. Each of these key assumptions is described in detail, followed by a brief discussion of practical limitations associated with implementation of IRT methods.

Unidimensionality

The IRT assumption of unidimensionality of scales is common to many CTT applications as well, and thus is familiar to many developers and users of psychometric instruments. While the concept of unidimensionality is not without controversy (see McDonald, 1981), the assumption generally refers to the notion that a set of items measures a single latent construct (Lord & Novick, 1968). In IRT applications, this assumption is often clarified to specify that the data obtained in response to a set of test items are “unidimensional enough,” in that one dominant latent construct accounts for patterns of participant responses (Hambleton & Swaminathan, 1985; Reckase, 1979). While progress has been made toward enabling application of IRT methods to multidimensional scales, the vast majority of research efforts, as well as the availability of software to implement IRT analyses, have been focused on models assuming unidimensionality (Hambleton & Swaminathan, 1985). No consensus exists as to the best way to evaluate whether this assumption has been met in a particular application, but approaches such as exploratory factor analysis appropriate for categorical data have been proposed and used (Hambleton & Swaminathan, 1985; C. K. Parsons & Hulin, 1982).

Local Independence

The IRT assumption of local independence is related to that of unidimensionality, but incorporates subtle differences. Specifically, local independence refers to the requirement that the latent construct fully explains all relationships between items (Hambleton & Swaminathan, 1985). This means that given a respondent’s level of θ , the conditional probability of obtaining any pair of scores for any pair of items is the product of the probabilities for the individual items (Yen, 1993):

$$P(X_1 = x_1, X_2 = x_2 | \theta) = P(X_1 = x_1 | \theta)P(X_2 = x_2 | \theta). \quad (2)$$

Equation 2 specifies that when holding θ constant, any selected pair of items should be statistically independent of one another—thus, the measured latent construct fully accounts for any relationships between items. Evaluation of local independence is often overlooked in applications of IRT, but it is crucial in the derivation of IRT models, and violations can result in problems with model misfit (Yen, 1993) and reliability (Wainer & Thissen, 1996). As with unidimensionality, no consensus exists as to the best way to assess whether a set of data meets the assumption of local independence, but several promising approaches have been highlighted (Hambleton & Swaminathan, 1985; Wainer & Thissen, 1996; Yen, 1993).

Trace Line Functions

Finally, each IRT model assumes a specific trace line function, or ICC (Hambleton & Swaminathan, 1985). For example, the 2PL assumes that the function of Equation 1 accurately represents the ICC for dichotomous items which can be adequately characterized by two parameters. Similarly, Rasch models, the 3PL, and polytomous IRT models all assume particular trace line functions to represent response data. To check this assumption for a given model, no universally accepted test of model fit exists (Hambleton & Swaminathan, 1985). However, as for the other assumptions of IRT, a variety of model-fit testing approaches have been proposed (Bock & Aitkin, 1981; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Hambleton & Swaminathan, 1985), and the importance of attempting to check this assumption is highlighted in Maldonado and Greenland's (1993) discussion of the implications of model misfit for parameter

interpretation. Visual inspection of nonparametric graphs of trace line functions is one simple approach for investigating the appropriateness of specific trace line functions for a given dataset (Ramsay, 2000).

Practical Limitations

In addition to the limitations associated with IRT methods based upon their strong underlying assumptions, a practical difficulty posed by these analyses is the large sample size required for many applications (DeVellis, 2003; Hambleton & Swaminathan, 1985). Most IRT applications require at least 500 respondents, the minimum number recommended for obtaining stable parameter estimates (Reise & Yu, 1990). For applications beyond item calibration, such as analyses comparing item performance between groups, a minimum sample size of 250 per group has been proposed (Bolt, 2002), though many authors have analyzed datasets with fewer participants. Unfortunately, power analysis techniques for determining required sample size to achieve stability in parameter estimates have not been developed in IRT. Sample size guidelines, such as those reviewed above, also depend on item characteristics and model choice, and this area of study has been identified as one requiring much more theoretical and practical progress (Fayers, 2004; Tay-Lim & Harwell, 1997).

Finally, the lack of familiarity with IRT outside the fields of education and psychology can pose problems for researchers using these methods. IRT requires a conceptual leap from the familiar ground of CTT, and its primary applications have been centered in educational, industrial, and psychological testing (DeVellis, 2003; Hambleton et al., 1991). Expertise and familiarity with IRT methods are lacking in many areas of the social sciences, including social work. Exceptions include burgeoning efforts in the area

of health-related quality of life (Ware, 2003), as well as several recent social work applications (DeRoos & Allen-Meares, 1992; Nugent, 2003, 2005, 2006; Nugent & Hankins, 1992). Despite the challenges and limitations associated with the implementation of IRT approaches, these methods hold much promise as additional tools for the advancement and improvement of measurement of psychosocial constructs in many fields—including the measurement of externalizing behavior problems among very young children in primary care settings.

Theoretical Advantages of IRT

When the assumptions underlying IRT methods can be met and practical barriers to their implementation can be overcome, IRT models have several significant advantages over traditional CTT approaches to scale development and evaluation. The following discussion of the theoretical advantages of IRT models further explicates the rationale for applying such models to answer the research questions posed in Chapter II regarding measurement of externalizing behavior problems in very young children.

When an IRT model can be appropriately fit to data capturing patterns of responses to items, there are three primary theoretical advantages of IRT, as summarized by Hambleton and Swaminathan (1985): First, item parameter estimates (i.e., indices of item difficulty and discrimination) are statistically independent of the particular sample of respondents drawn to examine the item. Second, an estimate of a particular respondent's score (i.e., level of the latent construct) is statistically independent of the particular set of items used for measurement. Finally, a statistic indicating the degree of precision of a score estimate is available, which is free to vary depending on the level of the latent construct and the characteristics of the item(s) used for measurement.

These advantages are based upon the theoretical property of *parameter invariance*, in which item parameter estimates do not depend mathematically on the distribution of the latent construct in the sample of interest (Hambleton & Swaminathan, 1985). This is in direct contrast to the situation found in CTT analyses. For example, the CTT-based method of estimating the difficulty level of an item is to assess the proportion of respondents who answered the item correctly (for knowledge-based items) or who endorsed the item (for symptom-measurement items; Lord & Novick, 1968). If a dichotomous item measuring a particular externalizing behavior were administered to a sample of parents of children with very few externalizing behavior problems, very few parents would endorse the item. This low proportion, in CTT, would suggest that the item is *difficult*, in that very few respondents endorse it. If the same item were administered to a sample of parents of children diagnosed with ODD or CD, however, a much higher proportion would likely endorse the item, suggesting that the item is *easy*. Thus, assessment of item difficulty using CTT-methods leads to estimates which are dependent on the distribution of the latent construct of interest within the particular sample studied.

In IRT, estimates of item parameters (i.e., item difficulty and item discrimination) are theoretically invariant when the IRT model adequately represents the data (Hambleton & Swaminathan, 1985). Therefore, an item's parameter estimates do not change when the item is administered to groups with different distributions of the latent construct: An easy item is always easy, and a difficult item is always difficult. While it has been shown that CTT-based estimation of item difficulty can be very stable when normal distributions of the latent construct are present in different samples, CTT-based item discrimination estimates (i.e., the ability of the item to differentiate between

respondents at different levels of the latent construct) are extremely variable, while IRT item discrimination estimates are quite stable (MacDonald & Paunonen, 2002). The accuracy of item parameter estimates obtained using IRT methods has also been demonstrated to be superior to that of CTT methods in Monte Carlo simulation studies (MacDonald & Paunonen, 2002).

Practical benefits of IRT methods are related to the theoretical property of item parameter invariance. For example, in all IRT models, estimates of respondents' levels of the latent construct do not vary with the characteristics of the sample measured; measurement error is conditional upon level of the latent construct; and item content is specifically targeted at a particular range of the latent construct (Embretson & Reise, 2000). Further, IRT methods allow more comprehensive evaluation of item characteristics, as the item comprises the basic unit of analysis. Because the scale metric is not dependent on a specific set of items but rather on the level of the latent construct, considerable flexibility in scale development is possible, allowing (a) variations in item response formats within the same scale, (b) reduction of number of items required, and (c) variation in combinations of items presented (Ware, 2003). Adaptive testing (e.g., computerized adaptive tests such as the GRE and others offered by Educational Testing Service) is one exciting application of these possibilities (Ware, 2003). In addition, sets of items can be tailored to measure specific ranges of the latent construct of interest, either broadly or narrowly, eliminating ceiling and floor effects if desired (Hambleton & Swaminathan, 1985). Construction of truly parallel measures consisting of entirely different sets of items is possible (Hambleton & Swaminathan, 1985; Ware, 2003). Precision of measurement can be adjusted as needed for different intended uses of sets of

items, including research and clinical applications at either group or individual levels (Crocker & Algina, 1986). Assessment of group differences in item and scale functioning can be accomplished, and identification of problematic individual response patterns is possible (Hambleton & Swaminathan, 1985). Finally, different measures addressing the same latent construct can be equated, placing scores on the same metric and allowing cross-instrument comparisons (Ware, 2003).

Potential benefits conferred by the theoretical advantages of IRT analyses could be significant if applied to evaluate instruments measuring externalizing behavior problems in very young children. In evaluating the items measuring externalizing behaviors comprising subscales of the PSC-17 and the BPI, the use of IRT methodology to calibrate each item with a large sample of preschool-aged children in primary care clinics should yield theoretically invariant item parameter estimates. Such parameter estimates would allow comparisons of the level of externalizing behavior problems in 3 to 5 year old children best measured by each item. Items could be identified which measure above average (including clinical and sub-clinical) levels of externalizing behavior problems, potentially facilitating early identification of children in need of further assessment. Additional analyses relevant to the research questions posed in Chapter II are discussed later in this chapter.

The Graded Response Model

When item responses can be ordered into more than two categories along a continuum—as seen in the Likert-type item response options comprising the PSC-17 and BPI—Samejima's (1969) graded response model (GRM) is an appropriate polytomous IRT model to use. While dichotomization of polytomous item responses is often

conducted to allow fitting of simpler IRT models (e.g., the Rasch or 2PL models), preservation of the ordinal nature of item responses provides more psychometric information than yielded by dichotomous models with comparable item parameters (Agresti, 2002; Samejima, 1977). The two-parameter polytomous GRM is an extension of the 2PL described earlier in this chapter (Reise & Yu, 1990), and, as with the 2PL, use of the logistic function in the model is generally preferred to the cumulative normal function to preserve computational efficiency.

In this overview of the GRM, hypothetical items with three ordered response options are used, to illustrate how the model may be applied to items found in the PSC-17 and the BPI. Each hypothetical item, therefore, has $K = 3$ ordered response options, coded $k = 0, 1,$ and 2 . Parallel to the manner in which ICCs are estimated for dichotomous items, in the GRM, option characteristic curves (OCCs) must be estimated for each response option in an item (Samejima, 1969). The OCCs are derived from the 2PL presented in Equation 1, by estimating item responses as one of the two dichotomies captured in the response thresholds: (a) response option 0 versus options 1 and 2; and (b) response options 0 and 1 versus option 2. The probability of endorsing option 0 or higher is defined as 1.0, and the probability of endorsing an option higher than option 2 is defined as 0, since no option higher than 2 is provided (Samejima, 1969). The *option characteristic functions* associated with a hypothetical item with $K = 3$ ordered response options ($k = 0, 1, 2$) are as follows:

$$P(k_i | \theta) = \begin{cases} 1 - \frac{e^{Da_i(\theta-b_{i,1})}}{1 + e^{Da_i(\theta-b_{i,1})}} & \text{if } k_i = 0 \\ \frac{e^{Da_i(\theta-b_{i,1})}}{1 + e^{Da_i(\theta-b_{i,1})}} - \frac{e^{Da_i(\theta-b_{i,2})}}{1 + e^{Da_i(\theta-b_{i,2})}} & \text{if } k_i = 1 \\ \frac{e^{Da_i(\theta-b_{i,2})}}{1 + e^{Da_i(\theta-b_{i,2})}} & \text{if } k_i = 2. \end{cases} \quad (3)$$

In Equation 3, $P(k_i | \theta)$ represents the probability of the endorsement of response option k for item i , given a particular level of the latent construct, represented by θ . The mathematical constants e and D are identical to their values in the 2PL. The parameter $b_{i,1}$ represents the value of θ at the threshold (i.e., intersection) between response options 0 and 1, and the parameter $b_{i,2}$ represents the value of θ at the threshold between response options 1 and 2. In the two-parameter polytomous GRM, item discrimination is assumed to be constant within item response options, but may vary between items; thus, the parameter a_i refers to the discrimination level of all response options of item i .

A graphical illustration of the GRM for a hypothetical item with three ordered response options clarifies the interpretation of the option characteristic functions presented above. Figure 3 is a graph of the probabilities of endorsement of the response options associated with one such item, conditional on the level of the latent construct being measured. Note that for the lowest levels of θ , the most likely response option to be selected is option 0 (often labeled as *not at all* or *never* in symptom-type items). As the level of θ increases, the probability that option 0 will be selected gradually lowers, until at $\theta = -0.5$, the probability of endorsing option 0 is equal to the probability of endorsing option 1 (often labeled *sometimes* or *somewhat true* in symptom-type items). This level

of θ is equal to the parameter $b_{i,1}$, the threshold between response options 0 and 1. As the level of θ increases, the probability of endorsement of option 1 initially increases but gradually begins to decrease, until at $\theta = 1.5$, the probability of endorsing option 1 is equal to the probability of endorsing option 2 (often labeled *always* or *often true* in symptom-type items). This level of θ is equal to the parameter $b_{i,2}$, the threshold between response options 1 and 2. From this level of θ on, the probability of endorsement of option 2 increases, asymptotically approaching 1.0.

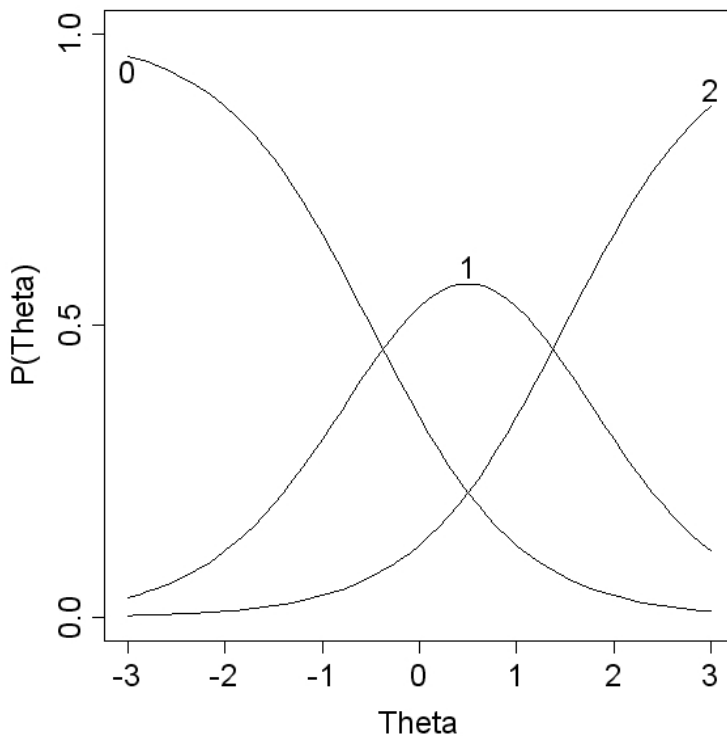


Figure 3. Graded response model option characteristic curves (OCCs) for a hypothetical item with three response options ($a_i = 1.3$, $b_{i,1} = -0.5$, $b_{i,2} = 1.5$).

Model-fitting and estimation of the item parameters $b_{i,k}$ and a_i can be efficiently achieved using marginal maximum likelihood estimation procedures with an expectation

maximization algorithm (Bock & Aitkin, 1981). These procedures are available in a Windows-based software program, MULTILOG 7.03 (Thissen, Chen, & Bock, 2003), which has been demonstrated to recover stable and accurate parameters using the GRM (Reise & Yu, 1990). Once items with ordinal response options, such as those comprising the PSC-17 and BPI, are calibrated using these techniques, they can be described in terms of the levels of θ measured by their response options, as well as in terms of their abilities to discriminate between respondents at different levels of θ . In addition, the item parameter estimates obtained by fitting the GRM can be used for at least two other valuable purposes: (a) estimating item and test information and precision, and (b) investigating biases in item performance between different groups. These applications, described below, can be used to answer the research questions posed in Chapter II regarding measurement of externalizing behavior problems in very young children.

Information and Precision

A very useful feature of IRT models is the evaluation of the test information offered by a set of items, as well as of the item information offered by individual items (Hambleton & Swaminathan, 1985). The test information function, $I(\theta)$, is defined for a particular set of items at each point along the continuum of the latent construct. It is influenced by both the number of items included in the scale as well as the discrimination parameters of each item (Hambleton & Swaminathan, 1985).

$$I(\theta) = -E \left\{ \frac{\partial^2}{\partial \theta^2} [\ln L(\mathbf{u}|\theta)] \right\} = \sum_{i=1}^n \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad (i = 1, 2, \dots, n). \quad (4)$$

In Equation 4, the test information function at a given value of θ is defined as the negative expected value of the second derivative of the log-likelihood function, $\ln L(\mathbf{u} | \theta)$. This value is equivalent to the sum of the individual item information functions of the n items comprising the test, with each item contributing independently to the total test information function. Each item i 's information function can be derived by squaring the first derivative of the probability function of item i at θ and dividing it by the product of the probability function of item i at θ and the quantity 1.0 minus that probability. The first derivative of the probability function of item i is equal to the slope of the function at each point along the θ continuum. Thus, information is affected by item discrimination, in that higher discrimination values are associated with higher levels of item and test information (Hambleton & Swaminathan, 1985). In addition, item characteristic functions with smaller variance—captured by $P_i(\theta)Q_i(\theta)$, in the denominator of Equation 4—also yield higher levels of information (Hambleton & Swaminathan, 1985).

It is noteworthy that the information function $I(\theta)$ is equal to the reciprocal of the variance of the maximum likelihood estimator of the level of the latent construct, θ . In IRT models, the analogous concept to the CTT standard error of measurement (SEM) is the standard error of estimation (SEE), computed as the square root of the variance of the maximum likelihood estimator of θ at each point along the latent construct continuum (Hambleton & Swaminathan, 1985). Thus,

$$SEE = \frac{1}{\sqrt{I(\theta)}}. \quad (5)$$

As suggested in Equation 5, at levels of θ where test and item information are high, the SEE is low, and measurement precision is high. Conversely, at levels of θ where test and item information are low, the SEE is high, and measurement precision is low. The advantage of the IRT approach to assessing SEE is that, unlike with CTT methods of estimating SEM, the use of a sample-dependent reliability coefficient is avoided, as is the use of an *average* estimate of standard error across all values of the latent construct (Hambleton & Swaminathan, 1985). Thus, as highlighted by Nugent (2005), estimation of SEE with IRT methods allows the standard error to vary across different levels of the latent construct, accounting for differences in the quality of measurement of particular levels of θ offered by different items and sets of items. The IRT item and test information functions supplant the CTT-based concepts of reliability and SEM, allowing item-level or test-level precision of measurement to be assessed independently at any desired level of θ (Hambleton & Swaminathan, 1985).

Test and item information functions can be represented graphically for ease of comparison (see Figures 4 and 5). Figure 4 portrays the item information functions associated with the same three hypothetical 2PL items with ICCs depicted in Figure 2. Figure 5 presents the test information function of the same three items as a set. The relationship between item and test information can be seen in these figures, and the potential utility of these functions for selecting items to improve measurement precision at particular ranges of the latent construct is illustrated.

As can be observed by comparing Figures 4 and 5, the test information function is the sum of the individual item information functions. Of the three hypothetical items, Figure 4 demonstrates that Item 3, with the highest discrimination parameter, offers the

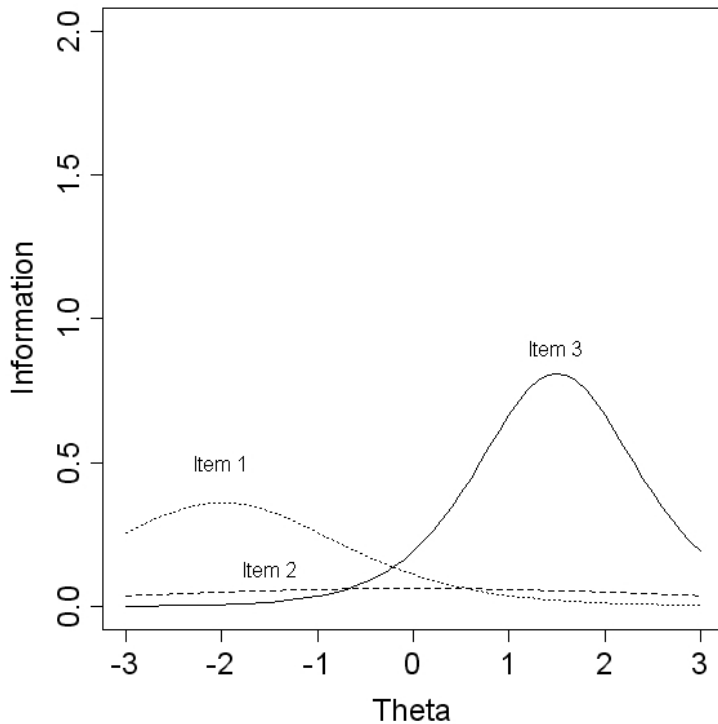


Figure 4. Item information functions for three hypothetical items ($b_1 = -2.0$, $a_1 = 1.2$; $b_2 = 0.0$, $a_2 = 0.5$; and $b_3 = 1.5$, $a_3 = 1.8$).

most information for measurement of the latent construct, followed by Item 1. Item 2, with a very low discrimination parameter, provides very little information for the measurement of any range of the latent construct of interest. When these three items are combined into a scale, Figure 5 illustrates the levels of the latent construct at which the set of items provides the most information. While these examples are for dichotomous items in the 2PL, analogous information functions can be generated for polytomous items, as represented in the GRM.

The importance of test information, item information, and SEE is in their application to the assessment of the quality of measurement offered by specific items. For example, in the PSC-17 and the BPI, multiple items are thought to measure externalizing

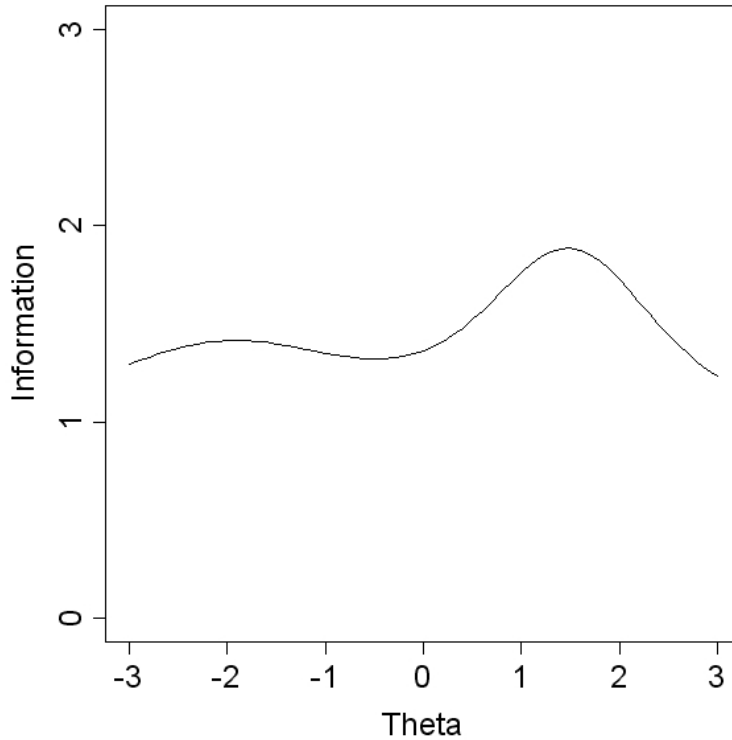


Figure 5. Test information function for a set of three hypothetical items ($b_1 = -2.0$, $a_1 = 1.2$; $b_2 = 0.0$, $a_2 = 0.5$; and $b_3 = 1.5$, $a_3 = 1.8$).

behavior problems. When administered to caregivers of children between the ages of 3 and 5, the amount of information provided by these items for measurement of various levels of externalizing behavior problems is unknown. Similarly, data regarding the degree of standard error associated with the items' measurement performance at different levels of externalizing behavior problems are not available. It is possible that some items included in these scales may be much more informative and precise than others, when applied to this age group. By fitting the GRM to data capturing response patterns to these items among the population of interest, item and test information functions can be obtained to allow identification of highly informative and precise items within a specific range of externalizing behavior problems. Use of highly informative items targeting

appropriate levels of externalizing behavior problems could improve efforts toward early identification of very young at-risk children in the primary care setting. Visual inspection of graphical representations of item and test information functions can provide valuable information in selection of items which best measure the desired levels of externalizing behavior problems in very young children.

Differential Item Functioning

Item bias is a serious concern in the measurement of any psychosocial latent construct (Teresi, 2001). Such bias is defined as the tendency of an item to perform differently with different groups of respondents. For example, in an educational test such as the Scholastic Aptitude Test (SAT), a quantitative item which appears to be harder for African-American students than it is for white students, when math ability is held constant, would be a biased item. Item bias, then, is a systematic error in the measurement process (Osterlind, 1983). Such error in screening instruments can lead to broader social injustices, incompatible with equitable primary and secondary prevention efforts.

The systematic measurement error introduced when biased items are included in a scale threatens the scale's construct validity (Osterlind, 1983), and thus the validity of conclusions drawn from respondent scores on that scale. In a scale developed to measure externalizing behavior problems in children, for example, unidimensionality of the scale is assumed—the only latent construct influencing item responses should be externalizing behavior problems. If other, unknown latent constructs associated with group membership (e.g., by sex, race, or SES) also influence responses to certain items, then the items in question do not purely measure the latent construct of interest. The result is an

incongruence of meaning of scores across the groups, and a scale which measures externalizing behavior problems more accurately in some groups than in others.

Efforts to detect item bias with CTT methods suffer the same shortcomings associated with CTT approaches to many scale development and evaluation tasks: sample-dependence and test-dependence (Osterlind, 1983). Often, CTT approaches to investigating item bias involve comparisons of traditional item difficulty (i.e., proportion of respondents endorsing an item) and item discrimination (i.e., item-total biserial correlations) between groups of interest (Lord & Novick, 1968). As reviewed in the discussion of the limitations associated with CTT methods, any differences in these indices assessed by traditional approaches cannot be extricated from differences in the distribution of the latent construct present in the group samples. In CTT, item difficulty and discrimination indices always depend on the sample of respondents with whom they are generated (Lord & Novick, 1968).

Modern measurement theory, however, offers advanced methods to detect item bias, referred to as *differential item functioning* (DIF) in the IRT framework. A clear definition of an *unbiased* item from an IRT perspective asserts that all individuals with equal levels of the latent construct of interest should have equal probabilities of item endorsement, regardless of group membership (Hambleton & Swaminathan, 1985). The IRT relationship between the probability of item endorsement and the level of the latent construct, combined with the theoretical property of invariance of item parameters, thus necessitates that, “A test item is unbiased if the item characteristic curves across different subgroups are identical” (Hambleton & Swaminathan, 1985, p. 285). When ICCs (or

OCCs, for polytomous items) across groups differ, then the item under investigation exhibits DIF.

Investigations of DIF have revealed two types of item bias: *uniform* and *non-uniform* DIF (Camilli & Shepard, 1994). Uniform DIF involves differences in item responses between groups which are consistent across all levels of the latent construct. For example, an item may consistently be easier for one group as compared to another. In non-uniform DIF, however, interactions occur between level of the latent construct, item response, and group. In this case, an item which is easier for one group at low levels of the latent construct may be harder for that group at high levels of the latent construct. Either type of DIF results in a biased item which could influence the construct validity of the scale in which it is included. While no consensus exists regarding the best way to test for DIF, several authors have reviewed and evaluated a myriad of methods (e.g., Bolt, 2002; Camilli & Shepard, 1994; Hambleton & Swaminathan, 1985; Teresi, 2001), and methods specific to GRM applications for detecting uniform and non-uniform DIF have been empirically supported (e.g., Crane, van Belle, & Larson, 2004; Maldonado & Greenland, 1993). Visual inspection of ICCs (or OCCs) generated separately for each group also provides an intuitive approach for screening items for DIF (Hambleton & Swaminathan, 1985).

As discussed in Chapter II, questions regarding the quality (i.e., precision and utility) of measurement of externalizing behavior problems among children in various subgroups remain unanswered. Differences in scores yielded by the PSC-17 and the BPI between groups differing by sex (Jellinek et al., 1999; Parcel & Menaghan, 1988), race (Jutte et al., 2003; Simonian & Tarnowski, 2001; Simonian et al., 1991; Spencer et al.,

2005), and SES (Jellinek et al., 1995; Jellinek et al., 1999; Simonian et al., 1991) have not been investigated at the level of item bias. Thus, the quality of measurement offered by these instruments for particular sociodemographic groups is unknown. While total score differences between groups may be due to true variations in levels of externalizing behavior problems, unbiased measurement by items comprising scales must be assured in order to avoid both over-identification and under-identification of children in need of further assessment and services (Spencer et al., 2005). DIF detection using IRT methods could help achieve this goal and improve efforts toward early identification of very young children with externalizing behavior problems in primary care settings.

Summary and Hypotheses

The applicability of IRT-based methods to improve the measurement of externalizing behavior problems in very young children is promising. Limitations inherent in CTT-based methods, such as the sample-dependent and test-dependent nature of traditional indices of reliability and item characteristics, pose problems for the utility and interpretation of scores obtained by measures developed and evaluated only with traditional methods (Hambleton & Swaminathan, 1985). Other shortcomings associated with CTT-based methods include their inability to provide information regarding (a) the precision of measurement offered at various levels of externalizing behavior problems; (b) the range of externalizing behavior problems adequately measured by these scales; and (c) biases in item performance between groups, when controlling for level of externalizing behavior problems. Given these limitations, substantial questions remain regarding the quality of measurement provided by the PSC-17 and BPI for externalizing behavior problems in very young children.

The theoretical and practical development of IRT was in response to limitations of CTT-based methods. The mathematical modeling of the probability of item (or response option) endorsement as a function of the level of the latent construct being measured constitutes the underlying theory common to all IRT methods. Definitions and derivations of the item difficulty and discrimination parameters estimated via IRT models, presented earlier in this chapter, guide the interpretation of such estimates for both dichotomous and polytomous items. The potential utility of IRT models for evaluation of the quality of measurement provided by a given item at specific levels of a latent construct is great. Item calibration via IRT model-fitting offers descriptive item-level information useful for several types of item and scale evaluation.

While capable of analyses beyond those possible with CTT-based methods, IRT approaches share a stringent set of assumptions which can pose limitations to their utility (Hambleton & Swaminathan, 1985). The key assumptions of unidimensionality, local independence, and specific item trace line functions must be evaluated in IRT applications, but several challenges complicate assessment of these requirements. Further, practical limitations associated with the implementation IRT methods are possible, including demands for large sample sizes and lack of familiarity with IRT outside of the fields of education and psychology (DeVellis, 2003).

The theoretical advantages offered by IRT include (a) the statistical independence of item parameter estimates from the particular sample of respondents; (b) the statistical independence of the estimate of a particular respondent's score from the particular set of items used for measurement; and (c) the availability of a statistic indicating the degree of precision of a score estimate, free to vary depending on level of the latent construct and

characteristics of the item in question (Hambleton & Swaminathan, 1985). These advantages stem from the theoretical property of item parameter invariance, in which item parameter estimates are independent of the distribution of the latent construct in the sample of interest (Hambleton & Swaminathan, 1985). These theoretical advantages suggest potential uses of IRT for scale improvement efforts with the PSC-17 and the BPI.

The GRM (Samejima, 1969) is an IRT model appropriate for use with the types of items included in the PSC-17 and the BPI. Use of this model for calibrating items with polytomous ordinal rating scales would allow comparisons among various items' parameter estimates, potentially revealing differences in (a) the levels of externalizing behavior measured by items' response options, as well as (b) items' abilities to discriminate between children at different levels of externalizing behavior problems.

Two additional exciting applications of IRT models for scale evaluation are possible: the use of item and test information functions, and methods for the detection of DIF. Item and test information functions obtained by fitting the GRM to response data from PSC-17 and BPI items would allow identification of highly informative items to measure precisely a defined range of externalizing behavior problems (Hambleton & Swaminathan, 1985). Items biased between groups of interest (i.e., groups differing by sex, race, or SES) could be identified using IRT DIF detection analyses (Crane, van Belle, & Larson, 2004; Teresi, 2001). Together, these analyses would allow the identification of a set of items offering the most precise, informative, and unbiased measurement of externalizing behavior problems among very young children in primary care settings.

Study Hypotheses

Information reviewed in this chapter supports the use of IRT-based methods to evaluate the items of the PSC-17 and BPI intended to measure externalizing behavior problems. Accurate measurement with preschool-aged children is crucial to efforts to improve early identification in primary care settings, an important component of effective and efficient primary and secondary prevention. To assess the accuracy and utility of the measurement provided by relevant items in the PSC-17 and BPI with the target population, specific hypotheses were developed to answer the two research questions posed in Chapter II. Though direct statistical tests for each hypothesis are not available in the IRT framework, decisions regarding the relative value of each item are possible, and an approach to such decisions is described in Chapter IV.

Research Question 1: What is the quality (i.e., precision and utility) of measurement provided by items in the PSC-17 and BPI measuring externalizing behavior problems in very young children?

Hypothesis 1.1: Items in the externalizing subscales of the PSC-17 and BPI will have differing difficulty and discrimination parameter estimates, when administered to primary caregivers of children between the ages of 3 and 5 and analyzed using the GRM.

Hypothesis 1.2: Items in the externalizing subscales of the PSC-17 and BPI will have differing item information functions (and hence differing degrees of precision at various levels of the latent construct), when administered to primary caregivers of children between the ages of 3 and 5 and analyzed using the GRM.

Research Question 2: Do any items measuring externalizing behavior problems in the PSC-17 and BPI exhibit measurement bias with very young children by (a) sex, (b) race, or (c) SES?

Hypothesis 2.1: Items in the externalizing subscales of the PSC-17 and BPI will exhibit differing degrees of bias between groups of male and female children, when administered to primary caregivers of children between the ages of 3 and 5 and analyzed using the GRM.

Hypothesis 2.2: Items in the externalizing subscales of the PSC-17 and BPI will exhibit differing degrees of bias between groups of white and minority children, when administered to primary caregivers of children between the ages of 3 and 5 and analyzed using the GRM.

Hypothesis 2.3: Items in the externalizing subscales of the PSC-17 and BPI will exhibit differing degrees of bias between groups of children of low versus high SES, when administered to primary caregivers of children between the ages of 3 and 5 and analyzed using the GRM.

CHAPTER IV

METHOD

The description of study methods is divided into five sections, summarizing (a) participants, (b) procedures, (c) measures, (d) data analyses, and (e) integration of findings. Regarding participants, details are provided describing sample size, recruitment sites, and inclusion and exclusion criteria. The procedure is outlined regarding data collection and gift card drawing specifications. Three sets of measures are delineated, including the externalizing subscale of the PSC-17 (Gardner et al., 1999); the headstrong, antisocial, and peer problems subscales of the BPI (Peterson & Zill, 1986; Zill, 1990); and a sociodemographic questionnaire developed by the author. Three stages of data analysis are described in detail, including descriptive analyses, assessment of CTT-based psychometric properties of the subscales, and analyses based on IRT. The description of IRT analyses includes testing of IRT assumptions, fitting of the IRT GRM, and detection of DIF. Power and sample size considerations are also addressed. Finally, a brief summary is provided of an approach for integrating findings from the three phases of data analysis.

Participants

Recruitment Sites

Caregivers of preschool-aged children ($N = 900$) were recruited to participate from four pediatric primary care settings: University Child Health Specialists (UCHS),

University Child Health Specialists South (UCHS-S), Children and Youth Project (C&Y), and Oldham County Pediatrics (OCP). These clinics are university-affiliated and serve a large population of diverse children and families, with patient demographics and clinic capacity as follows:

UCHS and UCHS-S. UCHS is the primary practice arm of the University of Louisville Pediatrics Department. This clinic is located in an urban center and provides ambulatory care and resident training in all aspects of pediatric practice, primarily serving a low SES, minority population. UCHS-S is a satellite clinic located on a hospital campus in a suburban setting, serving a combination of urban, suburban, and rural families of diverse races and SES. Together, the clinics serve over 7,000 infants, children, and adolescents, with nearly 20,000 outpatient visits per year.

C&Y. Located on the University of Louisville health sciences campus, the C&Y clinic provides comprehensive health care to inner city, high-risk infants and children from birth through 17 years of age, a population identified with substantial medical and socioeconomic challenges. C&Y serves over 8,000 active patients with an average of 72 medical visits per day.

OCP. This pediatric primary care practice is located in a setting serving primarily rural and suburban families. The clinic is affiliated with the University of Louisville, offering resident training rotations in pediatrics. The population served by OCP is mostly white and of higher SES than seen in the other sites. OCP provides general pediatric primary care services to approximately 6,000 children, with an average of 85 outpatient visits per day.

To maximize diversity among participants, targeted recruitment was equally divided among the four sites. Patient demographic characteristics among these sites vary considerably, so enrollment from all four clinics was needed to provide adequate group sizes for analyses.

Inclusion and Exclusion Criteria

Those eligible for the study were primary caregivers of at least one child between the ages of 3 and 5 years. In addition, participants were required to be age 18 or older and able to understand and read English, in order to complete the informed consent process and respond to the survey. All participants were in attendance at pediatric primary care appointments at one of the four designated clinics, but it was not necessary for the identified child to be present (i.e., a caregiver may have been attending an appointment for an older or younger sibling, but would still be invited to complete the survey regarding the child in the target age range).

Exclusion criteria included a) already having responded to the survey regarding another child in the home and b) presenting for an emergency appointment. Emergency appointments included those at which urgent care was being provided (e.g., breathing treatments, injuries), but did not include standard sick appointments (e.g., sore throats, low-grade fevers). These exclusion criteria were identical to those used in the largest study to date on screening for child mental health issues in primary care settings (Jellinek et al., 1999); preserved the independence of individual responses; and reflected the population of very young children seeking non-emergency primary care services.

Procedure

All study procedures were approved by the University of Louisville Institutional Review Board (IRB). The informed consent process included a preamble consent format provided at the beginning of the study questionnaire (see Appendix B). A complete waiver of HIPAA authorization was granted, in order to facilitate screening of potential participants in the clinics for eligibility and willingness to enroll. No HIPAA authorization forms were necessary since no personal health information was collected.

Data Collection

For this cross-sectional survey study, a convenience sample of caregivers of preschool-aged children from each clinic was selected. Recruitment was conducted at various days and times of the week over the course of 8 months. During this phase, the researcher or IRB-approved assistant approached all available caregivers in the waiting areas of each clinic to determine study eligibility and request participation (see Appendix C for the eligibility checklist and script used to screen and invite eligible participants). Potential participants were informed of the chance to win one of five gift cards valued at \$100 each at the conclusion of all data collection. If an approached individual met the eligibility requirements and was willing to participate, following informed consent procedures, the participant was asked to complete the survey while in a quiet area of the waiting room. Any participant with more than one child in the target age range was asked to select the child with the most recent birthday as the one to consider while responding to items. While this sampling procedure may have resulted in certain study limitations, the number of participants required and the goal of recruiting caregivers who were actually attending pediatric appointments made random sampling procedures untenable.

Questionnaires were color-coded by clinic and numbered to track participant responses. No personally identifiable information was recorded on questionnaires. The researcher or assistant was available during survey completion to answer questions as needed. If participants required more time to complete the questionnaires, they were able to bring them to the exam rooms during appointments and/or complete them in the waiting room following the appointments. Once a survey was completed, the researcher or assistant collected it from the participant and reviewed its contents for completeness, requesting responses to missed items if necessary. Upon completion of the survey, each participant was invited to provide contact information and seal it in an envelope to enter the gift card drawing. Completed questionnaires were removed from the clinic by the researcher or assistant at the end of each day of data collection. Sealed envelopes containing contact information were stored separately from questionnaires, and no information existed linking contact information and questionnaires. Responses from completed questionnaires were entered into an SPSS 15.0 (SPSS, 2007) database by the researcher. All questionnaires were double-entered, allowing data-cleaning to maintain integrity of the data.

Gift Card Drawing

The gift card drawing was expected to increase the response rate to an acceptable level. The drawing for five winners of gift cards was held at the conclusion of data collection, when five sealed envelopes were randomly selected from the total number submitted by all participants. Gift cards were delivered by registered mail. Contact information for all participants was subsequently destroyed.

Measures

The study survey included three components: two commonly-used scales for measuring child behavior problems and one sociodemographic questionnaire. The order of the behavior rating scales was counterbalanced in the distributed surveys to avoid response set or order bias.

Pediatric Symptom Checklist-17 (PSC-17)

The PSC-17 (Gardner et al., 1999), a brief version of the PSC (Jellinek et al., 1986), was developed for use in pediatric clinics to screen children for early identification of possible psychosocial problems. This instrument consists of 17 items on which caregivers rate their child using a 3-point Likert-type scale (0 = *never*, 1 = *sometimes*, 2 = *often*). Traditional CTT-based scoring involves summing item responses for a total score, where higher scores indicate higher levels of dysfunction. Possible scores on the entire instrument range from 0 to 34.

Investigations of the factor structure of the PSC-17 suggested that the instrument can be separated into three subscales, including an externalizing subscale (7 items), an internalizing subscale (5 items), and an attention subscale (5 items; Gardner et al., 1999). Due to the brevity of the scale, the entire set of 17 items was administered (see Appendix D), though IRT analyses focused solely on the externalizing subscale. Possible scores on the externalizing subscale ranged from 0 to 14, and the PSC-17 authors recommended a cut-score of 7 on this subscale to indicate need for further assessment (Gardner et al., 1999). See Appendix E for PSC-17 scoring instructions.

Psychometric properties of the PSC-17 reported by its authors (Gardner et al., 1999) included high levels of internal consistency for the full scale (Cronbach's $\alpha = .89$),

as well as for the externalizing subscale of interest (Cronbach's $\alpha = .83$). When used to identify children with externalizing behavior problems, the externalizing subscale reportedly exhibited a sensitivity of 77% and specificity of 80%, as compared to classifications of problems yielded by the parent-completed Iowa-Connors aggression subscale (Loney & Milich, 1982), a modification of the Conner's Teacher Rating Scale (Conners, 1969) with an author-reported internal consistency reliability coefficient of .86. The authors of the PSC-17 estimated the time required to complete all 17 items to be approximately 4 minutes (Gardner et al., 1999).

Behavior Problems Index (BPI)

The BPI (Peterson & Zill, 1986; Zill, 1990) was developed for use in national longitudinal surveys to measure behavioral problems in children and was standardized on a random sample of 6,000 children (P. C. Baker, Keck, Mott, & Quinlan, 1993). Its items were derived from the CBCL (Achenbach & Edelbrock, 1981) in order to provide a shorter scale appropriate for use in survey research. The BPI consists of 28 items (26 for preschool-aged children) on which caregivers rate their child using a 3-point Likert-type scale (0 = *not true*, 1 = *sometimes true*, 2 = *often true*). Total scores are computed via traditional CTT-based methods, by summing item responses. Higher scores indicate higher levels of dysfunction. Possible scores on the entire instrument range from 0 to 52 for preschool-aged children.

The BPI has six subscales, measuring headstrong behaviors, antisocial behaviors, peer problems, anxious/depressed mood, hyperactivity, and immature dependency (Zill, 1990). Three of these subscales are relevant to the measurement of externalizing behavior problems: the headstrong subscale (5 items), the antisocial subscale (4 items), and the

peer problems subscale (2 of 3 items are relevant to externalizing behaviors). These three subscales (minus 1 internalizing peer problems item) were combined into a BPI externalizing subscale consisting of 11 items for the purposes of this study. This measure of externalizing behaviors was similar to a 15 item measure developed by Cooksey, Menaghan, and Jekielek (1997) from the BPI, but excluded 2 items targeting impulsive and inattentive behaviors (associated primarily with ADHD) and 2 items measuring school behavior (not included in the preschool version of the BPI).

While clinical cut-scores have not been set for this instrument, most authors use the raw subscale scores associated with the 90th percentile for a given age group as indicative of clinically significant behavior problems (Zill, 1990). These scores are based on dichotomized coding of each item, in which a response of *not true* is coded 0, and a response of either *sometimes true* or *often true* is coded 1. For 4 and 5 year old children, dichotomized raw scores associated with the 90th percentile are 5 for the headstrong subscale, 3 for the antisocial subscale, and 1 on the original 3-item peer problems subscale (Center for Human Resource Research, 2000). Due to the brevity of the scale, the full set of 26 items appropriate for preschool-aged children was administered (see Appendix F), though IRT analyses focused solely on the externalizing subscale. Possible scores on the BPI externalizing subscale ranged from 0 to 22. See Appendix G for scoring instructions for the BPI.

Psychometric properties of the BPI reported in previous studies included high estimates of internal consistency for the full instrument (Cronbach's α ranging from .89 to .90; Gortmaker et al., 1990; Zill, 1990), and lower estimates for individual subscales (Cronbach's α ranging from .63 to .75; Gortmaker et al., 1990; Spencer et al., 2005). Test-

retest reliability has not been reported for the full instrument nor for subscales of interest. Only one published study has evaluated construct validity (Spencer et al., 2005), concluding based on factor analysis that the BPI appeared valid for measurement of behavior problems primarily among white children. The estimated time required to complete all 26 items was approximately 6 minutes.

Sociodemographic Questionnaire

The final section of the study survey included several items measuring relevant sociodemographic characteristics of the sample (see Appendix H).

Caregiver characteristics. Participants were asked to report their own demographic characteristics. These included age, sex, race, level of household income, years of education completed, and relationship to the child.

Child characteristics. Participants were also asked to report the child's age (in years), sex, race, family structure (i.e., one- or two-parent household, caregiver other than parent, and so on), number of siblings in the home, type of health insurance, and number of hours per week spent in daycare and preschool. Child SES was operationalized by creating an index combining responses regarding household income level, caregiver education level, and child's type of health insurance. First, the ordinal-level variable of household income was recoded into three categories with roughly equal frequencies: \$0-\$20,000; \$20,001-\$50,000; and \$50,001 and higher. As a point of reference, according to the U.S. Department of Agriculture Economic Research Service (2008), the 2004 median household income level in Kentucky was approximately \$37,000. Second, the ordinal-level variable of caregiver education level was also recoded into three categories: less than high school, high school, and more than high school. Similarly, child's type of

health insurance was recoded into three categories: none; public (i.e., Medicaid, K-CHIP, and Medicare); and private. All three variables were coded 0, 1, and 2, with higher values assigned to higher levels of income, education, and insurance (private was rated as higher than public, which was rated higher than none). Next, the recoded income, education, and insurance variables were summed for each participant, yielding possible SES index scores from 0 to 6. Finally, index scores from 0 to 2 were classified as low SES; those from 3 to 4 were classified as medium SES; and those from 5 to 6 were classified as high SES. Crosstabulations of these classifications with the original data for household income, caregiver education, and child health insurance suggested that the SES designations were appropriate.

Child sex, race, and SES were independent variables in bivariate and IRT analyses. All other sociodemographic variables measured were used for sample description only.

Other relevant factors. Finally, for descriptive purposes, participants were asked to respond to several questions regarding the reason for the appointment on the day of recruitment (i.e., illness of child, well child check-up, sibling's appointment) and history of behavioral concerns (i.e., whether the parent believed the child has behavior problems; whether the child had received services from a mental health or behavioral provider; whether the parent had ever expressed concern to the child's physician regarding behavioral problems; whether the physician had ever expressed concern to the parent regarding child behavioral problems; and whether any other adults had ever expressed concern to the parent regarding child behavioral problems).

Estimated time to complete the entire sociodemographic questionnaire was approximately 3 minutes. Thus, the estimated total time required for completion of the entire study survey was approximately 14 minutes, though most participants finished more quickly.

Data Analysis

The focus of this study was on item-level analyses of the individual items included in the externalizing subscales of the PSC-17 and the BPI, with the purpose of identifying a set of the most informative and unbiased items suitable for screening in pediatric primary care of preschool-aged children for externalizing behavior problems. In order to accomplish this goal, data analysis involved three stages: (a) descriptive analyses, (b) CTT-based analyses, and (c) IRT-based analyses.

Descriptive Analyses

Simple descriptive statistics were employed to describe the sample. Summary measures of demographic characteristics of children and caregivers, as well as of other factors from the sociodemographic questionnaire (e.g., proportion of children who have received mental health services, reasons for clinic visit on day of recruitment, and so on), were obtained. Descriptive analyses were conducted using SPSS 15.0 for Windows (SPSS, 2007).

CTT-based Analyses

The psychometric properties of the PSC-17 and the BPI have been previously studied and reported in the literature. To determine whether the performance of these instruments with the study sample was comparable to previous investigations, several analyses were conducted based upon CTT methods. These included (a) assessment of

distributional properties of each externalizing subscale, including mean scores, standard deviation, skewness, kurtosis, frequency and patterns of missing data, and possible ceiling or floor effects; (b) assessment of the internal consistency of each externalizing subscale, as represented by Cronbach's α ; (c) computation of inter-item correlations and item-test correlations within each externalizing subscale; (d) investigation of item performance in terms of drop in externalizing subscale coefficient alpha when the item is removed; (e) exploration of concurrent and known groups validity of each externalizing subscale; and (f) bivariate analyses exploring relationships between externalizing subscale scores and child sex, race (white versus minority), and SES, respectively. SPSS 15.0 for Windows (SPSS, 2007) was used for all analyses based on CTT.

IRT-based Analyses

The crux of this study lay with analysis methods based on IRT. As the study results were intended to facilitate the combination of items from each subscale into a single measure of externalizing behavior, IRT analyses required both subscales to be analyzed together so that patterns of responses to all items could be considered. In the remainder of the text, the 18 investigated items are referred to as the *combined externalizing subscale*. In order to identify which items performed best in measuring externalizing behavior problems in very young children, several steps were necessary. These included (a) testing IRT model assumptions; (b) fitting an IRT model to the data to obtain item parameter estimates, item information functions, and subscale information functions; and (c) testing each item for differential item functioning (DIF) between identified groups of interest. Results of these analyses guided selection of a set of items

most appropriate for measurement of the latent construct of interest in the target population.

Evaluation of IRT model assumptions. Testing the strong assumptions inherent in IRT methods was key to appropriate use of this approach. As discussed in Chapter III, three primary assumptions are made for all IRT models: unidimensionality, local independence, and specific trace line functions. There are several available methods for testing each assumption, but no consensus exists regarding the best approach. Thus, when possible, more than one test of an assumption was conducted. Any discrepancies in findings were weighed in terms of the IRT literature and interpreted accordingly.

To assess unidimensionality of the combined externalizing subscale, the results of the CTT methods of assessing item performance and internal consistency were considered. However, these methods alone are insufficient to demonstrate unidimensionality, as high levels of internal consistency are possible with multidimensional data (McDonald, 1981). As an additional step in testing unidimensionality, exploratory factor analysis was conducted on the combined externalizing subscale. Reckase (1979) and others have recommended that in order for a scale to be “unidimensional enough” for IRT analyses, the first factor should be dominant and account for at least 20% of the variance. Magnitudes of eigenvalues for additional factors, correlations among factors, and strength of factor loadings, in combination with visual evaluation of a scree plot (Bjorner et al., 2003a, 2003b) and indicators of internal consistency, were reviewed to assess the dimensionality of the combined externalizing subscale. SPSS 15.0 for Windows (SPSS, 2007) was used for these analyses.

As described in Chapter III, the assumption of local independence refers to the independence of item responses in a scale conditional upon the level of the latent trait. In other words, once the level of externalizing behavior is controlled, item responses should be statistically independent from one another (Steinberg & Thissen, 1996; Wainer & Thissen, 1996; Yen, 1993). Assessment of local independence involved examination of the residual correlation matrix from the exploratory factor analysis. According to Reeve and colleagues (2007), violations of local independence are suggested when $|r| \geq .20$. The residual correlation matrix was generated using SPSS 15.0 for Windows (SPSS, 2007).

The assumption of specific trace line functions, as applied to the GRM, refers to the requirement that the probability of selecting progressively higher item response options increases with higher levels of the latent trait, and never decreases. This assumption was assessed by fitting a non-parametric IRT model to the data from the combined subscales and graphing the results, in effect generating a trace line from the observations. The trace lines for each item were then visually inspected for the expected form. This assessment was conducted using TestGraf software (Ramsay, 2000).

Fitting the IRT model. Samejima's (1969) GRM was fit to the observed data for the combined externalizing subscale in order to obtain item parameter estimates. The two-parameter polytomous GRM was used. This model provided a flexible framework in which both the difficulty threshold parameters and the item discrimination parameters were free to vary between items, while item discrimination was constrained to be constant within each item, thus reducing the number of estimated parameters and simplifying computations and interpretation. MULTILOG 7.03 (Thissen et al., 2003) software was used to fit the GRM and obtain item parameter estimates.

No consensus exists regarding methods of determining goodness of fit for the GRM (Hambleton & Swaminathan, 1985); most existing approaches utilize χ^2 statistics, which are problematic when there are many response patterns and large samples. Thus, a combination of graphical and statistical procedures was used to investigate model fit, using the MODFIT computer program (Stark, 2002). The sample was split evenly by odd versus even identification numbers into calibration and cross-validation samples, allowing the GRM to be fit to the calibration sample while the cross-validation sample was retained for assessment of goodness-of-fit. Model fit was evaluated graphically using sets of fit plots for each item, depicting (a) the model-derived OCCs estimated from the calibration sample, and (b) the empirical OCCs observed in the cross-validation sample. Close correspondence between the sets of curves for each item would suggest good model-data fit.

In addition, a statistical procedure based on χ^2 tests recommended by Drasgow and colleagues (1995) was used to compare expected counts from the model-fitting with the calibration sample to observed counts from the cross-validation sample. Drasgow and colleagues recommended that to alleviate the problems of sensitivity to sample size typically encountered with χ^2 statistics, as well as their insensitivity to certain types of misfit, ratios of χ^2 divided by degrees of freedom (*df*) be calculated for single items, pairs of items, and triples of items. Items with similar types of misfit would be expected to generate large χ^2/df ratios; per Drasgow and colleagues, ratios ≤ 3 generally indicate good fit. While Drasgow and colleagues suggested adjusting large samples sizes (i.e., $N > 3,000$) down to 3,000 in order to enable comparisons across studies with different sample sizes, the current study already had a sample size below that criterion. Thus, unadjusted

χ^2/df ratios were used. Drasgow and colleagues also cautioned that all IRT models will be misspecified to some degree, resulting in frequent rejection of models based upon statistical tests of significance. To remedy this situation, combining statistical and graphical procedures can be helpful in interpreting model fit. In general, when model assumptions are deemed to be satisfactorily met and graphical assessment appears satisfactory, interpretation of the model is useful even when statistical tests suggest poor model fit (C. K. Parsons & Hulin, 1982).

Item parameter estimates, OCCs, and item information curves were inspected and compared. Marginal maximum likelihood estimation with an expectation maximization algorithm was used to estimate item parameters (Bock & Aitkin, 1981). Item information curves graphically represented the amount of information offered by an item at various levels of the measured construct. In other words, item information curves demonstrated at what levels of externalizing behavior problems each item was most informative. Precision of measurement was highest where information was greatest; conversely, SEE was highest where information was lowest (Hambleton & Swaminathan, 1985). Visual inspection of item information curves allowed identification of items which offered the greatest amount of precision (i.e., reliability) of measurement at various locations along the continuum of externalizing behavior problems for this population.

For the purposes of early identification and screening in a prevention context, it was important to identify combinations of items that were informative at clinical as well as sub-clinical ranges of externalizing behavior problems (E. J. Costello & Shugart, 1992). The test information function, generated by summing individual item information functions, was plotted for visual inspection of the precision of measurement at various

levels of the latent construct provided by a given set of items (Hambleton & Swaminathan, 1985). In IRT model-fitting, the theta metric (i.e., the scale of measurement of externalizing behavior problems) is generally standardized with a mean of 0 and standard deviation of 1.0. Item difficulty parameters are measured on the same metric as theta. Thus, item difficulty parameters, and their graphical location on plots of OCCs and item information curves, were directly relatable to levels of the latent construct, interpretable in relation to the mean (i.e., 1.5 standard deviations above the mean, 0.8 standard deviations below the mean, etc.). This allowed clear interpretation of the utility of each item for measurement at various levels of theta (Hambleton & Swaminathan, 1985). The investigations of item parameter estimates, OCCs, and item and test information functions were conducted using MULTILOG 7.03 (Thissen et al., 2003) software. Additional graphing results were produced with PlotIRT (C. D. Hill & Langer, 2007) freeware using the R platform (R Development Core Team, 2007).

Detection of DIF. There are many approaches to assessing DIF, and again, no consensus exists as to the best method (Bolt, 2002; Teresi, 2001). For this reason, two approaches were employed in this study, and the results from each method were compared. To answer the research questions and test the hypotheses delineated in Chapters II and III, comparisons of interest were for male children versus female children; for white children versus minority children; and for low SES children versus medium/high SES children. Operational definitions of these grouping variables were provided in the Measures section, above.

The first method for DIF detection was the IRT-based likelihood ratio test (IRT-LR; Thissen, 2001). This test was used to identify both uniform (i.e., in item difficulty

parameters) and non-uniform (i.e., in item discrimination parameters) DIF in items yielding different parameter estimates for reference and focal groups. The IRT-LR method involved several steps. First, for each set of group comparisons, an iterative process allowed identification of an anchor set of items exhibiting no DIF (Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006). Next, several hierarchically nested models were fit for each remaining item one at a time, comparing item parameter differences between groups to those seen in the no-DIF anchor items. Initially, all parameters were allowed to vary between groups; in subsequent nested models, discrimination and difficulty parameters were constrained to be equal between groups. A likelihood ratio test statistic (G^2) was generated for each model, distributed as χ^2 with degrees of freedom equal to the difference in the number of parameters estimated in each nested model (Thissen, 2001). A Bonferroni correction for multiple comparisons was used to preserve overall α at .05. Statistically significant values of G^2 indicated improved model fit when a given item's parameters could vary between groups. This situation was suggestive of DIF. The IRT-LR DIF detection method was implemented using IRTLRDIF freeware (Thissen, 2001).

The second method for detecting DIF was the ordinal logistic regression approach (OLR), developed by Crane and colleagues (2004). For this approach, three nested ordinal logistic regression models were fit for each item, predicting the cumulative logit of item responses: (a) a model including the main effect of theta (i.e., level of externalizing behavior problems) as the only predictor; (b) a model including the main effects of theta as well as group membership as predictors; and (c) a model including main effects of theta and group membership, as well as the interaction effect between

theta and group membership, as predictors. Theta was represented by participants' IRT scores on the combined externalizing subscale, computed using MULTILOG 7.03 (Thissen et al., 2003) software. Statistical significance of the main effect of group and/or the interaction effect between group and theta were indicative of uniform and non-uniform DIF, respectively. A Bonferroni correction for multiple comparisons was used to preserve overall α at .05. Statistically significant uniform DIF suggested that group membership was predictive of item responses while controlling for level of externalizing behavior problems. Statistically significant non-uniform DIF suggested that item responses were predicted by an interaction between group membership and level of externalizing behavior problems, captured by different item discrimination parameters for each group (Crane et al., 2004). The OLR analyses were completed using SPSS 15.0 (SPSS, 2007) software, based upon the approach designed for the DIFdetect (Crane, Jolley, & van Belle, 2003) computer program.

Different DIF detection methods often yield disparate identifications of biased items (Teresi, 2001). Thus, descriptive comparisons of items identified by either or both methods were conducted, in order to identify items detected both ways and/or with high levels of potential bias. Interpretation of statistically significant DIF was aided by examining the item parameter estimates and OCCs generated for salient items for each group of interest. Finally, item parameters were re-estimated for those items with the highest degrees of DIF, and IRT scores were re-calculated for all participants. These adjusted scores were compared to the IRT scores obtained without adjustment for DIF using paired *t*-tests, in order to determine whether item-level DIF affected measurement at the level of the combined externalizing subscale. The above steps were consistent with

recent recommendations for assessing DIF effect sizes at the levels of items as well as of scales (Steinberg & Thissen, 2006). Analyses were implemented using MULTILOG 7.03 (Thissen et al., 2003) and SPSS 15.0 (SPSS, 2007) software.

Power and Sample Size Considerations

CTT-based analyses. Independent and paired samples *t*-tests, Pearson correlations, Pearson chi-square tests, and one-way ANOVA are powerful analysis methods for which the planned sample size was more than adequate (Nunnally & Bernstein, 1994). Similarly, the ratio of sample size to number of items analyzed was sufficient for computing Cronbach's α and conducting exploratory factor analysis (A. B. Costello & Osborne, 2005).

IRT-based analyses. Sample size considerations in IRT analyses are not as well-established as for traditional CTT-based methods of investigating psychometric properties of scales. In fact, sample size and its relation to stability of parameter estimation has been identified by numerous authors as an important and potentially rich area of future investigation and development (Fayers, 2004; Tay-Lim & Harwell, 1997). However, results of several simulation studies have led to "rule of thumb" recommendations regarding sample sizes needed for stable parameter estimates and detection of DIF. In general, a minimum of 500 participants is suggested in order to attain relatively stable parameter estimates (Reise & Yu, 1990; Tay-Lim & Harwell, 1997), with 1,000 participants identified as a desirable sample size, when possible. For DIF detection, a minimum of 250 participants per group has been suggested, though lower numbers of participants may be acceptable without loss of reliability of results if

parametric procedures are used (Bolt, 2002). Based on these recommendations, the sample size was sufficient for these analyses.

Integration of Findings

The final step in the study was to integrate the findings from the above set of analyses to address each study hypothesis. Items were compared and classified based on the amount of information they provided, areas of the latent construct continuum they measured most precisely, and the amount (if any) of DIF detected between groups. Items were identified which appeared to (a) measure sub-clinical to clinical levels of externalizing behavior problems in preschool-aged children most precisely, and (b) exhibit the least amount of bias between groups split by child sex, race, and SES. These items were proposed as a set suitable for improved measurement of externalizing behavior problems among very young children in the primary care setting.

CHAPTER V

RESULTS

Sample Characteristics

Caregivers

Of the 938 eligible participants approached in pediatric primary care waiting rooms, 900 primary caregivers of children between the ages of 3 and 5 years agreed to participate, yielding a response rate of 96%. Approximately equal numbers were recruited from each site: UCHS (25%), C&Y (24%), UCHS-S (22%), and OCP (29%). Reasons reported for visits at each site included well child check-ups (26%), sick visits (33%), siblings' appointments (28%), and others (13%), including a wide range of issues from allergy shots to minor injuries to dental care.

Participant ages ranged from 18 to 78 years with a mean of 31 years ($SD = 8$ years). The majority of participants (87%) were female. Most identified themselves as either white (55%) or African-American (42%), with only 3% identifying other racial backgrounds. Participants were not found to differ significantly from non-responders by sex, race, or clinic, the only variables recorded to describe those who declined to participate. Most participants (88%) identified themselves as parents of the children about whom they responded to survey questions, while other reported caregiving relationships included grandparents, step-parents, foster parents, adoptive parents, legal guardians, and other relatives. See Table 1 for more detailed information on caregiver characteristics.

Table 1***Caregiver Characteristics (N = 900)***

Variable	Frequency	(%)
Caregiver Sex		
Male	118	(13)
Female	776	(87)
Caregiver Race		
White	491	(55)
African-American	375	(42)
Other	32	(3)
Caregiver Household Income		
< \$10,000	248	(28)
\$10,001 - \$20,000	187	(21)
\$20,001 - \$30,000	153	(17)
\$30,001 - \$40,000	71	(8)
\$40,001 - \$50,000	62	(7)
\$50,001 - \$60,000	42	(5)
\$60,001 - \$70,000	21	(2)
\$70,001 - \$80,000	22	(3)
\$80,001 - \$90,000	20	(2)
> \$90,000	58	(7)
Caregiver Education		
Less than high school	145	(16)
High school diploma/GED	388	(44)

(table continues)

Table 1 (continued)

Variable	Frequency	(%)
More than high school	355	(40)
Caregiver Relation to Child		
Parent	786	(88)
Step-parent	21	(2)
Grandparent	58	(7)
Foster parent	4	(0)
Other	27	(3)

Note. Percentages do not include missing data and may not sum to 100 percent due to rounding.

Children

Participants provided demographic and mental health information about the children of interest. Approximately equal numbers of children were 3 years (32%), 4 years (38%), and 5 years (29%) old. Just over half of the children of interest were male. Exactly half of the children were reported to be white, with 40% identified as African-American and 10% as other races (including Asian, Hispanic, and bi- or multi-racial). Most children (71%) were reportedly covered by either Medicaid or K-CHIP (Kentucky's SChip program), with more than a quarter covered by private health insurance, and only 1% lacking health insurance coverage. Using the operationalization of SES incorporating household income, parent education, and child health insurance type (see Chapter IV), 42% of children were classified as low SES, 33% as medium SES, and 25% as high SES. Child race (dichotomized as white versus minority) and SES (dichotomized as low versus medium/high) were significantly associated, $\chi^2(1, N = 872) = 52.83, p < .001$. A higher

than expected proportion of white children were of medium/high SES, while a higher than expected proportion of minority children were of low SES. See Table 2 for more detailed child demographic characteristics.

More than one in four participants reported that they believed that the child of interest had behavioral problems, though only one in ten reported that their child had received services from a mental health professional. Approximately 5% of children had reportedly been prescribed medications to treat behavioral problems. Nearly one in five participants indicated that they had expressed concerns about the child’s behavior to a primary care physician, while only a small fraction reported that a primary care physician had expressed concerns to them. A quarter of participants acknowledged that at least one other adult had expressed concerns to them regarding the child’s behavior. See Table 3 for more detailed results.

Table 2
Child Characteristics (N = 900)

Variable	Frequency	%
Child Sex		
Male	472	(53)
Female	424	(47)
Child Race		
White	450	(50)
African-American	362	(40)
Other	88	(10)

(table continues)

Table 2 (continued)

Variable	Frequency	%
Child Household Composition		
Two-parent	512	(57)
Single parent	339	(38)
Caregiver other than parent	47	(5)
Child Program Attendance		
None	218	(24)
Preschool/kindergarten only	454	(51)
Daycare only	145	(16)
Preschool/kindergarten and daycare	82	(9)
Child Health Insurance		
Public	634	(71)
Private	252	(28)
None	10	(1)
Socioeconomic Status (SES)		
Low	371	(43)
Medium	285	(33)
High	216	(25)

Note. Percentages do not include missing data and may not sum to 100 percent due to rounding. See Chapter IV for operationalization of SES.

Table 3***Caregiver-Reported Child Behavioral Health History (N = 900)***

Variable	Frequency	%
Believes child has behavior problems	232	(26)
Child has seen a mental health provider	85	(10)
Child has been prescribed medication(s) for behavior	42	(5)
By primary care provider	21	(2)
By psychiatrist	18	(2)
By other	4	(0)
Caregiver has expressed concerns to primary care provider	163	(18)
Primary care provider has expressed concerns to caregiver	58	(7)
Other adult has expressed concerns to caregiver	217	(24)
Relative	149	(17)
Daycare provider	54	(6)
Teacher/School personnel	47	(5)
Other	22	(2)

Note. Percentages do not include missing data and may not sum to 100 percent due to rounding. More than one response was accepted for the item asking whether other adults had ever expressed concerns to the caregiver.

Classical Test Theory Psychometric Analyses

Classical psychometric analyses were conducted to provide basic information on the measurement properties of the PSC-17 and BPI full scales and externalizing subscales, for comparison with previous studies investigating scale performance. As

outlined in Chapter IV, distributional properties; internal consistency reliability; concurrent and known groups validity; and group differences by sex, race, and SES were explored. For all statistical tests, the level of significance was set at $\alpha = .05$.

Distributional Properties

The distributional properties (i.e., means, standard deviations, skewness, and kurtosis) of the PSC-17 and the BPI full scales and externalizing subscales are presented in Table 4. It is noteworthy that responses to two of the BPI subscales used to create the BPI externalizing subscale (i.e., Peer Problems and Antisocial) demonstrated considerable variability, with the standard deviation of responses to the Peer Problems subscale exceeding the mean. With regard to missing data, fewer than one-half of a percent of participants failed to respond to one or more PSC-17 and BPI items. For both instruments, each full scale and externalizing subscale distribution exhibited mild but statistically significant positive skewness, suggesting the possibility of floor effects. In addition, the distributions of the PSC-17 total scale and BPI externalizing subscale exhibited mild but statistically significant positive kurtosis.

Reliability

Measures of internal consistency were used to investigate the reliability of each instrument and externalizing subscale. Cronbach's α , mean inter-item correlations, and mean corrected item-total correlations are presented in Table 5. Values of the coefficients suggested adequate internal consistency. For the PSC-17 total, PSC-17 externalizing subscale, and BPI externalizing subscale items, no items were identified which would increase Cronbach's α if deleted. For the BPI total scale, however, two items were identified which would not decrease Cronbach's α if deleted: items BPI 2 ("Feels or

complains that no one loves him/her”) and BPI 23 (“Clings to adults”). Neither of these items appeared in the BPI externalizing subscale.

Table 4

Descriptive Statistics for PSC-17, BPI, and Selected Subscales

Subscale	<i>M</i>	<i>SD</i>	Skewness/ <i>SE</i>	Kurtosis/ <i>SE</i>
PSC-17 Externalizing	5.06	2.86	0.47/0.08*	0.21/0.17
PSC-17 Total	9.99	5.51	0.64/0.08*	0.39/0.17*
BPI Externalizing	6.08	4.39	0.92/0.08*	0.50/0.17*
BPI Antisocial	1.97	1.86	1.07/0.08*	0.66/0.16*
BPI Headstrong	3.63	2.43	0.58/0.08*	-0.27/0.16
BPI Peer Problems	0.70	1.04	1.77/0.08*	3.52/0.16*
BPI Total	13.76	8.99	0.85/0.08*	0.32/0.17

Note. Positive skewness indicates a distribution with a long right tail and negative skewness indicates a distribution with a long left tail. Positive kurtosis indicates that the observations cluster more and have longer tails than the normal distribution, while negative kurtosis indicates that the observations cluster less and have shorter tails. In general, skewness and kurtosis estimates which are twice their standard errors are indicative of significant deviations from normality.

* $p < .05$.

Validity

Concurrent validity was explored with bivariate Pearson correlations between the PSC-17 and the BPI, as well as between the externalizing subscales of each instrument.

In addition, known groups validity was assessed using independent samples *t*-tests of mean differences in full scale and externalizing subscale scores between (a) participants who reported believing that their child had behavior problems and those who did not, and

(b) participants who reported that their child had been seen by a mental health professional and those who did not.

Table 5

Internal Consistency of PSC-17, BPI, and Selected Subscales

Subscale	Cronbach's α	Mean Inter-Item Correlation	Mean Corrected Item- Test Correlation
PSC-17 Externalizing	.79	.35	.51
PSC-17 Total	.86	.26	.47
BPI Externalizing	.85	.34	.54
BPI Antisocial	.71	.38	.49
BPI Headstrong	.77	.40	.54
BPI Peer Problems	.60	.33	.41
BPI Total	.91	.29	.51

Concurrent validity. Scores on the PSC-17 and BPI total scales were strongly significantly positively correlated ($r = .80, p < .01, N = 825$). Externalizing subscale scores of each instrument were also significantly positively correlated to a lesser degree ($r = .67, p < .01, N = 859$).

Known groups validity. Participants were divided into several groups indicative of possible child behavioral problems. First, responses to the survey item asking whether the respondent believed that the child had behavioral problems were used to divide the sample into those who did and did not hold that belief. Similarly, responses to the item inquiring whether the child had been seen by a mental health professional were used to

divide the sample into two additional groups. For each set of groups, mean differences in PSC-17 and BPI full scale and externalizing subscale scores were investigated using independent samples *t*-tests. Results are presented in Tables 6 and 7. Results consistently demonstrated significantly higher total and externalizing subscale scores among participants who believed their child had behavior problems and who reported that their child had been seen by a mental health professional.

Table 6

Known Groups Validity: Parent Belief that Child has Behavior Problems

Subscale and Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>
PSC-17 Externalizing						
Behavior Problems	226	7.38	2.92	14.47	332.60 ^a	< .001
None	653	4.27	2.36			
PSC-17 Total						
Behavior Problems	216	15.21	5.30	17.43	317.85 ^a	< .001
None	639	8.25	4.34			
BPI Externalizing						
Behavior Problems	224	10.72	4.34	19.84	304.58 ^a	< .001
None	650	4.49	3.09			
BPI Total						
Behavior Problems	223	23.65	8.17	22.18	319.03 ^a	< .001
None	637	10.30	6.30			

^aSatterthwaite's approximation for the degrees of freedom was utilized due to unequal variances between groups detected by Levene's test.

Table 7***Known Groups Validity: Differences in Mean Scores by Child History of Contact with Mental Health Professional (MHP)***

Subscale and Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>
PSC-17 Externalizing						
Contact with MHP	83	7.43	3.37	6.83	93.14 ^a	< .001
No Contact	796	4.82	2.68			
PSC-17 Total						
Contact with MHP	77	15.78	6.31	8.54	85.98 ^a	< .001
No Contact	778	9.44	5.07			
BPI Externalizing						
Contact with MHP	82	10.63	5.01	8.79	92.28 ^a	< .001
No Contact	792	5.61	4.05			
BPI Total						
Contact with MHP	81	24.06	9.37	11.65	858.00	< .001
No Contact	779	12.69	8.25			

^aSatterthwaite's approximation for the degrees of freedom was utilized due to unequal variances between groups detected by Levene's test.

Group Differences by Child Sex, Race, and Socioeconomic Status

Differences in participant responses as well as scale performance related to child sex, race, and SES were explored to provide additional context for the IRT item-level analyses. Differences in mean full scale and externalizing subscale scores were investigated using independent samples *t*-tests (for sex and race) and one-way ANOVA (for SES). Finally, CTT psychometric properties were reassessed after dividing the

sample by sex, race, and SES. Due to very low numbers (i.e., 10%) of participants identifying their child's racial background as one other than white or African-American, all classifications of "other" were combined with the African-American group and designated as *minority* in these analyses. (Findings were similar but power was lost in some analyses when three racial groups were used rather than two.)

Differences by sex. Differences in mean PSC-17 total, BPI total, PSC-17 externalizing subscale, and BPI externalizing subscale scores between boys and girls were investigated using independent samples *t*-tests. Results are reported in Table 8. Statistically significant differences in mean scores between boys and girls were found only on the PSC-17 total score, with boys scoring higher than girls on this scale.

Differences by race. Differences in mean scores between white and minority children were also investigated using independent samples *t*-tests. Results are reported in Table 9. No significant differences were found. The lack of significant differences in mean scores between white and minority children, however, did not exclude the possibility of item-level bias, explored in later IRT analyses.

Differences by SES. Differences in mean scores among low, medium, and high SES children were explored using one-way analysis of variance (ANOVA). Results are presented in Table 10. Significant group differences were detected in each mean full scale and externalizing subscale score. Post hoc analyses using the Scheffé criterion for significance indicated that low SES children consistently scored higher on each full scale and externalizing subscale score, as compared to medium and high SES children (who did not differ significantly from each other).

Table 8***Differences in Mean Scores by Child Sex***

Subscale and Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>
PSC-17 Externalizing						
Male	460	5.20	2.97	1.67	874.88 ^a	.10
Female	417	4.88	2.72			
PSC-17 Total						
Male	445	10.42	5.76	2.51	850.58 ^a	< .05
Female	408	9.48	5.17			
BPI Externalizing						
Male	459	10.63	6.29	1.64	869.00 ^a	.10
Female	412	5.61	5.81			
BPI Total						
Male	451	14.22	9.46	1.67	854.83 ^a	.10
Female	406	13.19	8.40			

^aSatterthwaite's approximation for the degrees of freedom was utilized due to unequal variances between groups detected by Levene's test.

Table 9***Differences in Mean Scores by Child Race***

Subscale and Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>
PSC-17 Externalizing						
White	442	4.99	2.74	-0.74	879	.46
Minority	439	5.13	2.98			

(table continues)

Table 9 (continued)

Subscale and Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>
PSC-17 Total						
White	434	10.08	5.29	0.47	855	.64
Minority	423	9.90	5.73			
BPI Externalizing						
White	438	6.21	4.36	0.91	873	.36
Minority	437	5.95	4.43			
BPI Total						
White	436	13.73	8.61	-0.09	859	.93
Minority	425	13.79	9.37			

Table 10***Differences in Mean Scores by Child Socioeconomic Status***

Subscale and SES Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>F</i>	<i>df1, df2</i>	<i>p</i>
PSC-17 Externalizing						
Low ^a	364	5.61	3.00	12.79	2, 851	< .001
Medium	278	4.69	2.79			
High	212	4.56	2.44			
PSC-17 Total						
Low ^a	354	11.04	6.11	11.28	2, 828	< .001
Medium	273	9.33	5.13			
High	204	9.11	4.55			

(table continues)

Table 10 (continued)

Subscale and SES Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>F</i>	<i>df1, df2</i>	<i>p</i>
BPI Externalizing						
Low ^a	358	6.89	4.87	11.28	2, 846	< .001
Medium	279	5.61	4.03			
High	212	5.29	3.89			
BPI Total						
Low ^a	352	15.83	10.01	17.26	2, 832	< .001
Medium	274	12.66	8.09			
High	209	11.70	7.79			

Note. SES = socioeconomic status.

^aPost hoc tests using the Scheffé criterion for significance revealed that in each case the low SES group scored significantly higher than the medium and high SES groups (p 's < .001). The medium and high SES groups did not differ significantly from each other.

Psychometric properties and sex, race, and SES. Indicators of internal consistency were re-examined after splitting the sample by sex, race, and SES. No salient differences were noted in Cronbach's α , mean inter-item correlations, or corrected item-total correlations among the groups, suggesting that in terms of classical psychometric analyses, the total scales and externalizing subscales performed fairly consistently.

Item Response Theory Analyses

To address the research questions and hypotheses posed in Chapter III, IRT analyses were conducted assessing the performance of individual items in the combined externalizing subscales of the PSC-17 and the BPI. As described in Chapter IV, several steps were required. First, IRT model assumptions were evaluated. Next, to answer the first research question and associated hypotheses, Samejima's (1969) GRM was fitted to

the observed data, yielding estimates of item parameters and information. Finally, to answer the second research question and associated hypotheses, each item was evaluated for DIF between groups split by child sex, race, and SES.

Evaluation of IRT Model Assumptions

As explained in Chapter IV, assessment of the strong assumptions underlying IRT was an important first step. Three primary assumptions are made for all IRT models: unidimensionality, local independence, and specific trace line functions. When possible, more than one strategy was used to evaluate each assumption.

Unidimensionality. An initial assessment of unidimensionality involved consideration of a CTT internal consistency reliability indicator. Cronbach's α for the combined externalizing subscale was .89, suggesting that the items correlated highly with each other. While not strictly a measure of unidimensionality, this finding revealed consistent within-subject responses, which can be considered one aspect of unidimensionality.

However, since high levels of internal consistency are possible with multidimensional data (McDonald, 1981), exploratory factor analysis (EFA) was also conducted on the combined externalizing subscale. Unidimensionality was evaluated by forcing a single factor using principal axis factoring as the extraction method. Results demonstrated that the single factor (eigenvalue = 6.53) accounted for 36% of the variance. This exceeded the minimum standard of 20% suggested by Reckase (1979) as sufficient for a scale to be "unidimensional enough" for IRT analyses. In addition, the first factor eigenvalue (6.53) was 5.05 times the second factor eigenvalue (1.29),

exceeding the criterion of 5 times suggested by Hambleton and colleagues (1991) for demonstrating a dominant single factor.

Magnitudes of eigenvalues for additional factors and strength of factor loadings, in combination with visual evaluation of a scree plot (Bjorner et al., 2003a, 2003b), were also reviewed to consider the unidimensionality of each subscale. Eigenvalues of additional factors “elbowed” beginning with the second factor, further supporting the dominance of the first factor. In addition, single factor structure coefficients ranged from .45 to .69 (see Table 11). Treating the combined externalizing subscale as a single measure, it appeared that the unidimensionality assumption was adequately met.

Local independence. As described in Chapter IV, the assumption of local independence requires that once the level of externalizing behavior is controlled, items should be statistically independent from one another (Steinberg & Thissen, 1996; Wainer & Thissen, 1996; Yen, 1993). This assumption was evaluated via examination of the residual correlation matrix from the EFA for the combined externalizing subscale, using Reeve and colleagues’ (2007) criterion of $|r| \geq .20$ for violation of local independence. After the single factor was extracted via EFA, absolute values of residual correlations for each pair of items ranged from .00 to .15, indicating that the assumption of local independence was adequately met.

Specific trace line functions. The assumption of specific trace line functions, as applied to the GRM, refers to the requirement that the probability of selecting higher item response options increases with higher levels of the latent trait, and never decreases. This assumption was assessed by fitting a non-parametric IRT model to the data from the combined externalizing subscale and graphing the results, generating a trace line from the

Table 11***Summary of Exploratory Factor Analysis Results for Combined Externalizing******Subscale (N = 861)***

Item	Short Wording	Factor Loading
PSC-17 4	Refuses to share	.50
PSC-17 5	Does not understand others' feelings	.47
PSC-17 8	Fights others	.62
PSC-17 10	Blames others	.55
PSC-17 12	Does not listen to rules	.65
PSC-17 14	Teases others	.52
PSC-17 16	Takes things	.56
BPI 3	High strung	.45
BPI 4	Cheats/lies	.51
BPI 6	Argues too much	.58
BPI 9	Bullies/cruel or mean	.69
BPI 10	Disobedient at home	.60
BPI 11	Not sorry after misbehaves	.59
BPI 12	Trouble getting along with others	.64
BPI 15	Not liked by others	.47
BPI 18	Stubborn, sullen, or irritable	.54
BPI 19	Very strong temper	.67
BPI 22	Breaks/destroys things	.62

Note. Results are for the forced single-factor solution using principal axis factoring.

observations. The trace lines for each item were then visually inspected for the expected form. This analysis was conducted using TestGraf software (Ramsay, 2000). The non-parametric trace line plots revealed that all items clearly exhibited the expected form. See Figure 6 for an example of a non-parametric trace line plot generated for a single item. As expected, the probability of selecting response options endorsing more behavioral problems increased as the level of externalizing behavior problems increased, suggesting that the specific trace lines assumption was met.

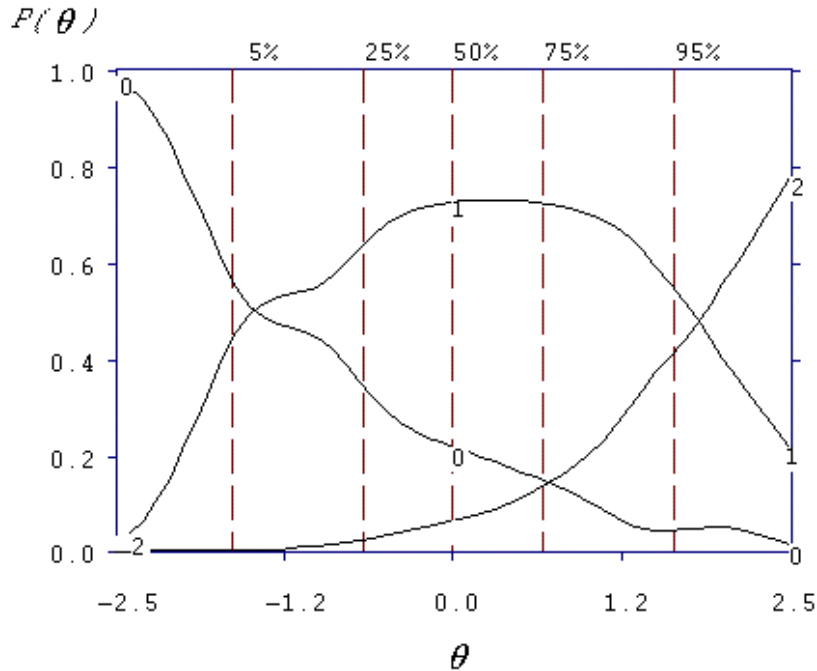


Figure 6. Non-parametric trace line plot for item PSC-17 4 (“Refuses to share”). Option 0 = *never*; option 1 = *sometimes*; option 2 = *often*.

In summary, all three assumptions underlying the application of IRT models appeared to be met. The items in the combined externalizing subscale were unidimensional, demonstrated local independence, and were characterized by the

expected trace line functions when a non-parametric model was fit. Following evaluation of the IRT model assumptions, a specific polytomous IRT model was fit to the data to address the first research question regarding the precision and utility of measurement offered by each item.

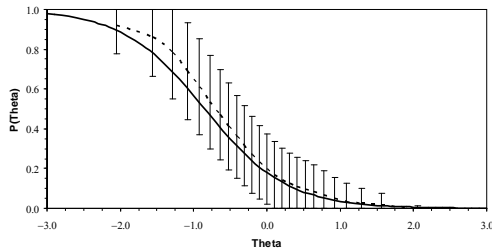
Research Question 1: Precision and Utility of Measurement

Samejima's (1969) GRM was fit to the data. Details regarding model fit are provided below. In addition, item parameter estimates, test information, and item information for data from the full sample are presented.

Model fit. The goodness-of-fit of the GRM was assessed graphically with fit plots as well as statistically with tests suggested by Drasgow and colleagues (1995). Fit plots depicting (a) the OCCs estimated with the GRM for the calibration sample, and (b) the empirical proportions of endorsed responses for each category for the cross-validation sample were produced using the MODFIT (Stark, 2002) computer program. Each item was represented by three fit plots, one for each response option (i.e., 0, 1, and 2). Examination of fit plots for each item suggested overall good fit, though several items displayed some degree of misfit. Figure 7 provides sample fit plots for 2 items: items PSC-17 8 ("Fights others") and BPI 15 ("Not liked by others"). The degree of misfit observed for item PSC-17 8 was typical of that seen for 6 of the 18 items, in that the overall fit appeared adequate with deviations noted in the tails of one or more OCCs. The remaining 12 items displayed negligible deviations, illustrated by the fit plots for item BPI 15. For these items, all cross-validation empirical curves fell within the 95% confidence intervals of the GRM parameter estimates.

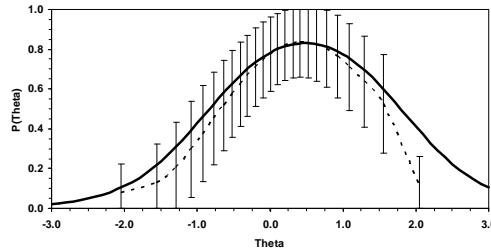
(a) PSC-17 Item 8 (“Fights others”)

Option 0



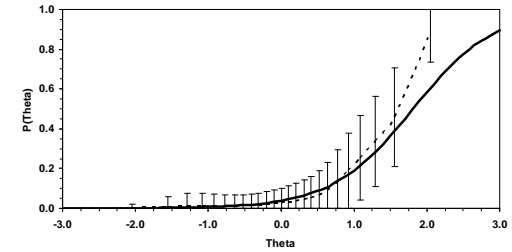
(b) PSC-17 Item 8 (“Fights others”)

Option 1



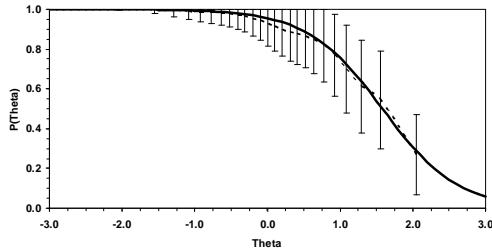
(c) PSC-17 Item 8 (“Fights others”)

Option 2



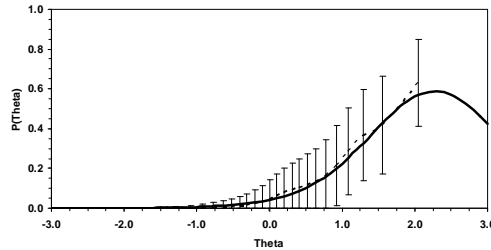
(d) BPI Item 15 (“Not liked by others”)

Option 0



(e) BPI Item 15 (“Not liked by others”)

Option 1



(f) BPI Item 15 (“Not liked by others”)

Option 2

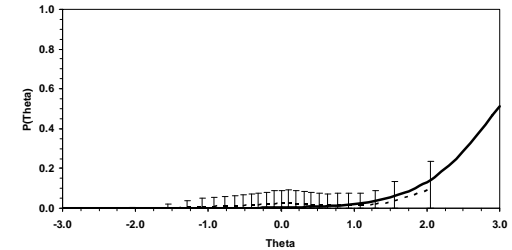


Figure 7. Sample fit plots for the graded response model option characteristic curves (OCCs) of two items. Solid curves represent the GRM OCCs estimated using the calibration sample ($n = 450$). Dashed curves represent the empirical proportions of responses for each option observed in the cross-validation sample ($n = 450$). Vertical bars represent the 95% confidence intervals for the model-based estimates of the OCCs.

Results of the statistical tests of model fit recommended by Drasgow and colleagues (1995) are presented in Table 12. The frequency distributions of χ^2 to degrees of freedom (df) ratios above and below 3 for singlets, doublets and triplets of items are included. Mean values and standard deviations of the χ^2/df ratio are also provided for each type of item combination. All mean ratios were below the cut-off of 3 recommended by Drasgow and colleagues (1995). Considering the magnitudes of the χ^2/df ratios and the fit suggested by the graphical fit plots, the fit of the GRM to the data was deemed acceptable.

Table 12

Goodness of Fit: Frequencies and Means of Chi Square to Degrees of Freedom Ratios

Item Groups	$\chi^2/df < 3$	$\chi^2/df > 3$	<i>M</i>	<i>SD</i>
Singlets	12	6	2.56	2.72
Doublets	10	8	2.75	1.42
Triplets	3	3	2.72	0.98

Note. χ^2 values were computed from expected counts from model-fitting with a calibration sample to observed counts from a cross-validation sample. Ratios of χ^2 divided by degrees of freedom (df) were calculated for single items, pairs of items, and triples of items. Ratios ≤ 3 generally indicate good fit (Drasgow et al., 1995).

Item parameter estimates. As discussed in Chapter III, in the current application of the GRM, each item is characterized by three parameter estimates: a (discrimination), b_1 (difficulty threshold between option 0 and option 1), and b_2 (difficulty threshold between option 1 and option 2). High values of a indicate highly discriminating items, meaning that items are better able to distinguish between participants at similar levels of

externalizing behavior problems, as compared to items with lower values of a . Guidelines for interpretation of the discrimination parameter were offered by Baker (1985), who suggested the following classification: $a < 0.20$, very low discrimination; $0.21 < a < 0.40$, low discrimination; $0.41 < a < 0.80$, moderate discrimination; $0.81 < a < 1.00$, high discrimination; $a \geq 1.00$, very high discrimination. Values of the parameters b_1 and b_2 provide the difficulty level of the item via the locations of the intersections of the OCCs along the continuum of externalizing behavior problems. Item parameter estimates and standard errors for each item in the combined externalizing subscale based on data from the full sample are presented in Table 13, as well as basic CTT descriptive information regarding item means and corrected item-total correlations. In addition, plots of OCCs for all 18 combined externalizing subscale items are provided in Figure 8, illustrating the meaning of the estimated item parameters.

According to Baker's (1985) guidelines, all 18 items demonstrated very high discrimination ($M = 1.62$, $SD = 0.34$). The highest quartile of discrimination parameters included those for items PSC-17 8 ($a = 1.94$, $se = 0.15$); BPI 19 ($a = 1.99$, $se = 0.16$); BPI 12 ($a = 2.02$, $se = 0.17$); PSC-17 12 ($a = 2.07$, $se = 0.16$); and BPI 9 ($a = 2.27$, $se = 0.19$). The lowest discrimination parameter estimate was for item BPI 3 ($a = 1.10$, $se = 0.12$). The effects of higher versus lower discrimination parameters can be seen in Figure 8 by comparing the OCC plots for items BPI 9 (part [k]) and BPI 3 (part [h]), in which the item with the highest discrimination parameter estimate (i.e., BPI 9) exhibits steeper curves than the item with the lowest discrimination parameter estimate.

Difficulty parameter estimates among items differed as well. The distribution of the b_1 difficulty parameter was centered just below the mean level of externalizing

Table 13*Item Descriptives and Graded Response Model Parameter Estimates for Total Sample (N = 900)*

Item	Short Wording	Item Descriptives		Parameter Estimates		
		<i>M (SD)</i>	<i>r_{it}</i>	<i>a_i (se)</i>	<i>b_{1i} (se)</i>	<i>b_{2i} (se)</i>
PSC-17 4	Refuses to share	0.85 (0.60)	.47	1.29 (0.12)	-1.12 (0.11)	1.89 (0.17)
PSC-17 5	Does not understand others' feelings	0.66 (0.62)	.44	1.21 (0.11)	-0.43 (0.09)	2.37 (0.23)
PSC-17 8	Fights others	0.81 (0.59)	.58	1.94 (0.15)	-0.82 (0.07)	1.65 (0.12)
PSC-17 10	Blames others	0.59 (0.66)	.52	1.47 (0.13)	-0.07 (0.07)	1.89 (0.16)
PSC-17 12	Does not listen to rules	1.00 (0.59)	.61	2.07 (0.16)	-1.28 (0.09)	1.11 (0.09)
PSC-17 14	Teases others	0.50 (0.60)	.49	1.34 (0.13)	0.11 (0.08)	2.53 (0.22)
PSC-17 16	Takes things	0.65 (0.64)	.53	1.50 (0.13)	-0.33 (0.07)	1.92 (0.16)
BPI 3	High strung	0.43 (0.65)	.43	1.10 (0.12)	0.64 (0.11)	2.49 (0.27)
BPI 4	Cheats/lies	0.67 (0.65)	.49	1.26 (0.11)	-0.37 (0.09)	2.08 (0.19)
BPI 6	Argues too much	0.79 (0.72)	.55	1.43 (0.13)	-0.54 (0.08)	1.37 (0.13)
BPI 9	Bullies/cruel or mean	0.45 (0.63)	.64	2.27 (0.19)	0.31 (0.06)	1.73 (0.12)

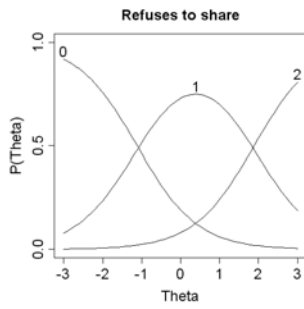
(table continues)

Table 13 (continued)

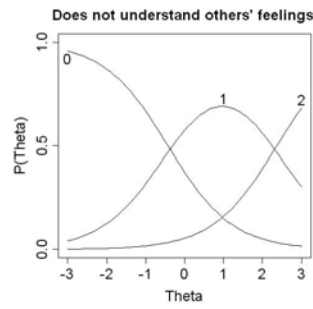
Item	Short Wording	Item Descriptives		Parameter Estimates		
		<i>M (SD)</i>	<i>r_{it}</i>	<i>a_i (se)</i>	<i>b_{1i} (se)</i>	<i>b_{2i} (se)</i>
BPI 10	Disobedient at home	0.86 (0.62)	.56	1.72 (0.15)	-0.92 (0.08)	1.52 (0.12)
BPI 11	Not sorry after misbehaves	0.49 (0.65)	.55	1.61 (0.14)	0.24 (0.07)	1.96 (0.16)
BPI 12	Trouble getting along with others	0.38 (0.57)	.60	2.02 (0.17)	0.45 (0.06)	2.22 (0.17)
BPI 15	Not liked by others	0.14 (0.39)	.44	1.65 (0.21)	1.65 (0.15)	3.17 (0.37)
BPI 18	Stubborn, sullen, or irritable	0.87 (0.67)	.52	1.41 (0.12)	-0.92 (0.10)	1.43 (0.13)
BPI 19	Very strong temper	0.70 (0.73)	.63	1.99 (0.16)	-0.22 (0.06)	1.21 (0.09)
BPI 22	Breaks/destroys things	0.37 (0.62)	.58	1.88 (0.17)	0.61 (0.07)	1.91 (0.14)

Note. r_{it} = corrected item-total correlation; a_i = item slope parameter; se = standard error; b_{1i} = item lower threshold difficulty parameter; b_{2i} = item upper threshold difficulty parameter.

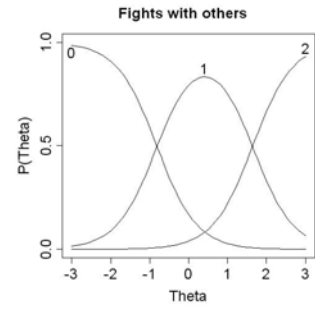
(a) PSC-17 Item 4



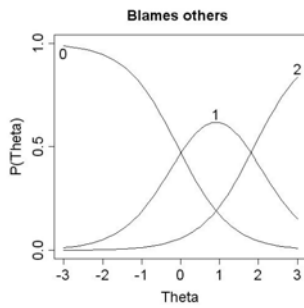
(b) PSC-17 Item 5



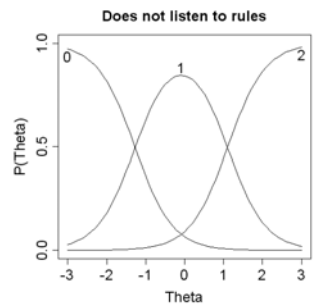
(c) PSC-17 Item 8



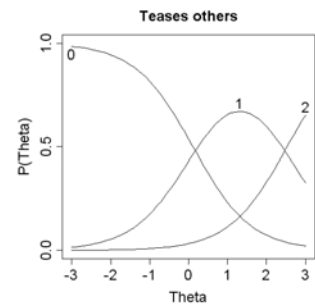
(d) PSC-17 Item 10



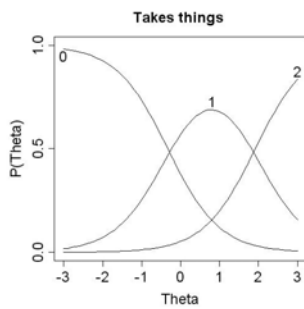
(e) PSC-17 Item 12



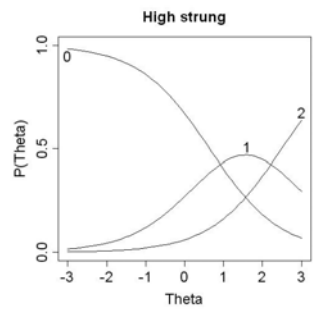
(f) PSC-17 Item 14



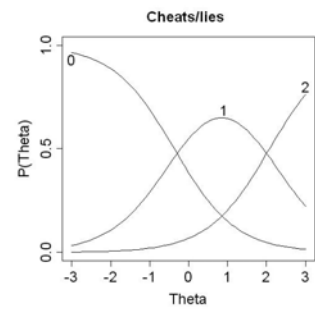
(g) PSC-17 Item 16



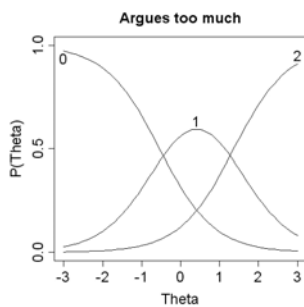
(h) BPI Item 3



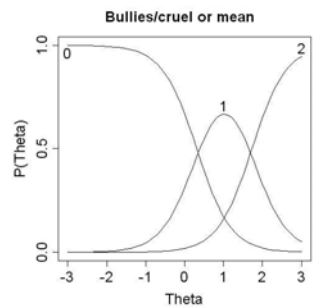
(i) BPI Item 4



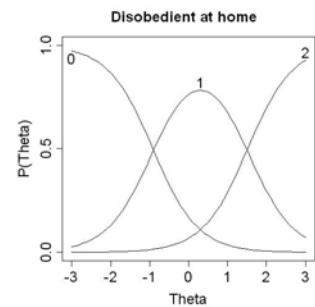
(j) BPI Item 6



(k) BPI Item 9



(l) BPI Item 10



(figure continues)

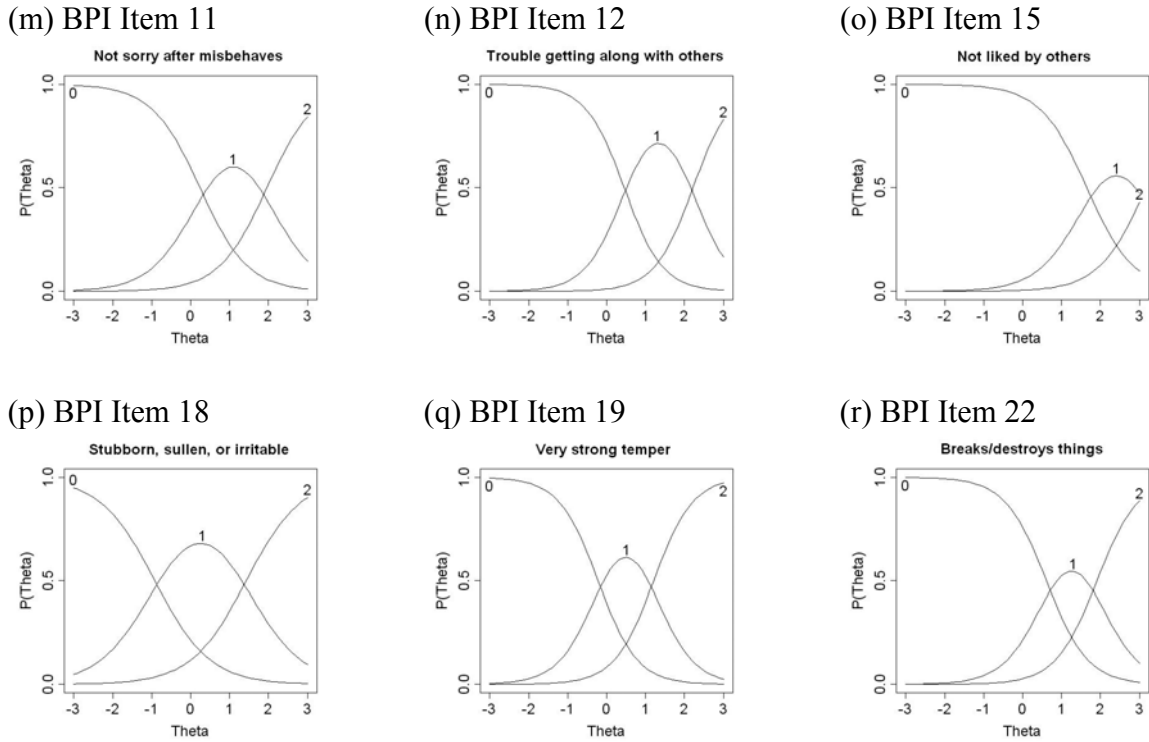


Figure 8. Plots of graded response model option characteristic curves (OCCs) for all items in the combined externalizing subscale.

behavior problems ($M = -0.17$, $SD = 0.74$). This suggests that the threshold level of externalizing behavior problems required for a randomly selected participant to select response option 1 (*sometimes* or *sometimes true*) rather than response option 0 (*never* or *not true*) was, on average, just below the mean level of externalizing behavior problems. The lowest b_I parameter estimate was for item PSC-17 12 ($b_I = -1.28$, $se = 0.09$), making this item the *easiest* of the set—in other words, very low levels of externalizing behavior problems were necessary for a caregiver to respond that the child *sometimes* does not follow rules, versus responding *never* to this item. Other items with low b_I parameter estimates included items PSC-17 4 ($b_I = -1.12$, $se = 0.11$); BPI 10 ($b_I = -0.92$, $se = 0.08$); and PSC-17 8 ($b_I = -0.82$, $se = 0.07$). In contrast, several items exhibited much higher difficulty levels for their lower thresholds: items BPI 22 ($b_I = 0.61$, $se = 0.07$); BPI 3 (b_I

= 0.64, $se = 0.11$); and BPI 15 ($b_1 = 1.65$, $se = 0.15$) had the highest b_1 parameter estimates.

Estimates for the upper difficulty threshold parameter b_2 were also disparate. The distribution of the b_2 difficulty parameter estimates clustered between 1.5 and 2 standard deviations above the mean ($M = 1.91$, $SD = 0.52$). Thus, the average threshold level of externalizing behavior problems required for a randomly selected participant to select response option 2 (*often* or *often true*) rather than response option 1 (*sometimes* or *sometimes true*) was in the sub-clinical to clinical range of externalizing behavior problems. The highest b_2 parameter estimate was for item BPI 15 ($b_2 = 3.17$, $se = 0.37$), making this item the most *difficult* of the set: Extremely high levels of externalizing behavior problems were necessary for a caregiver to respond that the child *often* is not liked by other children, versus responding *sometimes* to this item. The other items comprising the highest quartile of b_2 parameter estimates included items BPI 12 ($b_2 = 2.22$, $se = 0.17$); PSC-17 5 ($b_2 = 2.37$, $se = 0.23$); BPI 3 ($b_2 = 2.49$, $se = 0.27$); and PSC-17 14 ($b_2 = 2.53$, $se = 0.22$). Several items, however, exhibited much lower difficulty levels for their upper thresholds: items PSC-17 12 ($b_2 = 1.11$, $se = 0.09$); BPI 19 ($b_2 = 1.21$, $se = 0.09$); BPI 6 ($b_2 = 1.37$, $se = 0.13$); and BPI 18 ($b_2 = 1.43$, $se = 0.13$) all had b_2 parameter estimates lower than 1.5 standard deviations above the mean level of externalizing behavior problems.

The effects of lower versus higher b_1 and b_2 parameters on overall item functioning can be seen in Figure 8 by comparing the OCC plots for the least difficult (i.e., PSC-17 12, part [e]) versus the most difficult (i.e., BPI 15, part [o]) items. The difficulty parameter estimates for item PSC-17 12 locate its entire set of curves further to

the left on the continuum of externalizing behavior problems than is seen in more difficult items' plots. These plots illustrate the relationship between items' difficulty levels (as represented by their b_1 and b_2 parameter estimates) and the continuum of externalizing behavior problems.

Results suggested that, as hypothesized, the items from the combined externalizing subscales of the PSC-17 and the BPI exhibited different levels of discrimination and difficulty. Consideration of test and item information was the next step in assessing the precision and utility of measurement offered by each item.

Test information. As discussed in Chapter III, the test information function reveals at what levels of the latent variable a given set of items measures most precisely. Figure 9 provides a graphical illustration of the test information function yielded by retaining all items in the combined externalizing subscale. Information for measurement of externalizing behavior problems with this set of 18 items was highest from approximately 1.5 standard deviations below the mean to just over 3 standard deviations above the mean. Because the SEE is derived from the reciprocal of the information function, precision of measurement is high where information is high; error is high where information is low. The test information curve peaks between 1.5 and 2 standard deviations above the mean, a desirable range for precise measurement of clinical and sub-clinical levels of externalizing behavior problems.

Item information. Because test information functions are generated by summing the information functions of the individual items which comprise the test, the information functions of each item in the combined externalizing subscale were reviewed. Particular attention was paid to identification of items which most precisely measured clinical and

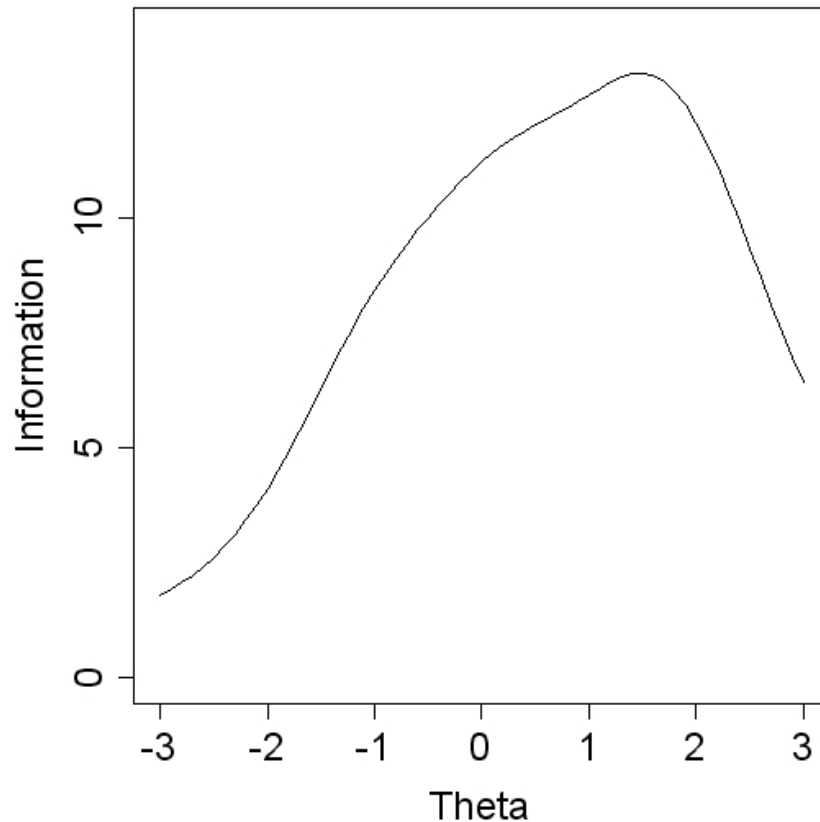


Figure 9. Test information function plot for all items in the combined externalizing subscale. Test information exceeds the standard error of estimation (SEE) between approximately 1.5 standard deviations below and 3 standard deviations above the mean level of externalizing behavior problems.

sub-clinical levels of externalizing behavior problems. See Table 14 for a summary of each item's (a) highest level of information, and (b) levels of externalizing behavior problems (i.e., θ values) at which information was greatest.

The 13 items in bold print in Table 14 demonstrated peaks in information within the sub-clinical to clinical range of externalizing behavior problems. Some, however, offered more information than others at similar levels of θ . The relative amounts of information offered by these items along the sub-clinical to clinical range of externalizing behavior problems is illustrated in Figure 10.

Table 14***Maximum Item Information Estimates and Locations***

Item	Short Wording	Maximum <i>I</i>	Theta Values ^a with Highest <i>I</i>
PSC-17 4	Refuses to share	0.42	-1.00, 1.80
PSC-17 5	Does not understand others' feelings	0.38	-0.40, 2.20
PSC-17 8	Fights others	0.95	-0.80, 1.60
PSC-17 10	Blames others	0.57	0.20, 1.70
PSC-17 12	Does not listen to rules	1.07	-1.20, 1.00
PSC-17 14	Teases others	0.47	0.20, 2.40
PSC-17 16	Takes things	0.58	-0.20, 1.80
BPI 3	High strung	0.35	1.40, 1.60
BPI 4	Cheats/lies	0.41	-0.20, 2.00
BPI 6	Argues too much	0.55	-0.30, 1.20
BPI 9	Bullies/cruel or mean	1.34	0.40, 1.60
BPI 10	Disobedient at home	0.74	-0.80, 1.40
BPI 11	Not sorry after misbehaves	0.69	0.40, 1.80
BPI 12	Trouble getting along with others	1.05	0.60, 2.20
BPI 15	Not liked by others	0.74	2.00, 2.80
BPI 18	Stubborn, sullen, or irritable	0.52	-0.80, 1.30
BPI 19	Very strong temper	1.05	0.00, 1.00
BPI 22	Breaks/destroys things	0.97	0.90, 1.60

Note. **Bolded** items indicate that information peaks at 1.5 standard deviations above the mean or more. *I* = Information.

^aTheta values are rounded within 0.05 standard deviations.

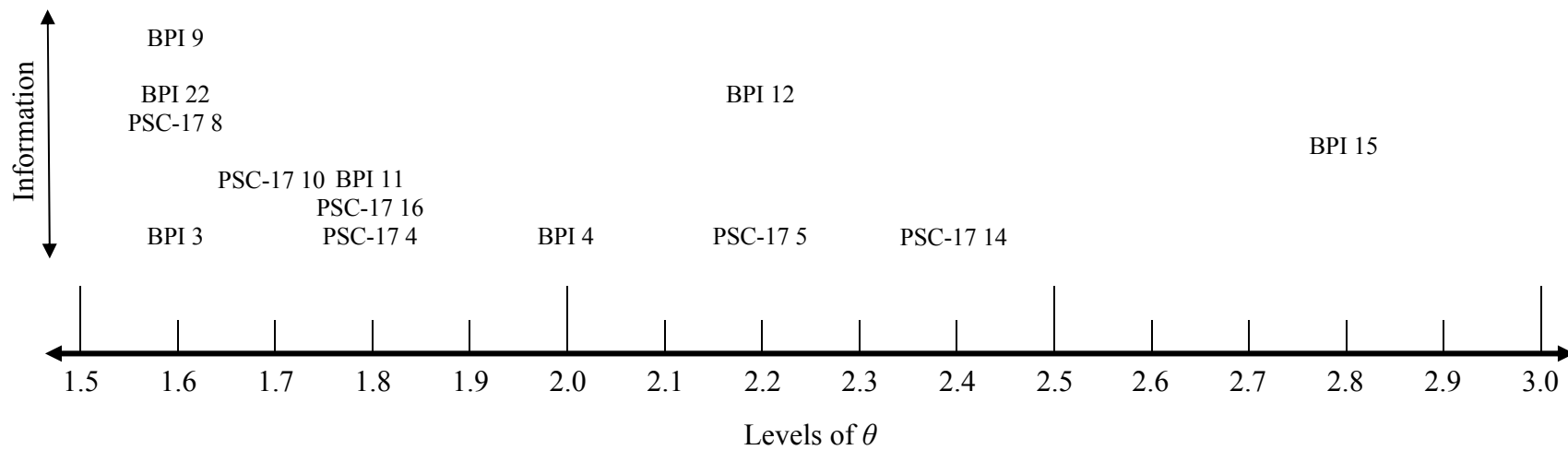


Figure 10. Relative levels of item information provided in the sub-clinical to clinical range of externalizing behavior problems.

As hypothesized, the items in the combined externalizing subscale provided disparate amounts of information in the measurement of externalizing behavior problems among very young children. Of the 18 items, 5 were most informative at levels below the sub-clinical range of externalizing behavior problems. The remaining 13 items yielded varying levels of information along the range of sub-clinical to clinical externalizing behavior problems.

Research Question 2: Item-level Measurement Bias

Two methods were used to examine each item in the combined externalizing subscale for DIF among groups differing by child sex, race, and SES. First, a likelihood-based model comparison approach (IRT-LR) was implemented using IRTLRDIF freeware (Thissen, 2001). Next, an ordinal logistic regression (OLR) technique was conducted using the approach outlined by Crane and colleagues (Crane et al., 2004). Results of each method of DIF detection are presented below, followed by a comparison of the findings yielded by each technique and provision of data regarding the extent of DIF observed.

IRT-LR. As described in Chapter IV, the IRT-LR method evaluated the statistical significance of differences between item parameters estimated for specific groups of interest: by child sex (male versus female), race (white versus minority), and SES (medium/high versus low). For each comparison, a likelihood ratio test statistic provided an overall significance test for the null hypothesis that none of the three parameters of an item's response function (i.e., a , b_1 and b_2) differed between groups. For a given item, if the overall likelihood ratio statistic G^2 with 3 degrees of freedom was greater than or equal to 3.84 (the critical value for a single degree of freedom test, used in this case to

minimize Type II error), then further tests were conducted using nested models to assess the significance of differences between the individual parameters. In interpreting these nested model tests, a significant difference in difficulty (b_1 and b_2) or discrimination (a) parameters for an item between groups required $p < .0027$, after a Bonferroni correction for multiple comparisons ($.05/18$) was implemented to preserve overall α at the .05 level. Results of the IRT-LR method are summarized in three tables: Table 15 presents results for DIF analyses comparing item parameters for male and female children; Table 16 presents results for white and minority children; and Table 17 presents results for low SES and medium/high SES children.

For groups defined by child sex, no items demonstrated DIF at the level of significance required after the Bonferroni adjustment for multiple comparisons. However, two items demonstrated DIF in difficulty parameters at the uncorrected $p < .05$ level of significance. Item PSC-17 4 (“Refuses to share”) was easier for male children than for female children; lower levels of externalizing behavior problems were needed in boys for the caregiver to endorse higher response options for this item. In contrast, item BPI 22 (“Breaks/destroys things on purpose”) was more difficult for male children than for female children. For this item, higher levels of externalizing behavior problems were needed in boys for the caregiver to endorse higher response options.

For groups defined by child race, three items exhibited DIF in difficulty parameters between white children and minority children at the more stringent level of significance set via the Bonferroni correction. Item PSC-17 14 (“Teases others”) was more difficult for white children than for minority children; higher levels of externalizing behavior problems were needed in white children for caregivers to endorse higher

response options for this item. Items BPI 3 (“High strung”) and BPI 6 (“Argues too much”), however, were easier for white children than for minority children. For each of these items, caregivers of white children required lower levels of externalizing behavior problems to select higher response options. In addition, nine other items demonstrated DIF by race at the uncorrected $p < .05$ level of significance. Items PSC-17 4 (“Refuses to share”), PSC-17 16 (“Takes things”), BPI 9 (“Bullies/cruel or mean”), BPI 11 (“Not sorry after misbehaves”), BPI 12 (“Trouble getting along with others”), and BPI 22 (“Breaks/destroys things”) were all more difficult for white children than for minority children, requiring higher levels of externalizing behavior problems for caregivers to select higher response options. Finally, items PSC-17 5 (“Does not understand others’ feelings”) and PSC-17 8 (“Fights others”) were more discriminating for white children than for black children, while for item PSC-17 10 (“Blames others”), the reverse was true.

For groups defined by child SES, three items exhibited DIF in difficulty parameters between low SES children and medium/high SES children at the Bonferroni-corrected level of significance. Items BPI 3 (“High strung”) and BPI 18 (“Stubborn, sullen, or irritable”) were both more difficult for low SES compared to medium/high SES children. Thus, higher levels of externalizing behavior problems were needed for caregivers of low SES children to select higher response options for these items. Item BPI 4 (“Cheats/lies”), however, was easier for low SES children than for medium/high SES children. In addition, DIF by SES was detected at the uncorrected $p < .05$ level of significance in five items. Item PSC-17 10 (“Blames others”) was easier for low SES children compared to medium/high SES children, while item BPI 10 (“Disobedient at

Table 15

Differential Item Functioning in Combined Externalizing Subscale Items by Child Sex

Item	Short Wording	IRT-LR			OLR	
		Overall DIF	<i>a</i> -DIF	<i>b</i> -DIF	Uniform	Non-Uniform
		G^2 (3 <i>df</i>)	G^2 (1 <i>df</i>)	G^2 (2 <i>df</i>)	OR (95% CI) ^a	β_3
PSC-17 4	Refuses to share	8.5*	0.1	8.4*	1.61 (1.21, 2.15)**	-0.06
PSC-17 5	Does not understand others' feelings	7.5	0.7	6.8*	0.71 (0.54, 0.94)*	0.19
PSC-17 8	Fights others	1.6	-	-	1.09 (0.79, 1.48)	-0.11
PSC-17 10	Blames others	2.1	-	-	0.85 (0.64, 1.13)	-0.06
PSC-17 12	Does not listen to rules	4.6	0.2	4.4	0.71 (0.51, 0.98)*	0.14
PSC-17 14	Teases others	1.6	-	-	1.06 (0.79, 1.42)	0.10
PSC-17 16	Takes things	2.4	-	-	1.18 (0.88, 1.56)	-0.14
BPI 3	High strung	1.2	-	-	0.92 (0.68, 1.24)	-0.01
BPI 4	Cheats/lies	1.7	-	-	1.11 (0.84, 1.46)	-0.15
BPI 6	Argues too much	3.2	-	-	1.27 (0.97, 1.67)	-0.10

(table continues)

Table 15 (continued)

Item	Short Wording	IRT-LR			OLR	
		Overall DIF	<i>a</i> -DIF	<i>b</i> -DIF	Uniform	Non-Uniform
		G^2 (3 <i>df</i>)	G^2 (1 <i>df</i>)	G^2 (2 <i>df</i>)	OR (95% CI) ^a	β_3
BPI 9	Bullies/cruel or mean	3.8	-	-	1.01 (0.72, 1.41)	0.12
BPI 10	Disobedient at home	0.7	-	-	0.91 (0.68, 1.22)	0.07
BPI 11	Not sorry after misbehaves	1.3	-	-	0.91 (0.67, 1.24)	-0.17
BPI 12	Trouble getting along with others	0.5	-	-	1.14 (0.81, 1.61)	0.00
BPI 15	Not liked by others	0.9	-	-	1.23 (0.76, 1.97)	-0.05
BPI 18	Stubborn, sullen, or irritable	6.8	1.7	5.1	1.41 (1.06, 1.86)*	0.07
BPI 19	Very strong temper	0.2	-	-	1.00 (0.75, 1.35)	-0.12
BPI 22	Breaks/destroys things	14.9	4.9	9.9*	0.57 (0.40, 0.80)**	0.29

Note. IRT-LR tests were conducted using IRTLRDIF (Thissen, 2001) freeware. This program does not conduct likelihood ratio tests for *a*- and *b*-DIF if the overall DIF test yields a G^2 statistic < 3.84; therefore, cells in this situation are empty. IRT-LR = likelihood-based model method; OLR = ordinal logistic regression method; DIF = differential item functioning; *a*-DIF = discrimination parameter DIF; *b*-DIF = difficulty parameters DIF; OR = odds ratio; CI = confidence interval; β_3 = beta coefficient for the interaction term for theta and group membership.

^aOdds ratios are presented only for the focal group (females), as reference group (males) odds ratios are equal to 1. The odds ratio represents how many times higher or lower the odds are for female children than for male children in selecting a higher versus lower response option for a given item, controlling for level of externalizing behavior problems.

* $p < .05$. ** $p < .0027$, denoting statistical significance after a Bonferroni correction for multiple comparisons.

Table 16

Differential Item Functioning in Combined Externalizing Subscale Items by Child Race

Item	Short Wording	IRT-LR			OLR ^a	
		Overall DIF	<i>a</i> -DIF	<i>b</i> -DIF	Uniform	Non-Uniform
		G^2 (3 <i>df</i>)	G^2 (1 <i>df</i>)	G^2 (2 <i>df</i>)	OR (95% CI) ^b	β_3
PSC-17 4	Refuses to share	10.3*	0.0	10.3*	1.18 (0.87, 1.58)	0.20
PSC-17 5	Does not understand others' feelings	8.3*	5.5*	2.8	0.71 (0.53, 0.95)*	-0.38*
PSC-17 8	Fights others	5.8	3.9*	1.9	1.17 (0.84, 1.62)	-0.13
PSC-17 10	Blames others	11.7*	6.0*	5.7	0.92 (0.68, 1.24)	0.60**
PSC-17 12	Does not listen to rules	5.6	3.3	2.3	0.81 (0.57, 1.13)	-0.13
PSC-17 14	Teases others	26.2**	0.2	25.9**	1.94 (1.43, 2.64)**	0.00
PSC-17 16	Takes things	9.5*	0.1	9.3*	1.43 (1.06, 1.93)*	0.00
BPI 3	High strung	13.6*	0.3	13.3**	0.58 (0.42, 0.80)**	0.09
BPI 4	Cheats/lies	4.7	0.5	4.2	1.02 (0.76, 1.34)	0.33
BPI 6	Argues too much	25.2**	0.4	24.7**	0.42 (0.31, 0.56)**	0.06

(table continues)

Table 16 (continued)

Item	Short Wording	IRT-LR			OLR ^a	
		Overall DIF	<i>a</i> -DIF	<i>b</i> -DIF	Uniform	Non-Uniform
		<i>G</i> ² (3 <i>df</i>)	<i>G</i> ² (1 <i>df</i>)	<i>G</i> ² (2 <i>df</i>)	OR (95% CI) ^b	β_3
BPI 9	Bullies/cruel or mean	9.9*	0.4	9.5*	1.54 (1.08, 2.19)*	0.00
BPI 10	Disobedient at home	5.6	1.4	4.2	0.84 (0.62, 1.15)	-0.04
BPI 11	Not sorry after misbehaves	10.6*	2.0	8.6*	1.26 (0.92, 1.73)	-0.33
BPI 12	Trouble getting along with others	7.4	0.6	6.8*	1.26 (0.88, 1.80)	-0.10
BPI 15	Not liked by others	4.2	3.6	0.6	1.18 (0.72, 1.92)	-0.53
BPI 18	Stubborn, sullen, or irritable	5.2	0.0	5.2	0.81 (0.61, 1.08)	-0.03
BPI 19	Very strong temper	5.2	1.0	4.2	0.71 (0.52, 0.97)*	-0.24
BPI 22	Breaks/destroys things	11.4*	0.1	11.3*	1.55 (1.08, 2.22)*	-0.01

Note. IRT-LR = likelihood-based model method; OLR = ordinal logistic regression method; DIF = differential item functioning; *a*-DIF = discrimination parameter DIF; *b*-DIF = difficulty parameters DIF; OR = odds ratio; CI = confidence interval; β_3 = beta coefficient for the interaction term for theta and group membership.

^aOLR analyses controlled for effects of SES. ^bOdds ratios are presented only for the focal group (minority children), as reference group (white children) odds ratios are equal to 1. The odds ratio represents how many times higher or lower the odds are for minority children than for white children in selecting a higher versus lower response option for a given item, controlling for level of externalizing behavior problems and SES.

* $p < .05$. ** $p < .0027$, denoting statistical significance after a Bonferroni correction for multiple comparisons.

Table 17***Differential Item Functioning in Combined Externalizing Subscale Items by Child Socioeconomic Status***

Item	Short Wording	IRT-LR			OLR ^a	
		Overall DIF	<i>a</i> -DIF	<i>b</i> -DIF	Uniform	Non-Uniform
		G^2 (3 <i>df</i>)	G^2 (1 <i>df</i>)	G^2 (2 <i>df</i>)	OR (95% CI) ^b	β_3
PSC-17 4	Refuses to share	2.3	-	-	0.88 (0.64, 1.19)	-0.17
PSC-17 5	Does not understand others' feelings	12.3*	8.3*	4.0	0.95 (0.71, 1.28)	-0.33
PSC-17 8	Fights others	13.8*	8.2*	5.7	1.03 (0.74, 1.45)	-0.11
PSC-17 10	Blames others	9.7*	1.3	8.4*	1.08 (0.80, 1.46)	0.50*
PSC-17 12	Does not listen to rules	4.0	2.1	1.9	0.96 (0.67, 1.36)	-0.02
PSC-17 14	Teases others	3.6	-	-	1.21 (0.89, 1.65)	0.15
PSC-17 16	Takes things	2.1	-	-	0.99 (0.73, 1.34)	-0.18
BPI 3	High strung	13.8*	0.7	13.0**	0.64 (0.46, 0.89)*	0.11
BPI 4	Cheats/lies	12.3	0.2	12.1**	1.21 (0.90, 1.63)	0.45*
BPI 6	Argues too much	4.9	1.3	3.6	1.02 (0.76, 1.37)	0.02

(table continues)

Table 17 (continued)

Item	Short Wording	IRT-LR			OLR ^a	
		Overall DIF	<i>a</i> -DIF	<i>b</i> -DIF	Uniform	Non-Uniform
		G^2 (3 <i>df</i>)	G^2 (1 <i>df</i>)	G^2 (2 <i>df</i>)	OR (95% CI) ^b	β_3
BPI 9	Bullies/cruel or mean	3.3	-	-	1.07 (0.75, 1.53)	-0.13
BPI 10	Disobedient at home	10.5*	0.9	9.6*	0.70 (0.51, 0.96)*	0.13
BPI 11	Not sorry after misbehaves	2.2	-	-	1.17 (0.85, 1.61)	-0.08
BPI 12	Trouble getting along with others	2.7	-	-	1.17 (0.81, 1.69)	-0.05
BPI 15	Not liked by others	0.4	-	-	1.08 (0.66, 1.78)	0.16
BPI 18	Stubborn, sullen, or irritable	24.4**	7.2*	17.2**	0.63 (0.47, 0.85)*	-0.40*
BPI 19	Very strong temper	1.1	-	-	1.11 (0.81, 1.52)	-0.24
BPI 22	Breaks/destroys things	2.0	-	-	1.29 (0.90, 1.85)	-0.06

Note. IRT-LR tests were conducted using IRTLRDIF (Thissen, 2001) freeware. This program does not conduct likelihood ratio tests for *a*- and *b*-DIF if the overall DIF test yields a G^2 statistic < 3.84; therefore, cells in this situation are empty. IRT-LR = likelihood-based model method; OLR = ordinal logistic regression method; DIF = differential item functioning; *a*-DIF = discrimination parameter DIF; *b*-DIF = difficulty parameters DIF; OR = odds ratio; CI = confidence interval; β_3 = beta coefficient for the interaction term for theta and group membership.

^aOLR analyses controlled for effects of race. ^bOdds ratios are presented only for the focal group (low SES children), as reference group (medium/high SES children) odds ratios are equal to 1. The odds ratio represents how many times higher or lower the odds are for low SES children than for medium/high SES children in selecting a higher versus lower response option for a given item, controlling for level of externalizing behavior problems and SES.

* $p < .05$. ** $p < .0027$, denoting statistical significance after a Bonferroni correction for multiple comparisons.

home”) exhibited difficulty DIF in the other direction. Finally, items PSC-17 5 (“Does not understand others’ feelings”), PSC-17 8 (“Fights others”), and BPI 18 (“Stubborn, sullen, or irritable”) were all more discriminating for medium/high SES children than for low SES children.

In summary, using the stringent Bonferroni-corrected significance level of $p < .0027$, the IRT-LR technique revealed five items with significant DIF: no items with DIF by child sex, two items by child race, two items by child SES, and one item by both child race and child SES. Each identified item demonstrated DIF in the difficulty parameters; no significant discrimination parameter DIF was detected using the Bonferroni-corrected criterion. Of the items displaying DIF only by race, item BPI 6 (“Argues too much”) was easier for white children than for minority children, while item PSC-17 14 (“Teases others”) exhibited the reverse effect. Of the items demonstrating DIF only by SES, item BPI 18 (“Stubborn, sullen, or irritable”) was more difficult for low SES children than for medium/high SES children, while item BPI 4 (“Cheats/lies”) had the opposite effect. The remaining item exhibited DIF by both race and SES: item BPI 3 (“High strung”) was more difficult for minority and low SES children than for white and medium/high SES children. Several other items were identified displaying DIF in the difficulty and discrimination parameters for groups differing by child sex, race, and SES when an uncorrected level of significance of $p < .05$ was utilized; however, the validity of these results is uncertain due to the likelihood of inflated Type I error attributable to multiple comparisons. To provide additional information regarding potential DIF and to assess possible replication of the findings from the IRT-LR method, an alternative technique was used: ordinal logistic regression.

OLR. As described in Chapter IV, the OLR approach also tested items for DIF by child sex, race, and SES. In this assessment of DIF, group membership was evaluated as to whether it affected the relationship between theta (in this case, level of externalizing behavior problems, obtained via IRT scoring) and response to a given item (i.e., choice of 0, 1, or 2 by the caregiver). Non-uniform DIF, analogous to effect modification, was assessed by considering the statistical significance of the interaction term (β_3) for theta and group membership in the following ordinal logistic regression equation, in which the left-hand term is the cumulative logit:

$$\text{clogit}(\text{item response}) = \alpha_i + \beta_1(\text{theta}) + \beta_2(\text{group}) + \beta_3(\text{theta*group}) \quad i = 0, 1. \quad (6)$$

If the interaction term in Equation 6 was statistically significant, then group membership affected the relationship between level of externalizing behavior problems and response to a given item. Uniform DIF, analogous to confounding, was evaluated by considering the statistical significance of the main effect of group membership (β_2) in the following ordinal logistic regression equation, in which the left-hand term is the cumulative logit:

$$\text{clogit}(\text{item response}) = \alpha_i + \beta_1(\text{theta}) + \beta_2(\text{group}) \quad i = 0, 1. \quad (7)$$

In considering the relevant effects in both the non-uniform and uniform models, a Bonferroni correction for multiple comparisons (.05/18) was implemented to preserve overall α at the .05 level, requiring $p < .0027$ for significance. Finally, proportional odds

ratios were computed from the group membership main effect coefficients to assist with interpretation of uniform DIF. Results of the OLR method are summarized in Tables 15 (by sex), 16 (by race), and 17 (by SES).

For groups defined by child sex, no non-uniform DIF was detected. However, significant uniform DIF ($p < .0027$) was found in two items. Item PSC-17 4 (“Refuses to share”) was more difficult for boys, as the odds of selecting a higher versus lower response option were over 60% higher for caregivers of girls than boys, controlling for level of externalizing behavior problems. In contrast, item BPI 22 (“Breaks/destroys things”) was easier for boys, with girls’ caregivers having lower odds of selecting a higher versus lower response option than boys’ caregivers. Three additional items displayed DIF by sex at the uncorrected $p < .05$ level of significance: items PSC-17 5 (“Does not understand others’ feelings”) and PSC-17 12 (“Does not listen to rules”) were easier for boys, while item BPI 18 (“Stubborn, sullen, or irritable”) was easier for girls.

For groups defined by child race¹, non-uniform DIF was detected in item PSC-17 10 (“Blames others”) at the stringent Bonferroni-corrected level of significance, while controlling for SES. At low levels of externalizing behavior problems, caregivers of both white and minority children were most likely to select *never* for this item. At the mean level of externalizing behavior problems, caregivers of minority children still tended to select *never*, while caregivers of white children were more likely to select *sometimes*. However, at high levels of externalizing behavior problems, caregivers of white children still tended to select *sometimes*, while caregivers of minority children were more likely to

¹ Analyses controlled for child SES. Caregiver race was not controlled due to small cell sizes (i.e., only 47 caregivers were of a different race than their children). However, OLR results were highly similar in analyses conducted only with cases in which caregiver and child race matched: the same items were identified with significant DIF either way.

select *always*. Non-uniform DIF was also detected at the $p < .05$ level of significance for item PSC-17 5 (“Does not understand others’ feelings”), with caregivers of white and minority children demonstrating similar response patterns at low and mean levels of externalizing behavior problems, but caregivers of white children being more likely to select *always* than caregivers of minority children at high levels of externalizing behavior problems.

In addition, three items displayed significant uniform DIF by race at the Bonferroni-corrected level of significance, controlling for SES: items PSC-17 14 (“Teases others”), BPI 3 (“High strung”), and BPI 6 (“Argues too much”). Compared to caregivers of white children, caregivers of minority children had nearly twice the odds of endorsing higher response options to item PSC-17 14. For items BPI 3 and BPI 6, however, the direction of the group effect was reversed, as these items were easier for caregivers of white children. Five additional items displayed DIF by race at the less stringent $p < .05$ level of significance: items PSC-17 5 (“Does not understand others’ feelings”) and BPI 19 (“Very strong temper”) were easier for white children, while items PSC-17 16 (“Takes things”), BPI 9 (“Bullies/cruel or mean”), and BPI 22 (“Breaks/destroys things”) were easier for minority children.

For groups defined by child SES², no non-uniform DIF was detected at the Bonferroni-corrected level of significance, controlling for race. However, three items were detected with non-uniform DIF by SES at the $p < .05$ level of significance, controlling for race: items PSC-17 10 (“Blames others”), BPI 4 (“Cheats/lies”), and BPI 18 (“Stubborn, sullen, or irritable”). In all three items, caregivers of low SES and medium/high SES children demonstrated similar response patterns at low and mean

² Analyses controlled for child race. See Footnote 1 regarding consideration of caregiver race.

levels of externalizing behavior problems. At high levels of externalizing behavior problems, however, caregivers of low SES children were much more likely to select *always* for items PSC-17 10 and BPI 4 than were caregivers of medium/high SES, while the pattern was reversed for item BPI 18.

Similarly, no uniform DIF by SES was found at the stringent Bonferroni-corrected level of significance, controlling for race. Three items, however, displayed uniform DIF by SES, controlling for race, at the $p < .05$ level of significance. Items BPI 3 (“High strung”), BPI 10 (“Disobedient at home”), and BPI 18 (“Stubborn, sullen, or irritable”) all demonstrated similar effects of SES: All three items were easier for caregivers of medium/high SES children, with caregivers of low SES children having lower odds of selecting higher versus lower response options than caregivers of medium/high SES children.

In summary, using the stringent Bonferroni-corrected significance level of $p < .0027$, the OLR technique revealed only one item with significant non-uniform DIF and five items with significant uniform DIF, including two items by child sex and three items by child race. No significant DIF by child SES was detected using the Bonferroni-corrected criterion. Item PSC-17 10 (“Blames others”) was the only item demonstrating significant non-uniform DIF, in which the relationship between item responses and child race changed as level of externalizing behavior problems increased, while controlling for SES. Of the items displaying uniform DIF by sex, item PSC-17 4 (“Refuses to share”) was more difficult for boys than girls, while item BPI 22 (“Breaks/destroys things”) was more difficult for girls than boys. Of the items demonstrating DIF by race, controlling for SES, item PSC-17 14 (“Teases others”) was more difficult for white children than

minority children, while items BPI 3 (“High strung”) and BPI 6 (“Argues too much”) were more difficult for minority children than white children. As with the IRT-LR method, the OLR method identified several other items displaying non-uniform and uniform DIF for groups differing by child sex, race, and SES when an uncorrected level of significance of $p < .05$ was utilized; however, false positive results at this level of significance were possible due to multiple comparisons. To further evaluate the status of each item in the combined externalizing subscale with regard to DIF, results from the OLR approach were compared to those from the IRT-LR method.

Comparisons of DIF findings. Table 18 presents a summary of the findings of both the IRT-LR method and the OLR approach, highlighting the types and levels of significance of DIF detected in each item. Only one item was completely free of DIF in all analyses: item BPI 15 (“Not liked by other children”). In several items, however, DIF was detected only at the $p < .05$ level of significance, and only by a single method. For example, item PSC-17 12 (“Does not listen to rules”) appeared to demonstrate uniform DIF by sex as detected by the OLR method, but not by the IRT-LR approach; items BPI 11 (“Not sorry after misbehaves”) and BPI 12 (“Trouble getting along with others”) displayed DIF in the difficulty parameters by race per the IRT-LR approach; and item BPI 19 (“Very strong temper”) demonstrated uniform DIF by race, as detected by the OLR method. In addition, item PSC-17 8 (“Fights others”) displayed DIF in the discrimination parameter in comparisons of groups by race as well as by SES; however, these findings were detected solely using the IRT-LR approach and were only at the $p < .05$ level of significance.

Table 18

Comparison of Results of DIF Detection by Two Methods

Item	Sex				Race				SES			
	IRT-LR		OLR		IRT-LR		OLR ^a		IRT-LR		OLR ^a	
	<i>a</i> -DIF	<i>b</i> -DIF	Unif	Non-U	<i>a</i> -DIF	<i>b</i> -DIF	Unif	Non-U	<i>a</i> -DIF	<i>b</i> -DIF	Unif	Non-U
PSC-17 4		○	●			○						
PSC-17 5		○	○		○		○	○	○			
PSC-17 8					○				○			
PSC-17 10					○			●		○		○
PSC-17 12			○									
PSC-17 14						●	●					
PSC-17 16						○	○					
BPI 3						●	●			●	○	
BPI 4										●		○
BPI 6						●	●					

(table continues)

Table 18 (continued)

Item	Sex				Race				SES			
	IRT-LR		OLR		IRT-LR		OLR ^a		IRT-LR		OLR ^a	
	<i>a</i> -DIF	<i>b</i> -DIF	Unif	Non-U	<i>a</i> -DIF	<i>b</i> -DIF	Unif	Non-U	<i>a</i> -DIF	<i>b</i> -DIF	Unif	Non-U
BPI 9						○	○					
BPI 10										○		○
BPI 11						○						
BPI 12						○						
BPI 15												
BPI 18			○						○	●	○	○
BPI 19							○					
BPI 22		○	●			○	○					

Note. DIF = Differential item functioning. SES = socioeconomic status; IRT-LR = likelihood-based model method; OLR = ordinal logistic regression approach; *a*-DIF = discrimination parameter DIF; *b*-DIF = difficulty parameters DIF; Unif = uniform DIF; Non-U = non-uniform DIF.

^aOLR analyses investigating race controlled for SES, and those investigating SES controlled for race. DIF is reported only when group membership of interest remained significant after controlling for the relevant covariate.

● = significant DIF detected after implementation of a Bonferroni correction, adjusted for multiple analyses of 18 items ($p < .0027$).

○ = significant DIF detected with no Bonferroni correction ($p < .05$). Both levels of significance are included due to inconsistencies in the literature regarding the necessity of correction for multiple analyses in DIF detection.

Several items were identified with DIF at the uncorrected $p < .05$ level of significance by both DIF-detection methods. Items PSC-17 16 (“Takes things”), BPI 9 (“Bullies/cruel or mean”), and BPI 10 (“Disobedient at home”) each displayed DIF in difficulty parameters, detected by both the IRT-LR approach and the OLR technique; the former two exhibited DIF by child race, and the latter by child SES. However, these findings were not significant at the Bonferroni-corrected level of $p < .0027$. Similarly, item PSC-17 5 (“Does not understand others’ feelings”) was found to demonstrate several types of DIF: *b*-DIF and uniform DIF were detected by child sex by both methods; *a*-DIF and non-uniform DIF by child race by both methods; uniform DIF by child race via the OLR approach; and *a*-DIF by child SES by the IRT-LR method.

The remaining items each exhibited at least one type of significant DIF after adjustment for multiple comparisons. Five items exhibited only one type of DIF detected at the Bonferroni-corrected level of significance, while three items demonstrated either multiple types of significant DIF or consistent findings of significant DIF by both methods. Of the five items with one type of significant DIF, item PSC-17 4 (“Refuses to share”) exhibited uniform DIF detected by the OLR approach, requiring higher levels of externalizing behavior problems among boys than girls for higher response options to be selected by caregivers. Item PSC-17 10 (“Blames others”) displayed non-uniform DIF by child race per the OLR approach, with the caregivers of minority children selecting higher response options than the caregivers of white children only at higher levels of externalizing behavior problems. Items BPI 4 and BPI 18 both exhibited DIF in difficulty parameters detected by the IRT-LR approach: BPI 4 (“Cheats/lies”) had a higher upper threshold for medium/high SES children than for low SES children, while BPI 18

(“Stubborn, sullen, or irritable”) was more difficult for low versus medium/high SES children. Item BPI 22 (“Breaks/destroys things on purpose”) demonstrated uniform DIF by sex as detected by the OLR approach, requiring higher levels of externalizing behavior problems in girls than in boys for caregivers to select higher response options. Each of these five items also exhibited at least one additional type of DIF at the less stringent level of significance.

The three remaining items demonstrated significant DIF at the Bonferroni-corrected level of significance duplicated by both methods. Item PSC-17 14 (“Teases others”) exhibited differing difficulty parameters by race, according to both DIF-detection approaches. This item required higher levels of externalizing behavior problems among white children than minority children for caregivers to select higher response options. In contrast, according to both DIF-detection methods, items BPI 3 (“High strung”) and BPI 6 (“Argues too much”) demonstrated DIF in difficulty parameters by race in the opposite direction: lower levels of externalizing behavior problems were necessary among white children than minority children for caregivers to endorse higher response options. Finally, item BPI 3 also exhibited significant DIF at the Bonferroni-corrected level of significance by child SES: Higher levels of externalizing behavior problems were required in low SES children than medium/high SES children for caregivers to select higher response options. These three items, combined with the five items demonstrating significant DIF detected by a single method, were examined further to enable interpretation of the meaning and effects of the DIF in the context of screening for externalizing behavior problems.

Extent of DIF effects. The extent to which DIF affects an item's measurement performance can be assessed in several ways, including (a) considering effect sizes of differences in an item's parameter estimates by group; (b) visually comparing plots of the item's OCCs representing each group of interest; and (c) assessing changes in IRT-based scores for each group after adjusting item parameters for DIF. Each of these methods was used to examine the extent of DIF present in the eight items exhibiting statistically significant DIF according to the stringent Bonferonni-corrected criterion.

First, the externalizing subscale items were recalibrated by fitting the GRM while allowing the affected parameters of the eight items identified with significant DIF to differ by relevant groups. Tables 19, 20, and 21 present the item parameter estimates obtained for each set of group comparisons, allowing consideration of the direction and size of the effects detected in the DIF analyses. Mean differences in difficulty parameter estimates ranged from 0.14 to 0.72 standard deviations in magnitude.

Next, the recalibrated parameter estimates described above were used to plot the OCCs for the eight items in question by group. Visual examination of these plots assisted with interpretation of the extent and effects of DIF detected in each item. The plots are presented in Figures 11, 12, and 13. Greater differences between OCCs were synonymous with larger differences in item parameters, as discussed above.

Finally, changes were examined in IRT-based scores for each group after adjusting item parameters for DIF. Paired *t*-tests were used to compare theta scores generated before the parameters of the eight items of concern were adjusted for DIF to theta scores obtained after recalibration. For the sample as a whole, no significant differences were found between the unadjusted ($M = -0.05$, $SD = 0.92$) and DIF-adjusted

($M = -0.05$, $SD = 0.91$) theta score estimates, $t(899) = -0.09$, $p = 0.93$. However, when the sample was split into the relevant groups of interest, several significant differences were observed. Results for these analyses are presented in Table 22. While no significant differences were detected in adjusted versus DIF-adjusted theta score estimates within groups of male or female children, significant differences were found within groups of white, minority, low SES, and medium/high SES children.

Table 19***Graded Response Model Item Parameter Estimates Adjusted for Items Displaying DIF by Child Sex***

Item	Short Wording	Male			Female		
		<i>a (se)</i>	<i>b₁ (se)</i>	<i>b₂ (se)</i>	<i>a (se)</i>	<i>b₁ (se)</i>	<i>b₂ (se)</i>
PSC-17 4	Refuses to share	1.30 (0.08)	-0.90 (0.12)	1.95 (0.17)	1.30 (0.08)	-1.35 (0.13)	1.81 (0.18)
BPI 22	Breaks/destroys things	1.90 (0.14)	0.47 (0.08)	1.75 (0.14)	1.90 (0.14)	0.78 (0.10)	2.07 (0.18)

Note. DIF = differential item functioning.

Table 20***Graded Response Model Item Parameter Estimates Adjusted for Items Displaying DIF by Child Race***

Item	Short Wording	White			Minority		
		<i>a (se)</i>	<i>b₁ (se)</i>	<i>b₂ (se)</i>	<i>a (se)</i>	<i>b₁ (se)</i>	<i>b₂ (se)</i>
PSC-17 10	Blames others	1.24 (0.18)	-0.18 (0.12)	2.30 (0.30)	1.75 (0.21)	0.03 (0.09)	1.58 (0.18)
PSC-17 14	Teases others	1.38 (0.10)	0.32 (0.11)	2.86 (0.28)	1.38 (0.10)	-0.10 (0.10)	2.20 (0.21)
BPI 3	High strung	1.15 (0.07)	0.33 (0.13)	2.20 (0.20)	1.15 (0.07)	0.95 (0.14)	2.68 (0.23)
BPI 6	Argues too much	1.50 (0.09)	-0.84 (0.11)	1.10 (0.12)	1.50 (0.09)	-0.22 (0.10)	1.60 (0.15)

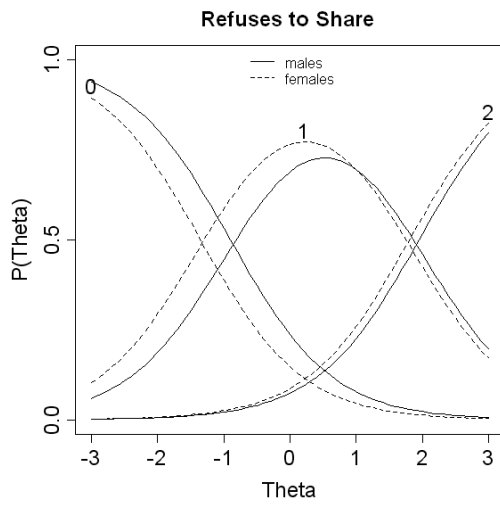
Note. DIF = differential item functioning.

Table 21*Graded Response Model Item Parameter Estimates Adjusted for Items Displaying DIF by Child Socioeconomic Status*

Item	Short Wording	Low SES			Medium/High SES		
		<i>a (se)</i>	<i>b₁ (se)</i>	<i>b₂ (se)</i>	<i>a (se)</i>	<i>b₁ (se)</i>	<i>b₂ (se)</i>
BPI 3	High strung	1.24 (0.08)	0.84 (0.14)	2.54 (0.22)	1.24 (0.08)	0.40 (0.11)	2.05 (0.19)
BPI 4	Cheats/lies	1.26 (0.08)	-0.38 (0.13)	1.79 (0.18)	1.26 (0.08)	-0.38 (0.11)	2.50 (0.24)
BPI 18	Stubborn, sullen, or irritable	1.55 (0.09)	-0.69 (0.11)	1.48 (0.15)	1.55 (0.09)	-1.03 (0.10)	1.21 (0.13)

Note. DIF = differential item functioning. SES = Child socioeconomic status.

(a) PSC-17 4



(b) BPI 22

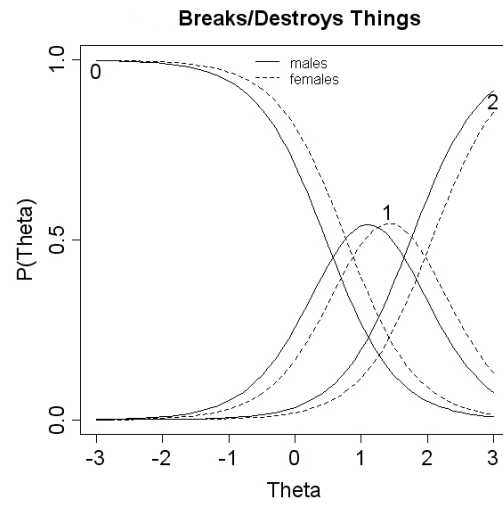
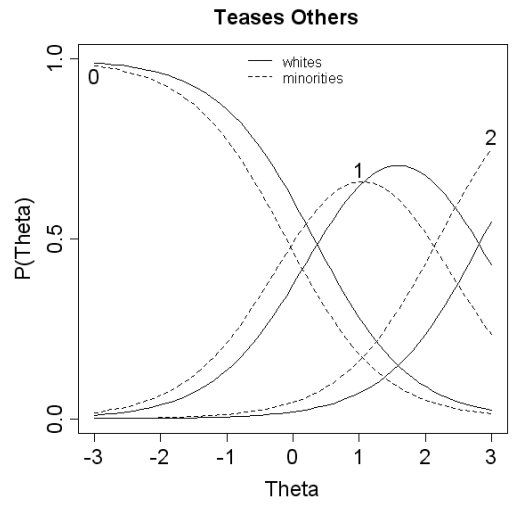


Figure 11. Plots of graded response model option characteristic curves (OCCs) by group for items exhibiting differential item functioning by child sex.

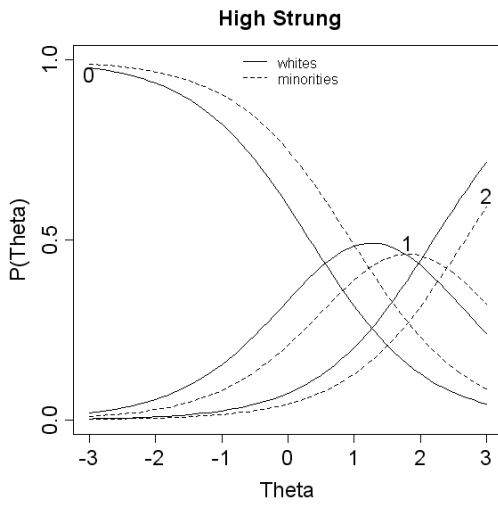
(a) PSC-17 10



(b) PSC-17 14



(c) BPI 3



(d) BPI 6

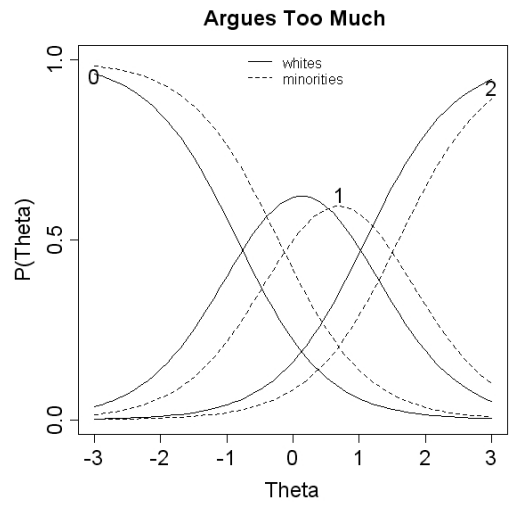
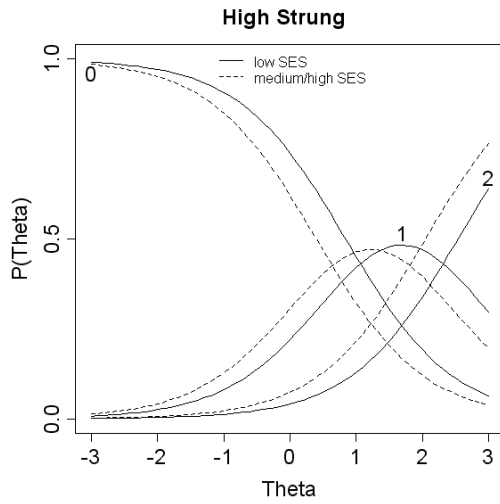
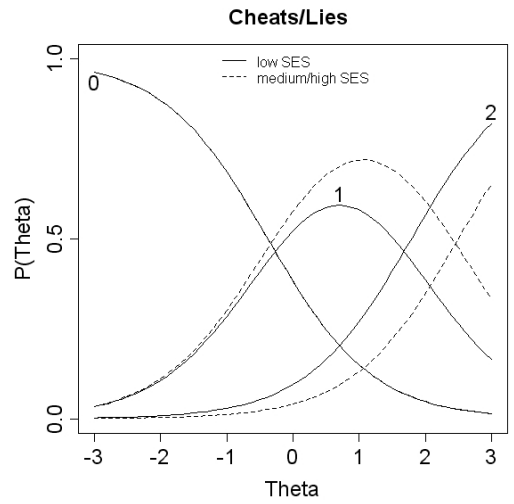


Figure 12. Plots of graded response model option characteristic curves (OCCs) by group for items exhibiting differential item functioning by child race.

(a) BPI 3



(b) BPI 4



(c) BPI 18

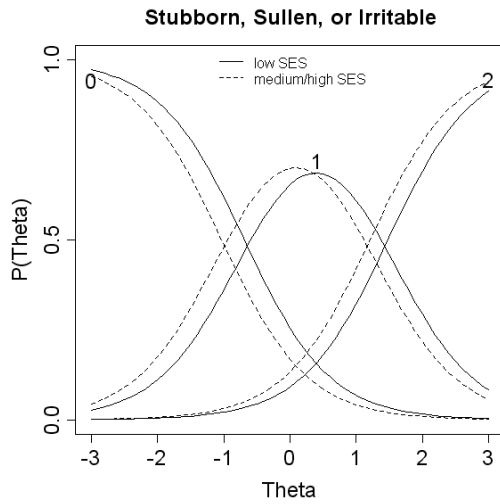


Figure 13. Plots of graded response model option characteristic curves (OCCs) by group for items exhibiting differential item functioning by child socioeconomic status.

Table 22***Differences in Unadjusted and DIF-Adjusted Theta Score Estimates within Sociodemographic Groups***

Group	N	Unadjusted Theta Score		DIF-Adjusted Theta Score		t	df	p
		M	SD	M	SD			
Male	472	-0.00	0.97	0.00	0.96	-0.62	471	.53
Female	424	-0.11	0.85	-0.12	0.86	0.52	423	.60
White	450	-0.04	0.88	-0.06	0.88	13.20	449	< .001
Minority	450	-0.06	0.96	-0.04	0.94	-9.88	449	< .001
Low SES	371	0.14	0.99	0.15	0.97	-8.72	370	< .001
Med/High SES	501	-0.19	0.84	-0.20	0.84	9.36	450	< .001

Note. Paired *t*-tests were conducted on unadjusted and DIF-adjusted theta score estimates previously obtained by fitting the graded response model with MULTILOG 7.03 (Thissen, 2003) software. DIF = differential item functioning. SES = socioeconomic status.

CHAPTER VI

DISCUSSION

Externalizing behavior problems in preschool-aged children are associated with a range of negative long-term social and public health consequences (Hann & Borek, 2001). Primary and secondary prevention efforts aimed at early identification may reduce these unfavorable outcomes (Hoagwood & Johnson, 2003). Mental health screening in pediatric primary care has been shown to be effective in increasing referrals to and uptake of mental health services (Lavigne, Arend, Rosenbaum, Binns, Christoffel, Burns, et al., 1998); thus, improved early identification of externalizing behavior problems in the pediatric primary care setting may decrease the prevalence of these problems and their associated outcomes (AHRQ, 2002). To identify accurately those children in need of further evaluation and intervention, brief screening measures are needed which (a) precisely measure behavior problems at clinical and sub-clinical levels, and (b) perform consistently across populations of very young children. Use of unbiased screening instruments could contribute to the elimination of sociodemographic disparities in identification of children with externalizing behavior problems.

The purpose of this study was to evaluate the measurement properties of items measuring externalizing behaviors in two commonly-used parent-report questionnaires: the PSC-17 (Gardner et al., 1999) and the BPI (Peterson & Zill, 1986; Zill, 1990). The target population included caregivers with preschool-aged children seen in primary care

practices. A cross-sectional survey design was utilized. Data were collected via pediatric primary care waiting rooms, where primary caregivers (i.e., parents/guardians) of 900 children between the ages of 3 and 5 from diverse socioeconomic and racial backgrounds completed the PSC-17, the BPI, and a sociodemographic questionnaire. IRT analyses allowed the identification of items which best measured clinical and sub-clinical levels of externalizing behavior problems in young children, as well as those which demonstrated measurement bias across groups who differed by child sex, race, and SES. IRT analyses were particularly well-suited for this investigation because they permitted (a) evaluation of the amount of measurement information offered by each item along the spectrum of externalizing behavior problems, and (b) scrutiny of item-level bias (DIF) not detectable with CTT methodologies.

In this discussion, the findings from two research questions are summarized, addressing the measurement properties of 18 PSC-17 and BPI items assessing externalizing behavior problems. Item content is examined as related to the findings from each research question. Results are integrated to identify a set of items most promising for use in screening very young children for externalizing behavior problems in diverse pediatric primary care settings. Implications of results, limitations of the study, and directions for future research are also addressed. As a preface to this discussion, the current results of traditional CTT analyses are reviewed to place this study in the context of the existing literature regarding the psychometric properties of the PSC-17 and BPI.

Scale Performance in Context: Classical Test Theory Analyses

Traditional psychometric analyses were conducted to compare findings regarding CTT reliability and validity to previous reports in the literature. In this way, the

comparability of the current scale performance of the PSC-17 and BPI to existing studies could be evaluated. Regarding traditional summed scores for the PSC-17, the BPI, and their respective externalizing subscales, all means and standard deviations observed were similar to those reported in previous studies (Gardner et al., 1999; Gortmaker et al., 1990; Zill, 1990). Additionally, reliability coefficients—including Cronbach's α , inter-item correlations, and corrected item-total correlations—were similar to those reported in previous CTT psychometric analyses of each instrument (Gardner et al., 1999; Gortmaker et al., 1990; Zill, 1990).

An examination of concurrent validity using Pearson correlations between the externalizing subscales of the PSC-17 and BPI suggested that, as anticipated, both instruments measured the same constructs. In addition, investigations of known groups validity involved comparisons of mean total scale and subscale scores for each instrument between (a) those caregivers who believed versus did not believe that their child had a behavior problem, and (b) those caregivers who reported that their child had versus had not received mental health services for behavior problems. These comparisons revealed significant differences in total scale and subscale scores between each pair of groups, supporting the previously reported known groups validity of these instruments (Gardner et al., 1999; Gortmaker et al., 1990; Zill, 1990).

Finally, differences in mean total scale PSC-17 and BPI scores by child sex, race, and SES were assessed. Significant differences in mean scores by child sex were detected, consistent with previous reports in the literature describing higher mean scores among male children compared to female children (Jellinek et al., 1999; Parcel & Menaghan, 1988). In contrast, differences in mean PSC-17 and BPI scores were not

found between white and minority children, a finding incongruent with previous studies (Jutte et al., 2003; Simonian & Tarnowski, 2001; Simonian et al., 1991; Spencer et al., 2005). However, studies previously reporting disparities in mean scores by race have either failed to control for SES in analyses (e.g., Spencer et al., 2005), or have included only low income children in their samples (e.g., Jutte et al., 2003; Simonian & Tarnowski, 2001), precluding consideration of possible confounding effects of SES. The present results regarding significant differences in mean scores between low and medium/high SES children, however, were consistent with previous studies identifying the effects of SES on scale scores (Jellinek et al., 1995; Jellinek et al., 1999).

With the exception of the lack of significant differences in mean scores between white and minority children, the PSC-17 and BPI total scale and externalizing subscales appeared to perform similarly to previous investigations. Distributional properties and indicators of reliability and validity suggested that the current performance of these instruments—as evaluated with CTT methods—was congruent with prior studies, accentuating the implications of the findings for the two research questions employing IRT analyses.

Research Question 1: Precision and Utility of Measurement

The investigation of the precision and utility of items in the PSC-17 and BPI for measurement of externalizing behavior problems among very young children involved estimation of each item's difficulty and discrimination parameters, as well as assessment of the measurement information offered by each item along the continuum of the latent construct. Samejima's (1969) GRM was fit to the data to obtain item parameter and information estimates. The model fit was acceptable. Results revealed that, as

hypothesized, items in the combined externalizing subscale were characterized by (a) differing item discrimination and difficulty parameter estimates, and (b) disparate levels of information provided along the continuum of externalizing behavior problems. These results are best interpreted via consideration of item- and test-level information, as the amount and location of measurement information offered were directly related to the difficulty and discrimination levels of each item.

Precision of Measurement along the Continuum

As suggested in Chapters III and IV, a screening instrument for externalizing behavior problems intended for use in the pediatric primary care setting would benefit from inclusion of the fewest items possible offering the most information at sub-clinical and clinical levels of externalizing behavior problems. Using the standard normal scale upon which IRT measurement of the latent construct is based, desirable items would be highly informative at levels of externalizing behavior problems 1.5 standard deviations above the mean and higher. The test information curve evaluated in Research Question 1 showed that, as a set, the 18 items in the combined externalizing subscale were most informative between 1.5 and 2 standard deviations above the mean level of externalizing behavior problems (see Figure 9). Test information exceeded the SEE from approximately 1.5 standard deviations below to 3 standard deviations above the mean, suggesting that all 18 items used together precisely measured a wide range of the spectrum of externalizing behavior problems.

For use in screening efforts with the target population in pediatric primary care settings, however, it appeared that several items were superfluous, based on their locations below clinical and sub-clinical levels of externalizing behavior problems. Five

of the 18 items were most informative at levels below the sub-clinical range of externalizing behavior problems, making them undesirable for a brief screening instrument. These included items PSC-17 12 (“Does not listen to rules”), BPI 6 (“Argues too much”), BPI 10 (“Disobedient at home”), BPI 18 (“Stubborn, sullen, or irritable”), and BPI 19 (“Very strong temper”). In essence, these 5 items were revealed to be too easy for the purposes of screening: only low to average levels of externalizing behavior problems among preschool-aged children were necessary for their caregivers to endorse *sometimes* or *often*.

The remaining 13 items exhibited information peaks at sub-clinical to clinical levels of externalizing behavior problems, from 1.5 standard deviations above the mean to over 3 standard deviations above the mean. These items included PSC-17 4 (“Refuses to share”), PSC-17 5 (“Does not understand others’ feelings”), PSC-17 8 (“Fights others”), PSC-17 10 (“Blames others”), PSC-17 14 (“Teases others”), PSC-17 16 (“Takes things”), BPI 3 (“High strung”), BPI 4 (“Cheats/lies”), BPI 9 (“Bullies/cruel or mean”), BPI 11 (“Not sorry after misbehaves”), BPI 12 (“Trouble getting along with others”), BPI 15 (“Not liked by others”), and BPI 22 (“Breaks/destroys things”). In general, these items required sub-clinical to clinical levels of externalizing behavior problems for caregivers to select *often* rather than *sometimes* in describing their child.

Selecting Among Equally Informative Items

One benefit of IRT approaches to scale development is that knowledge of the levels of the latent construct best measured by each item allows the selection of fewer items, as multiple items at a given level are redundant. Thus, given two items located at

the same level of the latent construct, the item providing more information would generally be preferred to the item providing less information.

Among the 13 items in the combined externalizing subscale identified as most informative in the sub-clinical to clinical range of externalizing behavior problems, some duplication was noted in the levels best measured. Item information functions revealed that certain items provided more information than others at the same level of externalizing behavior problems, as depicted in Figure 10. For example, although 4 items exhibited information peaks at 1.6 standard deviations above the mean, item BPI 9 (“Bullies/cruel or mean”) was the most informative at this level of externalizing behavior problems. Similarly, item BPI 11 (“Not sorry after misbehaves”) was the most informative of 3 items which peaked at 1.8 standard deviations above the mean, and item BPI 12 (“Trouble getting along with others”) was more informative than item PSC-17 5 (“Does not understand others’ feelings”) at 2.2 standard deviations above the mean. More informative items, by definition, were those that provided more precision and better discrimination in measurement; thus, they would be preferable to less informative items for inclusion in a brief screening instrument.

Research Question 1 Summary

In summary, 13 items in the combined externalizing subscale were found to be informative in the desired range of the latent construct, with some offering more precision than others at similar levels of externalizing behavior problems. Five items clearly measured lower levels of externalizing behaviors, making them undesirable for inclusion in a brief screening instrument.

Results regarding the precision and utility of measurement of externalizing behavior problems provided by each item were considered in selecting promising items for screening preschool-aged children in pediatric primary care settings. However, additional facets of item performance were salient to the decision process. Specifically, the degree to which an item exhibited measurement bias, or DIF, also influenced its suitability for use in a screening context. Research Question 2 addressed this issue.

Research Question 2: Item-Level Measurement Bias

Item-level bias between groups in measurement is a serious concern in scale development (Camilli & Shepard, 1994; DeVellis, 2003; Osterlind, 1983). From the IRT perspective, when group membership influences item responses while controlling for level of the latent construct of interest, an item exhibits DIF. In the current study, when level of externalizing behaviors was controlled, responses to items exhibiting DIF were influenced by child sex, race, or SES. As discussed in Chapter III, screening instruments for externalizing behavior problems among very young children must be comprised of DIF-free items in order to avoid both over- and under-identification of children of particular group membership (e.g., females, minorities, and those of low SES) in need of further assessment and services (Spencer et al., 2005). Unbiased measurement is crucial to ensure just and equitable screening of children from all sociodemographic backgrounds.

Two approaches to DIF detection were employed: IRT-LR (Thissen, 2001) and OLR (Crane et al., 2004). Analyses compared item responses and parameters between male children and female children; white children and minority children; and low SES children and medium/high SES children. The IRT-LR method utilized likelihood ratio

tests to identify DIF in difficulty and discrimination parameters between groups of interest. In the OLR approach, three nested ordinal logistic regression models were fit for each item, predicting item responses with and without main effects and interaction effects of group membership. Uniform (i.e., statistically significant main effect of group membership, controlling for level of externalizing behavior problems) and non-uniform (i.e., statistically significant interaction effects of group membership and level of externalizing behavior problems) DIF were assessed. Due to a significant bivariate association between child race and child SES, OLR analyses controlled for SES in DIF analyses by child race, and vice versa. Results from each method were compared and combined in order to identify items which exhibited statistically significant DIF.

Detection of Significant DIF

As hypothesized, the 18 items comprising the combined externalizing subscale exhibited varying degrees of DIF by child sex, race, and SES. Only one item was completely free of any indication of DIF in all analyses: item BPI 15 (“Not liked by other children”). Typically, however, in studies of DIF, it is common for different detection methods to yield disparate results (Teresi, 2001). In the current study, DIF was detected in several items with significance levels not meeting the Bonferroni-corrected criterion, either by one or both methods. For example, items PSC-17 8 (“Fights others”), PSC-17 12 (“Does not listen to rules”), BPI 11 (“Not sorry after misbehaves”), BPI 12 (“Trouble getting along with others”), and BPI 19 (“Very strong temper”) were found to exhibit DIF by only one method each, with significance levels of $p < .05$. In these items, due to the possibility of inflated Type I error and the lack of concordance between DIF-detecting methods, it seems likely that the DIF detected may not be valid or meaningful.

Other items—PSC-17 5 (“Does not understand others’ feelings”), PSC-17 16 (“Takes things”), BPI 9 (“Bullies/cruel or mean”), and BPI 10 (“Disobedient at home”)—were each flagged with at least one type of DIF by both methods, but only at the uncorrected level of significance. Concerns regarding false positive findings extend to these items as well, despite the apparent duplication of results from both methods.

In contrast, while most items were unbiased between the groups of interest, eight items were identified with significant DIF by child sex, race, or SES at the stringent level of significance adjusted for multiple comparisons. Five items exhibited significant DIF detected by a single method: items PSC-17 4 (“Refuses to share”) and BPI 22 (“Breaks/destroys things on purpose”) demonstrated DIF by child sex; item PSC-17 10 (“Blames others”) by child race; and items BPI 4 (“Cheats/lies”) and BPI 18 (“Stubborn, sullen, or irritable”) by child SES. The remaining three items were of the greatest concern due to the detection of significant DIF duplicated by both methods: items PSC-17 14 (“Teases others”) and BPI 6 (“Argues too much”) exhibited DIF by child race, and item BPI 3 (“High strung”) by both child race and SES. Assessment of the magnitude and direction of DIF detected in each of these items provided additional information regarding item-level bias and potential effects on the measurement of externalizing behavior problems in the target population.

Magnitude and Direction of DIF Effects

To determine the extent of DIF present in the 8 items of concern, the GRM was refit to all 18 items in the combined externalizing subscale, allowing the parameters of the items identified with DIF to differ between salient groups. The DIF-adjusted parameter estimates for these recalibrated items were visually compared using plots of

item OCCs by group (see Tables 19-21 for re-estimated parameters and Figures 11-13 for OCC plots). In addition, the recalibrated item parameter estimates were applied in IRT scoring, enabling comparisons between unadjusted and DIF-adjusted IRT scores within the total sample as well as within groups split by child sex, race, and SES.

DIF by child sex. As seen in Figure 11, the DIF by child sex observed in items PSC-17 4 (“Refuses to share”) and BPI 22 (“Breaks/destroys things”) was not extensive. Item PSC-17 4, in fact, exhibited DIF primarily in the low to average range of externalizing behavior problems—levels not of great concern in a screening context. Interestingly, the DIF effects by sex for these two items were in opposite directions: item PSC-17 4 (“Refuses to share”) was more difficult for boys, while item BPI 22 (“Breaks/destroys things”) was more difficult for girls. If presented together, the effects of DIF in one item could offset the other.

DIF by child race. Effects of DIF by child race, however, were generally larger than those detected by child sex. In all four items demonstrating DIF by child race, examination of plots of the OCCs by racial group revealed measurement differences within the range of externalizing behavior problems most salient in a screening context (see Figure 12). The largest DIF effect observed overall was for item PSC-17 10 (“Blames others”), in which the upper difficulty threshold for white children was nearly three-quarters of a standard deviation higher than for minority children. This difference represented a substantial divergence in the measurement of externalizing behavior problems provided by this item between white and minority children. Noticeably lower levels of externalizing behavior problems were necessary for caregivers of minority children to report that the child *often* blamed others, as compared to those required for

caregivers of white children. Moreover, the ability of this item to discriminate well among children at similar levels of externalizing behavior problems was better with minority children than with white children. In short, endorsement of each response option for item PSC-17 10 (“Blames others”) by caregivers of white and minority children provided different information about the latent construct being measured.

The three remaining items displaying DIF by race—items PSC-17 14 (“Teases others”), BPI 3 (“High strung”) and BPI 6 (“Argues too much”)—also exhibited meaningful differences in difficulty threshold parameters for white and minority children, ranging from approximately one-half to two-thirds of a standard deviation in magnitude. As observed in the items displaying DIF by child sex, however, the direction of DIF effects was not consistent among all four items, suggesting that at the scale level, presentation of certain item combinations could either mitigate or exacerbate the observed item-level bias.

DIF by child SES. Regarding the three items exhibiting DIF by child SES, the largest effect size observed was for item BPI 4 (“Cheats/lies”), in which the upper difficulty threshold for low SES children was nearly three-quarters of a standard deviation lower than for medium/high SES children. This item’s DIF was primarily problematic in the sub-clinical to clinical range of externalizing behavior problems, raising concerns related to its use in a screening instrument. The remaining DIF effects by SES were of lesser magnitude, though item BPI 3 (“High strung”) also performed differently between SES groups within the range of the latent construct salient to screening. As with the DIF detected by child sex and race, DIF effects by child SES were not consistent in direction.

Unadjusted versus DIF-adjusted IRT scores. The final assessment of DIF effects involved paired *t*-tests comparing theta scores obtained with the original combined externalizing subscale item parameter estimates with those obtained once the 8 items exhibiting DIF were recalibrated for each relevant group. While no significant differences in mean IRT scores were noted for the total sample or within groups split by child sex, small (i.e., ranging from 0.01 to 0.02 standard deviations) but statistically significant ($p < .001$) differences were noted within groups split by child race and SES. These findings suggested that the DIF exhibited by the 8 identified items did have some minor effect on IRT estimates of the levels of externalizing behavior problems within racial and socioeconomic subgroups, as measured by all 18 items in the combined externalizing subscale. The clinical significance of the observed differences in these analyses, however, was negligible.

Notably, as discussed above, many of the items exhibiting DIF by child sex, race, and SES did so in opposing directions. The IRT score estimates obtained before and after adjustments for DIF were generated using all items in the combined externalizing subscale. Thus, it is likely that DIF in opposite directions diminished effects within a given group. For example, while the DIF effects by race were noted to be relatively large for all four identified items, the DIF in items PSC-17 10 (“Blames others”) and PSC-17 14 (“Teases others”)—both of which were easier for caregivers of minority children at above average levels of externalizing behavior problems—may have, in essence, canceled out the DIF in items BPI 3 (“High strung”) and BPI 6 (“Argues too much”)—both of which were easier for caregivers of white children. If the PSC-17 alone were administered to caregivers, scores for minority children could be inflated due to the

tendency of their caregivers to select higher response options than caregivers of white children at the same levels of externalizing behavior problems. The reverse would be true if the BPI alone were administered. Similar concerns exist regarding the items exhibiting DIF by child sex: If only the PSC-17 were administered, girls' scores could be artificially higher than boys' scores, while if only the BPI were administered, boys' scores could be inflated compared to girls' scores. In contrast, all three items exhibiting DIF by child SES were from the BPI, meaning that the direction of DIF in two items (i.e., easier for medium/high SES children) could still be offset by the direction of DIF in the third (i.e., easier for low SES children). This issue highlights the importance of avoiding DIF altogether in the construction of screening instruments, as various combinations of items demonstrating bias may have differing effects on scale-level measurement.

Research Question 2 Summary

In summary, 8 items in the combined externalizing subscale were identified with statistically significant DIF between groups split by child sex, race, or SES. Notably, within each category of DIF—by sex, race, and SES—the direction of DIF among items was not consistent. In addition, effect sizes of DIF ranged from very small to quite large. Therefore, at the scale level, various combinations of items exhibiting DIF could either exacerbate or reduce the effects of item-level bias on overall scores. In the present analyses of all 18 items in the combined externalizing subscale, comparisons of unadjusted and DIF-adjusted IRT score estimates revealed significant but very small differences within groups of white, minority, low SES, and medium/high SES children. These small effects may have been in part due to the presence of items exhibiting DIF in

opposing directions; larger effects could be possible with different combinations of items, especially via inclusion of items demonstrating DIF in the same direction.

Despite the findings of only slight differences between unadjusted and DIF-adjusted IRT scores, item-level measurement precision among these eight items was unsatisfactory. The use of items free of DIF would be preferable in a screening context, increasing confidence in the accuracy of measurement obtained and minimizing false positive or negative findings attributable to sociodemographic characteristics. Ultimately, DIF-free items could contribute toward alleviating disparities in identification of externalizing behavior problems among diverse groups. Thus, the findings obtained via DIF-detection analyses were integrated with the results regarding item information from Research Question 1 to suggest a set of items most promising for screening for externalizing behavior problems among very young children in the pediatric primary care setting. First, however, the relevance of item content was appraised as it related to issues in screening diverse preschool-aged children for externalizing behavior problems. Appreciation of the possible bearings of item content on measurement properties may provide a useful context for deliberations regarding “best” items.

Item Content: Relevance to Screening of Very Young Children

Consideration of the content of items found to be informative within particular ranges along the continuum of externalizing behavior problems may elucidate challenges in the assessment of very young children. Similarly, reflection regarding the content of items which did versus did not exhibit DIF may contribute to understanding of the inherent complexities of screening children differing by sex, race, and SES using

caregiver-report questionnaires. These issues are explored below, providing further background for the integration of the results of Research Questions 1 and 2.

Item Content and Item Information

Review of the content of easy versus difficult items in the combined externalizing subscale revealed several themes relevant to child development and the assessment of very young children. Shared themes are noted below for each group of items, potentially explaining why items measured best at the levels that they did.

Easy items. Of the five items found to be easy (i.e., informative primarily at lower levels of externalizing behavior problems), three appeared to share a theme of noncompliance in their content. According to the difficulty parameter estimates for items PSC-17 12, BPI 6, and BPI 10, not following rules, being argumentative, and being disobedient at home were behaviors which required relatively low levels of externalizing behavior problems in order for caregivers to describe their frequency as *often* rather than *sometimes*. The two remaining easy items also shared thematic content: Items BPI 18 and BPI 19 both referred to issues of temperament or mood, whether stubbornness and irritability or anger and losing one's temper. In fact, with the exception of item BPI 3 ("High strung"), all items in the combined externalizing subscale which referred to either noncompliance with authority or issues of temperament were found to be easy.

The classification of these five items as best measuring non-problematic levels of externalizing behaviors likely reflects the developmental stages of very young children, in whom such behaviors and moods are usually typical and not cause for concern (Merritt et al., 2003; Reijneveld et al., 2004; Task Force on Research Diagnostic Criteria, 2003; Thomasgard & Metz, 2004). While item content indicating noncompliance,

argumentativeness, and irritability may be helpful in screening for externalizing behavior problems in older children, these items do not appear to be informative for the target population. In a developmental context, behaviors eliciting concern at a particular developmental stage may be perfectly acceptable and expected at another. For the purpose of screening very young children, it appears that inclusion of these items would contribute to measurement error in the sub-clinical to clinical range of externalizing behavior problems.

Difficult items. The content of the 13 items identified as most informative at sub-clinical to clinical levels of externalizing behavior problems was also enlightening in the context of developmental stages. Compared to the 5 items which were more informative at lower levels of externalizing behavior problems, the content of several of the difficult items appeared to suggest behaviors exceeding the developmentally typical noncompliance observed in many preschool-aged children. While the easier 5 items tended to reflect issues with arguing and defying authority, many of the 13 more difficult items indicated problems in relationships with peers, including items PSC-17 4 (“Refuses to share”), PSC-17 5 (“Does not understand others’ feelings”), PSC-17 8 (“Fights others”), PSC-17 10 (“Blames others”), PSC-17 14 (“Teases others”), BPI 9 (“Bullies/cruel or mean”), BPI 12 (“Trouble getting along with others”), and BPI 15 (“Not liked by others”). Other difficult items, such as items PSC-17 16, BPI 4, BPI 11, and BPI 22, denoted antisocial behaviors and characteristics such as stealing, cheating or lying, showing lack of remorse, and destroying things on purpose—each of which is suggestive of the diagnostic criteria for behavioral disorders such as Oppositional Defiant Disorder, Conduct Disorder, or Disruptive Behavior Disorder Not Otherwise Specified

(see Appendix A; APA, 2000). Of the items best measuring sub-clinical to clinical levels of externalizing behavior problems, only one contained content not belonging to either of these categorizations: Item BPI 3 (“High strung”) was alone in referring to an issue of temperament. However, highly reactive temperament, as alluded to in this item, is a known correlate of the behavioral disorders (APA, 2000). How the content of this item differs from the content of the other easy items referencing temperament—items BPI 18 (“Stubborn, sullen, or irritable”) and BPI 19 (“Very strong temper”)—is not clear.

Summary: Item content along the continuum. In summary, the content of items best measuring sub-clinical and clinical levels of the latent construct in the target population appeared to tap behaviors and characteristics more severe than the typical noncompliance observed in very young children, including peer relationship problems and antisocial tendencies. Each of these issues has been identified as a risk factor for externalizing behavior problems in previous research (see Hann & Borek, 2001, for a review). Very young children exhibiting these behaviors or attributes frequently may benefit from further assessment to determine whether early intervention may be helpful or necessary; thus, inclusion of such items in a screening instrument intended for use with very young children could be advantageous.

Item Content and Differential Item Functioning

Review of the content of items in the combined externalizing subscale exhibiting DIF may aid in interpretation of item-level measurement bias. Possible explanations for DIF by child sex, race, and SES are considered in this section, leading to further questions regarding etiology of the observed disparities in item performance. In addition, appraisal of the content of several sample items in the combined externalizing subscale

found to exhibit minimal (if any) DIF may further explicate possible causes of the item-level bias observed in other items.

Item content and child sex. The two items exhibiting significant DIF by child sex did so in opposite directions, as discussed previously. Item PSC-17 4 (“Refuses to share”) was easier for girls at low to average levels of externalizing behavior problems, while item BPI 22 (“Breaks/destroys things”) was easier for boys at average to high levels. In essence, among children *not* exhibiting externalizing behavior problems, caregivers of girls were more likely to report problems sharing than were caregivers of boys. Sharing may be a social behavior expected more of girls than of boys (Maccoby, 1988), leading to heightened sensitivity among girls’ caregivers when difficulty sharing is observed in otherwise behaviorally typical children. Alternatively, the frequency and ease of girls’ and boys’ sharing may actually vary and be differentially related to levels of externalizing behavior problems. Other causes are possible as well. Any reason for DIF, however, undermines the utility of this item for general use with preschool-aged children differing by sex.

Similar possible explanations apply to the significant DIF observed regarding reports of purposeful destructive behaviors in response to item BPI 22 (“Breaks/destroys things”). For children displaying average to high levels of externalizing behavior problems, caregivers of boys may simply be more sensitive to destructive behaviors as compared to caregivers of girls, leading to greater likelihood of endorsing higher response options for this item. However, the frequency and meaning of the purposeful destruction of objects may in fact differ between boys and girls. In any case, the true

relationship of this item to the latent construct of externalizing behavior problems is unclear.

Item content and child race. Of the four items exhibiting significant DIF by child race, two were easier for minority children, while two were easier for white children. The content of items PSC-17 10 and PSC-17 14 suggested that blaming others and teasing others were more frequently observed behaviors among minority children, compared to white children at similar levels of sub-clinical to clinical externalizing behavior problems. As seen in the interpretation of items displaying DIF by child sex, the meaning of this finding is unclear. It is possible that caregivers of minority children were more sensitive to these behaviors than caregivers of white children, leading to their increased likelihood of endorsing *often* rather than *sometimes*. Alternatively, true rates of blaming and teasing behaviors may differ between minority and white children. Again, regardless of the reason for the observed DIF, the relationship of these items to the latent construct of externalizing behavior problems is problematic.

A parallel situation is observed regarding items BPI 3 (“High strung”) and BPI 6 (“Argues too much”), both of which suggest similarly reactive temperamental characteristics. In response to these items, caregivers of white children were more likely to endorse higher frequencies than caregivers of minority children, controlling for overall level of externalizing behavior problems. The same questions again arise: Are the observed differences due to (a) varying sensitivities between caregivers of white versus minority children, (b) actual divergences in characteristics of children in each racial group, or (c) other causes? Any explanation provokes concern regarding the use of these

items for measurement of externalizing behavior problems among diverse young children.

Item content and child SES. Three items from the BPI demonstrated significant DIF between low SES and medium/high SES children. Two of these items—BPI 3 and BPI 18—referred to issues of temperament, including being high strung or stubborn, sullen, and irritable. These items were both easier for medium/high SES children than for low SES children. Interpretation of this DIF in the context of item content regarding child temperament suggests that caregivers of medium/high SES children were more sensitive to these issues than caregivers of low SES children, or perhaps that children of medium/high SES exhibit higher rates of these temperamental concerns than do children of low SES, at similar levels of overall externalizing behavior problems. Other explanations may be possible as well.

The third item exhibiting DIF by child SES exposed large differences in the likelihoods of caregivers of low SES children to endorse higher frequencies of cheating or lying as compared to caregivers of medium/high SES children. Item BPI 4 (“Cheats/lies”) was much easier for low SES children compared to medium/high SES children, especially within the range of sub-clinical to clinical externalizing behavior problems. Interpretation of this effect, given the content of the item, is complicated by the same issues described above. Whether cheating and lying behaviors are more frequent among low SES than medium/high SES children, are more likely to be reported by caregivers of low SES children, or are influenced by some other factor is unclear from these results.

Item content and DIF-free items. In total, 10 items in the combined externalizing subscale exhibited no significant DIF. Findings indicated that items PSC-17 12 (“Does not listen to rules”), BPI 11 (“Not sorry after misbehaves”), BPI 12 (“Trouble getting along with others”), BPI 15 (“Not liked by others”), and BPI 19 (“Very strong temper”) were the least biased of the 10 DIF-free items in the combined externalizing subscale. Comparison and contrast of the content of these 5 items with those identified with significant DIF may highlight additional interpretive issues for consideration.

Of all items included in the combined externalizing subscale, only one was completely free of DIF in all analyses and at all levels of significance considered: item BPI 15 (“Not liked by others”) appeared to function consistently across all groups of children split by sex, race, and SES. The content of this item, therefore, is intriguing. In responding to item BPI 15, caregivers were required to assess other children’s feelings about their child. This perspective was in contrast to the task presented by all other items in the combined externalizing subscale, each of which presented a behavior or attribute of the child in question to be rated. The consistency of item performance across groups split by child sex, race, and SES suggests that in responding to this item, caregivers may have been able to maintain some degree of objectivity not always possible with items directly assessing their child’s behavior. An alternative interpretation may be that most caregivers, regardless of the child’s subgroup membership, were unwilling or unable to either discern or report peer rejection of their child. Notably, this item was previously identified as the most difficult item in the combined externalizing subscale, lending possible credence to this explanation.

Four other DIF-free items are particularly notable. Analyses revealed that any DIF exhibited by items PSC-17 12 (“Does not listen to rules”), BPI 11 (“Not sorry after misbehaves”), BPI 12 (“Trouble getting along with others”), and BPI 19 (“Very strong temper”) was (a) detected only by a single method, and (b) non-significant after adjustment for multiple comparisons. Interestingly, item PSC-17 12 (“Does not listen to rules”) was one of two items referencing issues of noncompliance, both of which were found to be DIF-free. No significant disparities in caregiver responses regarding noncompliance or disobedience were noted across sociodemographic groups. In contrast, item BPI 19 (“Very strong temper”) was the only temperament-related item in the combined externalizing subscale *not* found to exhibit significant DIF. The content of items BPI 11 and BPI 12, however, was not suggestive of any pattern in DIF-related themes. These items referenced antisocial and peer relationship issues, respectively, which were alluded to by items both with and without significant DIF.

Though a conclusive explanation of the DIF-free measurement provided by the above items is not possible given the current data, their utility in measuring externalizing behavior problems still surpassed that of any item exhibiting DIF. The consistency of item responses across groups split by child sex, race, and SES ensured that the relationships between the latent construct and the content of each DIF-free item was not unduly influenced by sociodemographic characteristics.

Summary: Item content and DIF. In summary, review of the content of items exhibiting DIF raised several questions regarding interpretation of the item-level bias detected. For each type of DIF observed, a pattern emerged regarding possible explanations for group differences in item responses, controlling for level of externalizing

behavior problems. Caregiver sensitivities to particular child behaviors could be related to sociodemographic characteristics of the child or family; for example, group norms, cultural issues, or societal expectations may influence the perceived acceptability of target behaviors, leading to over- or under-reporting by differing groups (Kagan et al., 2002; Simonian & Tarnowski, 2001). In contrast, actual differences in child behaviors or attributes could exist between certain groups, captured by disparate caregiver responses to items measuring such behaviors. Other contributing factors, such as idiosyncratic item wording, caregiver literacy, or other unmeasured child or caregiver characteristics, could be possible as well (Simonian et al., 1991). A similar lack of conclusiveness also characterized attempts to understand the lack of DIF demonstrated by the least biased items in the combined externalizing subscale.

Despite the unanswered questions generated by consideration of item content in the presence of DIF, biased items would clearly be inappropriate for generic use in screening a diverse population of very young children, as seen in pediatric primary care settings. The relationship between item response and level of externalizing behavior problems is unclear in such items, potentially leading to inequities in assessment efforts. With these considerations in mind, as well as awareness of the item content relevant to sub-clinical and clinical levels of the latent construct, results for each research question were integrated in an effort to identify the most promising items for screening the target population in pediatric primary care settings.

Integration of Results: Identification of “Best” Items

With regard to Research Question 1, assessment of the precision and utility of items in the combined externalizing subscale revealed 13 items with information peaks

within the sub-clinical to clinical range of externalizing behavior problems. Results for Research Question 2 identified 8 items with DIF between groups split by child sex, race, or SES. Substantial overlap was noted in these results: Of the items found to be most informative within the desired range of externalizing behavior problems, 6 also exhibited DIF. Items PSC-17 4 (“Refuses to share”), PSC-17 10 (“Blames others”), PSC-17 14 (“Teases others”), BPI 3 (“High strung”), BPI 4 (“Cheats/lies”), and BPI 22 (“Breaks/destroys things”), though highly informative in the sub-clinical to clinical range, each demonstrated item-level bias by child sex, race, or SES. As previously suggested, the observed DIF negated the value of these items for screening purposes.

Eliminating the six items demonstrating DIF left seven items for consideration, each of which provided DIF-free measurement within the desired range of the latent construct. Several of these seven items, however, demonstrated information peaks at identical levels of externalizing behavior problems. As discussed previously, given multiple items with information peaks at the same level of the latent construct, the item offering the most information is preferable to those offering less, thus eliminating redundancy and unnecessary measurement error. Figure 14 depicts the relative information levels provided by the remaining seven items along the continuum of externalizing behavior problems, replicating the data from Figure 10 but excluding the six items (listed above) identified with DIF. As illustrated in Figure 14, items PSC-17 5 (“Does not understand others’ feelings”), PSC-17 8 (“Fights others”), and PSC-17 16 (“Takes things”), though informative in the sub-clinical to clinical range of externalizing behavior problems, were each surpassed by other items measuring more precisely at the same levels.

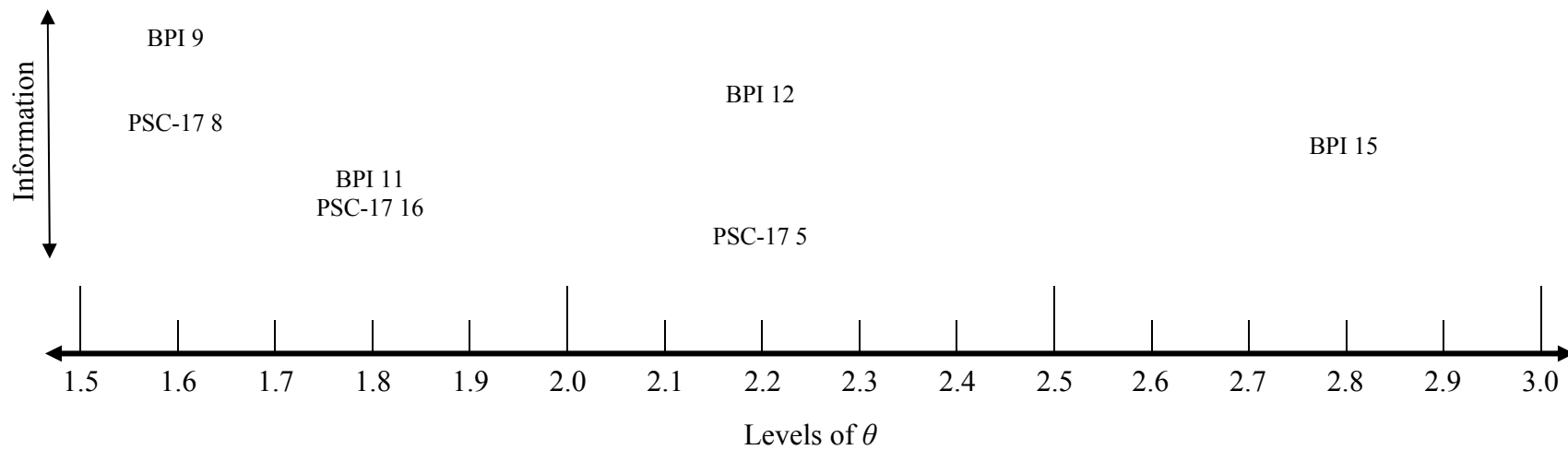


Figure 14. Relative levels of DIF-free item information provided by items in the sub-clinical to clinical range of externalizing behavior problems.

To attain the most efficient, most informative, and least biased measurement of externalizing behavior problems in the target population, four items appeared especially promising for use in screening: items BPI 9 (“Bullies/cruel or mean”), BPI 11 (“Not sorry after misbehaves”), BPI 12 (“Trouble getting along with others”), and BPI 15 (“Not liked by others”). Of all items in the combined externalizing subscale, these four were the most informative along the spectrum of sub-clinical to clinical levels of externalizing behavior problems. Crucially, none of these items demonstrated statistically significant DIF between groups split by child sex, race, or SES. Thus, they appeared to meet the two criteria previously set forth as essential for a brief screening instrument to be used in pediatric primary settings: (a) providing precise measurement of behavior problems at clinical and sub-clinical levels, and (b) demonstrating consistent measurement performance across diverse populations of very young children. It is interesting to note that all four selected items were drawn from the BPI; the three remaining PSC-17 items were possible, albeit slightly less informative, alternatives.

The content of these four items further supports their utility as elements of a brief screening instrument for externalizing behavior problems in very young children. Cruelty to others (BPI 9), lack of remorse (BPI 11), and conflict with peers (BPI 12) are each key diagnostic criteria for behavioral disorders including Oppositional Defiant Disorder, Conduct Disorder, and Disruptive Behavior Disorder NOS (see Appendix A; APA, 2000). Rejection by peers (BPI 15), while not specifically identified as diagnostic of behavioral disorders, has been associated with externalizing behavior problems even among preschool-aged children (Lochman et al., 1993; Lochman & Wayland, 1994). Together, these items allude to issues more severe than the developmentally appropriate

noncompliance referenced by several other items in the combined externalizing subscale, requiring at least sub-clinical levels of externalizing behavior problems for caregivers to endorse *often true* rather than *sometimes true* in describing their frequencies. Item BPI 15 (“Not liked by others”), in fact, required externalizing behavior problems nearly 2 standard deviations above the mean level in order for caregivers to endorse *sometimes true* rather than *not true*, illustrating the relative rarity of this circumstance being reported. The content of items BPI 9, BPI 11, BPI 12, and BPI 15 appears to elicit concerns regarding developmentally inappropriate externalizing behaviors—a challenging task in the assessment of very young children, who typically exhibit behaviors which would be troubling if observed frequently in older children (Keenan & Wakschlag, 2000). Further, the content of each of these items is consistent with recommendations by the Task Force on Research Diagnostic Criteria – Preschool Age (2003) regarding developmentally appropriate assessment of disruptive behavior disorders in very young children.

Implications

Results of this study suggest several important implications for behavioral screening in primary care of very young children from diverse backgrounds. One purpose of the study was to identify the “best” PSC-17 and BPI items for screening externalizing behavior problems in very young children; it follows that the identified set of four most informative and least biased items could be further investigated as a measure appropriate for screening the target population in pediatric primary care settings. Improvements in precision of measurement, accuracy of identification, response burden, and time required for scoring and interpretation could ensue if these items performed well as a stand-alone

screening instrument. While the current study focused on pediatric primary care, the identified items could serve equally well as a screening tool for use in other venues, such as preschools, early child care settings, mental health agencies, and the child welfare system. Ultimately, primary and secondary prevention of the social and public health problem of externalizing behavior problems could be enhanced with such improved screening tools.

In addition, formal qualitative assessment of the content of these 4 items may offer valuable insights regarding the nature of externalizing behavior problems among diverse children in the target age range, as well as of caregivers' perceptions. Similarly, qualitative review of the remaining 14 items could also inform understanding of externalizing behaviors in very young children from varied backgrounds, along the continuum from typical to atypical levels. In particular, examination of the content of items exhibiting DIF could augment the knowledge base regarding the meaning and experience of externalizing behavior problems in preschool-aged children differing by sex, race, or SES. Further, item content found to be informative primarily at low to average levels of the latent construct could be studied to gain insights regarding assessment of typical behavioral development in very young children. Social work researchers with expertise in qualitative methods would be well-positioned to conduct such investigations.

Practically, results of this study could inform the use of the PSC-17 and BPI in practice and research with preschool-aged children. Researchers and clinicians—from social work as well as other fields—should be aware that with the target population of preschool-aged children, certain items in the PSC-17 and BPI exhibited significant

measurement bias between groups split by child sex, race, and SES. Though item-level DIF effects on IRT scoring based upon all 18 items in the combined externalizing subscale were found to be relatively small, particular items from these scales used in combination may exacerbate overall bias, especially with a traditional summed scoring approach. For example, 2 PSC-17 items displayed significant DIF characterized by selection of higher response options for minority children than for white children at the same levels of externalizing behavior problems. Thus, use of both items—as included in the PSC-17 externalizing subscale—may lead to inflated externalizing subscale scores among minority children. Similarly, 2 items in the BPI exhibited DIF by child SES in the same direction, potentially raising scores of medium/high SES children as compared to low SES children; use of a third BPI item, however, could influence scores in the other direction. Awareness of item-level bias is crucial in the interpretation of total and subscale scores for these instruments, as well as in conclusions regarding individual- and group-level measurement for research or clinical purposes. In particular, great caution should be exercised in interpreting discrete responses to any of the 8 individual items found to exhibit significant DIF.

Additionally, researchers and practitioners should be aware that the level of measurement error present in the PSC-17 and BPI may not be constant along the entire continuum of externalizing behavior problems, when used with preschool-aged children. In particular, when used primarily for screening or diagnostic purposes, the presence of easy items in each of these scales may contribute to measurement error within the sub-clinical to clinical range of the latent construct.

These implications lead directly to the issue of scale development relevant to externalizing behavior problems in very young children. Results of this study could be used in continued efforts to improve measurement of this latent construct in the target population for a variety of purposes. If measurement of a broad range of externalizing behaviors (i.e., from below to above average) were desired, the current IRT results have delineated which items are most informative along the entire continuum of the latent construct. If precise measurement of a more restricted range of externalizing behaviors were desired, reduction of measurement error could be achieved via selection of items most salient to the preferred levels. The results of this study could facilitate the development of brief, informative measures for any range of externalizing behavior problems in very young children, using items from the PSC-17 and BPI.

Concerns regarding less informative or biased items extend to instruments beyond the PSC-17 and BPI, when used with diverse populations of very young children. Generic use of items demonstrating DIF (i.e., with a single form instrument intended for use with both boys and girls of all racial and SES groups) may be inappropriate. The item content observed in items flagged as problematic in the current study is also seen in other frequently used scales measuring child behaviors, including the PBQ (Behar & Stringfield, 1974); the PKBS-2 (Merrell, 2003); the BBRB (Burks, 1996); the BASC-2 (Reynolds & Kamphaus, 1992); the CBCL/1.5-5 (Achenbach & Rescorla, 2000); and others. As discussed above, qualitative assessment of the content of these items may inform understanding of the presentation of externalizing behavior problems in preschool-aged children, leading to improved assessment approaches and possible tailoring of instruments to particular groups of children. Related to this possibility is the

promise of other IRT applications such as computerized adaptive testing, employing the item-level psychometric information obtained from the current analyses in the development of individualized, adaptive measurement approaches appropriate for both research and practice settings.

Interestingly, each of the PSC-17 and BPI items identified either as being too easy or as biased for use in screening efforts had relatively high factor loadings and corrected item-total correlations in CTT analyses. Thus, the current study also illustrated the ability of IRT analyses to assess the measurement performance of individual items in ways beyond those offered by traditional CTT psychometric studies. This translational research harnessed the advantages of advanced measurement theory and methods to improve clinical practice and could be replicated and extended in the future. Application of IRT analyses to other areas of assessment would likely be equally informative, potentially improving measurement for a wide array of issues.

Several of the above study implications are particularly relevant to social work education, practice, and research. Heightened attention to both classical and modern measurement theory in social work education could prepare social work practitioners to be cognizant of potential limitations of CTT-developed instruments. This educational focus would also enable social work researchers to increase their participation in the evaluation and development of measurement tools crucial to social work practice and research, especially via advanced measurement theory and applications such as IRT. Investigations of DIF are particularly relevant to efforts to reduce health disparities and promote social justice, activities mandated of all social work professionals by the profession's Code of Ethics (NASW, 2000). As a profession, social work calls for

cultural sensitivity and competence; thus, social work researchers, practitioners, and educators should ensure that the measurement instruments they use meet those standards.

Limitations

Several methodological limitations of this study are important to recognize. First, given the convenience sample necessitated by the study design and resources, there may be some concern regarding generalizability of findings. This concern, however, is mitigated by the sample descriptive statistics and CTT analyses, which suggest similarities between the current sample and instrument performance and the nationally representative samples reported in previous, larger studies (e.g., Gardner et al., 1999; Gortmaker et al., 1990). More importantly, as discussed in Chapter III, IRT methods theoretically yield “sample-free” stable parameter estimates (Hambleton & Swaminathan, 1985), meaning that as long as a broad distribution of externalizing behavior problems was represented in the sample, external validity concerns are unwarranted.

Another limitation of the current study relates to the final set of “best” items identified by integrating results regarding precise and DIF-free items. While IRT methods can identify informative and unbiased items for measurement of a given latent construct, further investigation is needed to assess various types of validity of the final set of items recommended. This limitation of the current study provides direction for future research on the measurement performance of the set of four recommended items in screening efforts.

Regarding limitations of specific study analyses, two variables in particular were coarsely categorized: child race and child SES. The original child race data were dichotomized into categories of white and minority children to achieve the largest

frequencies possible in each group, due to IRT requirements for DIF analyses. Children of minority races other than African-American were not adequately represented at the clinic sites, preventing analyses focused specifically on children of Hispanic ethnicity, Asian race, or other racial or ethnic backgrounds. As a result, however, the DIF results regarding comparisons between white and minority children were even more concerning, given that some effects may have been diluted as a result of the coarse categorization used. Similarly, the operationalization of SES was adequate but not ideal. While the use of three SES indicators (i.e., child's type of health insurance, caregiver's education, and household income) was reasonable, the dichotomization into low and medium/high SES groups was influenced by the distribution of income in the sample and the sociodemographic characteristics of the region. Though minority children of medium SES were adequately represented in the sample, minority children of high SES were not; better representation of this group would have allowed DIF comparisons among low, medium, and high SES children, controlling for child race.

Another limitation related to analyses involving child race was an inability to control for caregiver race. Of the 900 caregiver-child dyads represented by the data, only 47 cases were identified in which child and caregiver race differed. Of these, only 3 involved minority caregivers of white children. Thus, due to small cell sizes, OLR DIF-detection analyses controlling for caregiver race were not possible by stratification nor by inclusion as a covariate. To address this limitation, the 47 cases with unmatched caregiver-child race were excluded from DIF analyses by child race and SES, and odds ratios from these results were compared with those obtained from the total sample. Because only trivial differences in odds ratios were observed, and no differences in items

identified with DIF were noted, results from the total sample were ultimately reported. However, in order to address this issue more precisely, increased frequencies of non-matching caregiver-child dyads would be necessary; in particular, more minority caregivers of white children would be required.

A frequent criticism of many child assessment instruments is the sole reliance on caregiver-reported data (Kagan et al., 2002). Indeed, the inclusion of an objective, standardized measure of child externalizing problems would have facilitated additional analyses and comparisons in the current study. The difficulty interpreting the etiology of DIF findings could be alleviated if such additional data were available. However, caregiver-report questionnaires are ubiquitous in clinical and research settings. Despite the absence in this study of an external objective measure of the latent construct, the current results would suggest that items exhibiting DIF be excluded from screening instruments regardless of the reason for the observed bias. In other words, whether item-level bias by child sex, race, or SES is attributable to child behaviors, caregiver perceptions, or any of a vast array of other possible reasons, the recommendation to exclude biased items remains the same, given the widespread use of caregiver-report instruments.

Related to concerns regarding sole reliance on caregiver-report data is a limitation regarding the concept of a latent construct in IRT. The current study's focus is on the latent construct of externalizing behavior problems. In classical psychometric theory, a latent construct is not directly measurable and requires at least one observable proxy for its assessment (Nunnally & Bernstein, 1994): in this case, caregiver responses to the items under investigation. In IRT, however, the scope of the latent construct may be more

specific, as its scale or continuum is determined by the particular items used for measurement. Thus, in this case, the standard normal scale of the latent construct was determined solely by the patterns of caregiver responses to the items in the combined externalizing subscale. Strictly speaking, then, the latent construct measured was *caregiver-observed* externalizing behavior problems, as indicated by caregiver report. This limitation of the present investigation is especially salient in the context of interpretation of results, in which acknowledgment is necessary that the latent construct of interest is caregiver-observed externalizing problems. Similar limitations are inherent in all psychometric studies utilizing IRT, requiring clear understanding of the latent construct as a prerequisite for interpretation of results.

Other limitations were associated with the reliance on a cross-sectional design with data collection via questionnaires. Follow-up assessments and cognitive interviewing regarding item content were not possible due to the study design. Thus, the stability and predictive value of particular items or combinations of items were not assessed, nor were caregiver perceptions of item content and responses explored. Longitudinal data would be needed in order to assess CTT test-retest reliability and predictive validity of item responses, and formal qualitative evaluation would be necessary to investigate caregivers' rationales for selecting particular response options. However, the current study design and data type were appropriate for the posed research questions and associated analyses.

Finally, in the absence of protocols for power analysis in fitting the GRM to sets of polytomous items, some question may remain regarding sample size. Given previous research and simulation studies regarding item parameter estimation and DIF detection,

however, the sample size was likely sufficient for the analyses conducted (Bolt, 2002; Reise & Yu, 1990). A larger sample may have yielded more precise estimates of item parameters and DIF effects, but it is unclear whether increased precision would be clinically relevant or useful in answering the research questions directing the study.

Directions for Future Research

Opportunities for future translational research building on the results of this study are plentiful. For example, further investigations are recommended to explore the utility of the four items proposed for screening very young children for externalizing behavior problems. Psychometric examination of the potential screening instrument comprised of items BPI 9 (“Bullies/cruel or mean”), BPI 11 (“Not sorry after misbehaves”), BPI 12 (“Trouble getting along with others”), and BPI 15 (“Not liked by others”), especially validity studies, would provide further information regarding its possible use as a screening tool in pediatric primary care settings. Related topics of interest for future research include assessments of outcomes of actual use of the screening instrument in primary care (and possibly other) settings, as well as development of clinical cut-points and evaluation of scoring options. Data regarding the efficacy and effectiveness of the proposed screening instrument in efforts to improve early identification of children with externalizing behavior problems would be invaluable in assessing its performance.

As addressed previously in the discussion of study implications, formal qualitative analysis of the content of items in the combined externalizing subscale could yield important information regarding (a) measurement of externalizing behavior problems in the context of early childhood development, and (b) DIF exhibited between groups split by child sex, race, and SES. Social work researchers with qualitative research

skills would be uniquely suited to pursue such investigations. Focus groups, cognitive interviewing, and other qualitative research methods could be used to evaluate the meaning and perceptions of item content within salient groups. Improved understanding of caregiver perceptions of item content, as well as consideration of the relevance of item content to typical and atypical behavioral development of very young children, could enable further practical and theoretical advances in assessment of this latent construct. Well-conducted qualitative research on this topic would facilitate exploration of the broadest possible range of explanatory factors related to the item performance observed in the current study. These could include issues ranging from cultural and ethnic variations in child behavior and caregiver perceptions to the effects of literacy on item comprehension.

In a similar vein, follow-up studies investigating the significant DIF detected in eight items from the PSC-17 and BPI could address the observed group differences in item responses. In particular, studies designed to identify the sources of DIF—such as differences in caregiver perceptions versus actual disparities in child behavior between groups—would be beneficial. Structured comparisons of caregiver ratings within groups split by child sex, race, and SES to more objective measures of child behavior could explicate issues in this area. Such investigations would also contribute to improved measurement of externalizing behavior problems in the target population.

While the present study focused solely on preschool-aged children, comparisons of item functioning between younger and older children would be highly informative as well. By including children of varying age groups in the sample, analyses of responses to the items in the combined externalizing subscale could assess (a) DIF by child age, as

well as (b) differential DIF effects by child sex, race, and SES among younger versus older children. The same methods employed in the current study could be utilized with a sample comprised of children of a broader age range to address these issues. Results could further elucidate issues of item bias and behavioral assessment within the context of child development. Other changes in sample composition, such as inclusion of more children of minority races other than African-American as well as more cases in which caregiver and child race do not match, would allow even more specific investigations of DIF effects.

Inclusion of caregiver characteristics beyond basic sociodemographic information would be another potentially fruitful direction for future research, addressing a broader scope of known contributing factors to child externalizing behavior problems. Consideration of parent mental health, family functioning, and other relevant caregiver- and family-level attributes could connect the current results regarding measurement of child externalizing behavior problems to related findings in the literature.

Finally, the application of IRT analyses to investigate the measurement performance of items in the PSC-17 and BPI could be extended to the other latent constructs assessed by these instruments: specifically, internalizing behavior problems and Attention Deficit Hyperactivity Disorder. The quality and utility of measurement offered by relevant items in the assessment of preschool-aged children could be evaluated, providing direction to efforts toward improving screening of the target population for these issues as well.

Summary and Conclusions

Screening for externalizing behavior problems in very young children followed in pediatric primary care requires a brief, easily scored instrument which can detect sub-clinical to clinical levels of the latent construct within the context of early childhood development. Equally importantly, to ensure equitable efforts in primary and secondary prevention with the diverse populations of young children seen in primary care, each item utilized should be free of bias related to sociodemographic characteristics. Most measures currently in use suffer from a variety of drawbacks limiting their appropriateness for the primary care setting, including excessive length and norms developed with unrepresentative samples. Of particular concern, several studies have suggested that female, minority, and low SES children are identified with externalizing behavior problems at both higher and lower rates than expected by many screening instruments. Traditional CTT-based psychometric methods of evaluating measurement performance are insufficient to address these concerns. Analyses developed from the modern measurement theory of IRT, however, offer novel information regarding the psychometric performance of items used to measure a given latent construct.

This study investigated the precision, utility, and measurement bias of items measuring externalizing behavior problems in two commonly used caregiver-report questionnaires: the PSC-17 (Gardner et al., 1999) and the BPI (Peterson & Zill, 1986; Zill, 1990). A large, diverse sample of caregivers of preschool-aged children seen in pediatric primary care provided data which were analyzed using Samejima's (1969) GRM. All items comprising the instruments' combined externalizing subscales were evaluated for (a) levels of externalizing behavior problems best measured, and (b) item-

level measurement bias exhibited by child sex, race, and SES. Five items were found to measure only low to average levels of externalizing problems in the target population, while the remaining 13 were informative at sub-clinical to clinical levels. Significant DIF was detected in 8 of 18 items by child sex, race, or SES. These findings call into question the use of the respective externalizing subscales of the PSC-17 and BPI with diverse populations of very young children, as measurement error and disparities in item performance may affect both item- and scale-level performance. However, a set of 4 items found to be the most informative within sub-clinical to clinical levels of the latent construct, as well as the least biased between groups differing by sociodemographic characteristics, appears to be a promising tool for screening purposes with preschool-aged children in the primary care setting. Additional investigations of the measurement properties of this set of items are needed to assess its potential value in improving early identification of very young children with externalizing behavior problems. Moreover, formal evaluation of the content of these items—as well as of the items not selected for screening purposes—may provide crucial insights for theoretical and practical developments regarding assessment of externalizing behavior problems within the context of early childhood development.

In conclusion, primary and secondary prevention efforts are very promising approaches for reducing the detrimental effects of the social and public health problem of externalizing behavior problems in very young children. Improving early identification in the pediatric primary care setting is an important step in such efforts. Results of the present study may contribute to advances in screening technologies, ultimately enriching

endeavors to alleviate the distress experienced by children, families, communities, and society in response to early onset of externalizing behavior problems in children.

REFERENCES

- Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and the 1991 profile*. Burlington, VT: University of Vermont Department of Psychiatry.
- Achenbach, T. M., & Edelbrock, C. (1981). Behavioral problems and competencies reported by parents of normal and disturbed children aged four through sixteen. *Monographs of the Society for Research in Child Development, 46*, 1-82.
- Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for ASEBA preschool forms and profiles*. Burlington, VT: University of Vermont Department of Psychiatry.
- Agency for Healthcare Research and Quality. (2002). *Guide to clinical preventive service, 3rd edition: Systematic evidence reviews*. Rockville, MD: Agency for Healthcare Research and Quality.
- Agresti, A. (2002). *Categorical data analysis (2nd ed.)*. Hoboken, NJ: John Wiley & Sons.
- American Academy of Pediatrics. (2000). *Fellows Survey*. Elk Grove Village, IL: American Academy of Pediatrics.
- American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders (DSM-III)*. Washington, DC: American Psychiatric Association.
- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (DSM-IV-TR)*. Washington, DC: American Psychiatric Association.
- Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heineman.
- Baker, P. C., Keck, C. K., Mott, F. L., & Quinlan, S. V. (1993). *National longitudinal survey of youth handbook (revised edition)*. Columbus, OH: The Ohio State University, Center for Human Resource Research.
- Bates, J. E., Pettit, G. S., Dodge, K. A., & Ridge, B. (1998). Interaction of temperamental resistance to control and restrictive parenting in the development of externalizing behavior. *Developmental Psychology, 34*, 982-995.
- Behar, L. B., & Stringfield, S. (1974). A behavior rating scale for the preschool child. *Developmental Psychology, 10*, 601-610.

- Belsky, J., Hsieh, K. H., & Crnic, K. (1998). Mothering, fathering, and infant negativity as antecedents of boys' externalizing problems and inhibition at 3 years: Differential susceptibility to rearing experience? *Development and Psychopathology, 10*, 301-319.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Bishop, S. J., Murphy, J. M., Jellinek, M. S., & Dusseault, K. (1991). Psychosocial screening in pediatric practice: A survey of interested physicians. *Clinical Pediatrics, 30*, 142-147.
- Bjorner, J. B., Kosinski, M., & Ware, J. E. (2003a). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT). *Quality of Life Research, 12*, 913-933.
- Bjorner, J. B., Kosinski, M., & Ware, J. E. (2003b). The feasibility of applying item response theory to measures of migraine impact: A re-analysis of three clinical studies. *Quality of Life Research, 12*, 887-902.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.
- Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*, 21-33.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*.
- Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113-141.
- Borowsky, I. W., Mozayeny, S., & Ireland, M. (2003). Brief psychosocial screening at health supervision and acute visits. *Pediatrics, 112*, 129-133.
- Boyce, C. A., Hoagwood, K., Lopez, M. L., & Tarullo, L. B. (2000). The Head Start Mental Health Research Consortium: New directions for research partnerships. *Behavioral Disorders, 26*, 7-12.
- Brennan, P. A., Grekin, E. R., & Mednick, S. A. (1999). Maternal smoking during pregnancy and adult male criminal outcomes. *Archives of General Psychiatry, 56*, 215-219.

- Brody, G. H., Stoneman, Z., & Flor, D. (1996). Parental religiosity, family processes, and youth competence in rural, two-parent African American families. *Developmental Psychology, 32*, 696-706.
- Brown, R. T., Coles, C. D., Smith, I. E., Platzman, K. A., Silverstein, J., Erickson, S., et al. (1991). Effects of prenatal alcohol exposure at school age II: Attention and behavior. *Neurotoxicology and Teratology, 13*, 369-376.
- Burks, H. F. (1996). *Burks' Behavior Rating Scales: Manual*. Los Angeles: Western Psychological Services.
- Camilli, G., & Shepard, L. N. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.
- Campbell, S. B. (1994). Hard-to-manage preschool boys: Externalizing behavior, social competence, and family context at two-year followup. *Journal of Abnormal Child Psychology, 22*, 147-166.
- Campbell, S. B., Pierce, E. W., & Moore, G. (1996). Boys' externalizing problems at elementary school age: Pathways from early behavior problems, maternal control, and family stress. *Development and Psychopathology, 8*, 701-719.
- Center for Human Resource Research. (2000). *NLSY79: 1998 child and young adult data users guide*. Columbus, OH: Center for Human Resource Research, The Ohio State University.
- Cheney, C. O., & Sampson, K. (1990). Issues in identification and service delivery for students with conduct disorders: The "Nevada solution". *Behavioral Disorders, 15*, 174-179.
- Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523-562.
- Cicchetti, D., & Cohen, D. J. (Eds.). (1995). *Developmental psychopathology, Vol. 1: Theory and methods*. Oxford: John Wiley & Sons.
- Clarizio, H. F. (1992). Differentiating emotionally impaired from socially maladjusted students. *Psychology in the Schools, 24*, 237-243.
- Cohen, D., & Strayer, J. (1996). Empathy in conduct-disordered and comparison youth. *Developmental Psychology, 32*, 988-998.
- Coie, J. D., Watt, N. F., West, S. G., Hawkins, J. D., Asarnow, J. R., Markman, H. J., et al. (1993). The science of prevention: A conceptual framework and some

- directions for a national research program. *American Psychologist*, 48, 1013-1022.
- Coles, C. D., Brown, R. T., Smith, I. E., Platzman, K. A., Erickson, S., & Falek, A. (1991). Effects of prenatal alcohol exposure at school age I: Physical and cognitive development. *Neurotoxicology and Teratology*, 13, 357-367.
- Conger, R. D., Conger, K. J., Elder, G. H., Lorenz, F. O., Simons, R. L., & Whitbeck, L. B. (1992). A family process model of economic hardship and adjustment of early adolescent boys. *Child Development*, 63, 526-541.
- Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry*, 126, 884-888.
- Conrad, P., & Schneider, J. (1980). *Deviance and medicalization: From badness to sickness*. St. Louis, MO: C. V. Mosby Co.
- Cooksey, E. C., Menaghan, E. G., & Jekielek, S. M. (1997). Life course effects of work and family circumstances on children. *Social Forces*, 76, 637-667.
- Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10, <http://pareonline.net/getvn.asp?v=10&n=17>.
- Costello, E. J. (1986). Primary care pediatrics and child psychopathology: A review of diagnostic, treatment, and referral practices. *Pediatrics*, 78, 1044-1051.
- Costello, E. J., & Edelbrock, C. (1985). Detection of psychiatric disorders in pediatric primary care: A preliminary report. *Journal of the American Academy of Child Psychiatry*, 24, 771-774.
- Costello, E. J., Edelbrock, C., Costello, A. J., Dulcan, M. K., Burns, B. J., & Brent, D. (1988). Psychopathology in pediatric primary care: The new hidden morbidity. *Pediatrics*, 82, 415-424.
- Costello, E. J., & Shugart, M. A. (1992). Above and below the threshold: Severity of psychiatric symptoms and functional impairment in a pediatric sample. *Pediatrics*, 90(3), 359-368.
- Crane, P. K., Jolley, L., & van Belle, G. (2003). *DIFdetect (STATA software)*. Seattle: University of Washington.
- Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine*, 23, 241-256.

- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group/Thomson Learning.
- Cyphers, L. H., Phillips, K., Fulker, D. W., & Mrazek, D. A. (1990). Twin temperament during the transition from infancy to early childhood. *Journal of the American Academy of Child and Adolescent Psychiatry, 29*, 392-397.
- DeRoos, Y. S., & Allen-Meares, P. (1992). Rasch analysis: Its description and use analyzing the Children's Depression Inventory. *Journal of Social Service Research, 16*, 1-17.
- DeVellis, R. F. (2003). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage Publications.
- Dodge, K. A., Pettit, G. S., Bates, J. E., & Valente, E. (1995). Social information-processing patterns partially mediate the effect of early physical abuse on later conduct problems. *Journal of Abnormal Psychology, 104*, 632-643.
- Dragow, F., Levine, M. V., Tsein, S., Williams, B. A., & Mead, A. D. (1995). Fitting polytomous IRT models to multiple choice tests. *Applied Psychological Measurement, 19*, 143-165.
- Drotar, D. (1999). The Diagnostic and Statistical Manual for Primary Care (DSM-PC), Child and Adolescent Version: What pediatric psychologists need to know. *Journal of Pediatric Psychology, 24*(5), 369-380.
- Drotar, D. (2004). Detecting and managing developmental and behavioral problems in infants and young children. *Infants & Young Children, 17*, 114-124.
- Duncan, B. B., Forness, S. R., & Hartsough, C. (1995). Students identified as seriously emotionally disturbed in school-based day treatment: Cognitive, psychiatric, and special educational characteristics. *Behavioral Disorders, 20*, 238-252.
- Edelbrock, C., Rende, R., Plomin, R., & Thompson, L. A. (1995). A twin study of competence and problem behavior in childhood and early adolescence. *Journal of Child Psychology and Psychiatry, 36*, 775-785.
- Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the Mini-Mental State Examination. *Medical Care, 44*(11 Suppl 3), S134-S142.
- Eisenberg, N., Fabes, R. A., Guthrie, I. K., Murphy, B. C., Maszk, P., Holmgren, R., et al. (1996). The relations of regulation and emotionality to problem behavior in elementary school children. *Development and Psychopathology, 8*, 141-162.

- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Erickson, M. F., Sroufe, L. A., & Egeland, B. (1985). The relationship between quality of attachment and behavior problems in preschool in a high-risk sample. *Monographs of the Society for Research in Child Development, 50*, 147-166.
- Farmer, E. (1993). Externalizing behavior in the life course: The transition from school to work. *Journal of Emotional and Behavioral Disorders, 1*, 179-188.
- Farrington, D. P., & Hawkins, J. D. (1991). Predicting participation, early onset and later persistence in officially recorded offending. *Criminal Behavior and Mental Health, 1*, 1-33.
- Farver, J. M. (1996). Aggressive behavior in preschoolers' social networks: Do birds of a feather flock together? *Early Childhood Research Quarterly, 11*, 351-376.
- Fayers, P. (2004). *IRT: The way forward*. Paper presented at the NCI and DIA Advances in Health Outcomes Measurement Meeting: Exploring the Current State and the Future of Item Response Theory, Item Banks, and Computer-Adaptive Testing, Bethesda, MD.
- Feil, E. G., Severson, H. H., & Walker, H. M. (1998). Screening for emotional and behavioral delays: The Early Screening Project. *Journal of Early Intervention, 21*(3), 252-266.
- Feldman, S. S., & Weinberger, D. A. (1994). Self-restraint as a mediator of family influences on boys' delinquent behavior: A longitudinal study. *Child Development, 65*, 195-211.
- Fergusson, D. M., & Lynskey, M. T. (1993). Maternal age and cognitive and behavioural outcomes in middle childhood. *Paediatric and Perinatal Epidemiology, 7*, 77-91.
- Forness, S. R., Cluett, S. E., Ramey, C. T., Ramey, S. L., Zima, B. T., Hsu, C., et al. (1998). Special education identification of Head Start children with emotional and behavioral disorders in second grade. *Journal of Emotional and Behavioral Disorders, 6*, 194-204.
- Forness, S. R., & Kavale, K. A. (2001). Ignoring the odds: Hazards of not adding the new medical model to special education decisions. *Behavioral Disorders, 26*, 269-281.
- Forness, S. R., Kavale, K. A., MacMillan, D. L., Asarnow, J. R., & Duncan, B. B. (1996). Early detection and prevention of emotional or behavioral disorders: Developmental aspects of systems of care. *Behavioral Disorders, 21*, 226-240.

- Forness, S. R., & Knitzer, J. (1992). *A new proposed definition and terminology to replace "serious emotional disturbance" in individuals with disabilities education act*. Alexandria, VA: The National Mental Health and Special Education Coalition.
- Gamoran, A., Nystrand, M., Berends, M., & LePore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal*, 32, 687-715.
- Gardner, W., Kelleher, K. J., Pajer, K. A., & Campo, J. V. (2004). Primary care clinicians' use of standardized psychiatric diagnoses. *Child: Care, Health & Development*, 30, 401-412.
- Gardner, W., Lucas, A., Kolko, D. J., & Campo, J. V. (2007). Comparison of the PSC-17 and alternative mental health screens in an at-risk primary care sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, 46(5), 611-618.
- Gardner, W., Murphy, M., Childs, G., Kelleher, K., Pagano, M., Jellinek, M., et al. (1999). The PSC-17: A brief pediatric symptom checklist with psychosocial problem subscales. A report from PROS and ASPN. *Ambulatory Child Health*, 5, 225-236.
- Ge, X., Donnellan, M. B., & Wenk, E. (2003). Differences in personality and patterns of recidivism between early starters and other serious male offenders. *The Journal of the American Academy of Psychiatry and the Law*, 31, 68-77.
- Goodwin, R., Gould, M. S., Blanco, C., & Olfson, M. (2001). Prescription of psychotropic medications to youths in office-based practice. *Psychiatric Services*, 52, 1081-1087.
- Gorman-Smith, D., & Tolan, P. (1998). The role of exposure to community violence and developmental problems among inner-city youth. *Development and Psychopathology*, 10, 101-116.
- Gortmaker, S., Walker, D., Weitzman, M., & Sobol, A. (1990). Chronic conditions, socioeconomic risks, and behavioral problems in children and adolescents. *Pediatrics*, 85, 267-276.
- Greenspan, S. I. (1992). *Infancy and early childhood: The practice of clinical assessment and intervention with emotional and developmental challenges*. Madison, CT: International Universities Press.
- Griffin, K. W., Scheier, L. M., Botvin, G. J., Diaz, T., & Miller, N. (1999). Interpersonal aggression in urban minority youth: Mediators of perceived neighborhood, peer, and parental influences. *Journal of Community Psychology*, 27, 281-298.

- Guerin, D. W., Gottfried, A. W., & Thomas, C. W. (1997). Difficult temperament and behaviour problems: A longitudinal study from 1.5 to 12 years. *International Journal of Behavioral Development, 21*, 71-90.
- Gumpel, T. (1998). An item response theory analysis of the Conners Teacher's Rating Scale. *Journal of Learning Disabilities, 31*, 525-532.
- Haggerty, R. J. (1974). The changing role of the pediatrician in child health care. *American Journal of Diseases of Children, 127*, 545–549.
- Halfon, N., Regalado, M., McLearn, K. T., Kuo, A. A., & Wright, K. (2003). *Building a bridge from birth to school: Improving developmental and behavioral health services for young children*. New York: The Commonwealth Fund, publication no. 564.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38-47.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.
- Hann, D., & Borek, N. (2001). *Taking stock of risk factors for child/youth externalizing behavior problems*. NIH publication no. 02-4938. Washington, DC: National Institute of Mental Health.
- Hawkins-Walsh, E., & Stone, C. (2004). A national survey of PNP curricula: Preparing pediatric nurse practitioners to meet the challenge in behavioral mental health. *Pediatric Nursing, 30*, 72-78.
- Hawkins, J. D., Catalano, R. F., Kosterman, R., Abbott, R. D., & Hill, K. G. (1999). Preventing adolescent health-risk behaviors by strengthening protection during childhood. *Archives of Pediatrics & Adolescent Medicine, 153*, 226-234.
- Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9 Supp. 2), II-28-II-42.
- Hill, C. D., & Langer, M. M. (2007). PlotIRT: A collection of R functions to plot curves associated with Item Response Theory. *Applied Psychological Measurement, 31*(5), 456.

- Hill, L. G., Coie, J. D., Lochman, J. E., & Greenberg, M. T. (2004). Effectiveness of early screening for externalizing problems: Issues of screening accuracy and utility. *Journal of Consulting and Clinical Psychology, 72*, 809-820.
- Hinshaw, S. P. (2005). The stigmatization of mental illness in children and parents: Developmental issues, family concerns, and research needs. *Journal of Child Psychology & Psychiatry, 46*, 714-734.
- Hoagwood, K., & Erwin, H. D. (1997). Effectiveness of school-based mental health services for children: A 10-year research review. *Journal of Child and Family Studies, 6*, 435-451.
- Hoagwood, K., & Johnson, J. (2003). School psychology: A public health framework. I. From evidence-based practices to evidence-based policies. *Journal of School Psychology, 41*, 3-21.
- Hooven, C., Gottman, J. M., & Katz, L. F. (1995). Parental meta-emotion structure predicts family and child outcomes. *Cognition and Emotion, 9*, 229-264.
- Horwitz, S. M., Leaf, P. J., & Leventhal, J. M. (1998). Identification of psychosocial problems in pediatric primary care: Do family attitudes make a difference? *Archives of Pediatrics & Adolescent Medicine, 152*, 367-371.
- Horwitz, S. M., Leaf, P. J., Leventhal, J. M., Forsyth, B., & Speechley, K. N. (1992). Identification and management of psychosocial and developmental problems in community-based, primary care pediatric practices. *Pediatrics, 89*, 480-485.
- Hudley, C., & Graham, S. (1993). An attributional intervention to reduce peer directed aggression among African-American boys. *Child Development, 64*, 124-138.
- Individuals with Disabilities Education Act of 1990, Pub. L. 101-476, (1990), Pub. L. 105-17.20 C.F.R. § 1400 et seq. (reauthorized 1997, 2004).
- Jellinek, M. S. (1997). DSM-PC: Bridging pediatric primary care and mental health services. *Journal of Developmental & Behavioral Pediatrics, 18*, 173-174.
- Jellinek, M. S., Little, M., Murphy, J. M., & Pagano, M. (1995). The Pediatric Symptom Checklist: Support for a role in a managed care environment. *Archives of Pediatrics & Adolescent Medicine, 149*, 740-746.
- Jellinek, M. S., & Murphy, J. M. (1988). Screening for psychosocial disorders in pediatric practice. *American Journal Diseases of Children, 142*, 1153-1158.
- Jellinek, M. S., Murphy, J. M., & Burns, B. J. (1986). Brief psychosocial screening in outpatient pediatric practice. *Journal of Pediatrics, 109*, 371-378.

- Jellinek, M. S., Murphy, J. M., Little, M., Pagano, M. E., Comer, D. M., & Kelleher, K. J. (1999). Use of the Pediatric Symptom Checklist to screen for psychosocial problems in pediatric primary care: A national feasibility study. *Archives of Pediatrics & Adolescent Medicine, 153*, 254-260.
- Jutte, D. P., Burgos, A., Mendoza, F., Ford, C. B., & Huffman, L. C. (2003). Use of the Pediatric Symptom Checklist in a low-income, Mexican American population. *Archives of Pediatrics & Adolescent Medicine, 157*, 1169-1176.
- Kagan, J. (1992). Behavior, biology, and the meaning of temperamental constructs. *Pediatrics, 90*, 510-513.
- Kagan, J. (1997). Conceptualizing psychopathology: The importance of developmental profiles. *Development and Psychopathology, 9*, 321-334.
- Kagan, J., Snidman, N., & Arcus, D. (1998). Childhood derivatives of high and low reactivity in infancy. *Childhood Development, 69*, 1483-1493.
- Kagan, J., Snidman, N., McManis, M., Woodward, S., & Hardway, C. (2002). One measure, one meaning: Multiple measures, clearer meaning. *Development and Psychopathology, 14*, 463-475.
- Kataoka, S. H., Zhang, L., & Wells, K. B. (2002). Unmet need for mental health care among U.S. children: Variation by ethnicity and insurance status. *American Journal of Psychiatry, 159*, 1548-1555.
- Kauffman, J. M. (1999). How we prevent the prevention of emotional and behavioral disorders. *Exceptional Children, 65*, 448-468.
- Kazdin, A. E., & Wassell, G. (1999). Barriers to treatment participation and therapeutic change among children referred for conduct disorder. *Journal of Clinical Child Psychology, 28*, 160-172.
- Keenan, K., Shaw, D. S., Walsh, B., Delliquadri, E., & Giovannelli, J. (1997). DSM-III-R disorders in preschool children from low-income families. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*, 620-627.
- Keenan, K., & Wakschlag, L. S. (2000). More than the terrible twos: The nature and severity of behavior problems in clinic-referred preschool children. *Journal of Abnormal Child Psychology, 28*, 33-46.
- Kellam, S. G., Ling, X., Merisca, R., Brown, C. H., & Ialongo, N. (1998). The effect of the level of aggression in the first grade on the course and malleability of aggressive behavior into middle school. *Development and Psychopathology, 10*, 165-185.

- Kelleher, K. J., McInerney, T. K., Gardner, W. P., Childs, G. E., & Wasserman, R. C. (2000). Increasing identification of psychosocial problems: 1979-1996. *Pediatrics, 105*, 1313-1321.
- Kelleher, K. J., & Wolraich, M. L. (1996). Diagnosing psychological problems. *Pediatrics, 97*, 899-901.
- Kupersmidt, J. B., Burchinal, M., & Patterson, C. J. (1995). Developmental patterns of childhood peer relations as predictors of externalizing behavior problems. *Development and Psychopathology, 7*, 825-843.
- Lambert, M. C., Samms-Vaughan, M. E., Fairclough, M., Schmitt, N., Jeong Shin An, N., & Nutter, C. A. (2003). Is it prudent to administer all items for each Child Behavior Checklist cross-informant syndrome? Evaluating the psychometric properties of the Youth Self-Report dimensions with confirmatory factor analysis and item response theory. *Psychological Assessment, 15*, 550-568.
- Lavigne, J. V., Arend, R., Rosenbaum, D., Binns, H. J., Christoffel, K. K., Burns, A., et al. (1998). Mental health service use among young children receiving pediatric primary care. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*, 1175-1183.
- Lavigne, J. V., Arend, R., Rosenbaum, D., Binns, H. J., Christoffel, K. K., & Gibbons, R. D. (1998). Psychiatric disorders with onset in the preschool years: I. Stability of diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry, 37*, 1246-1254.
- Lavigne, J. V., Binns, H. J., Christoffel, K. K., Rosenbaum, D., Arend, R., Smith, K., et al. (1993). Behavioral and emotional problems among preschool children in pediatric primary care: Prevalence and pediatricians' recognition. *Pediatrics, 91*, 649-656.
- Lewis, T. J., Sugai, G., & Colvin, G. (1998). Reducing problem behavior through a school-wide system of effective behavioral support: Investigation of a school-wide social skills training program and contextual interventions. *School Psychology, 27*, 446-459.
- Little, M., Murphy, J. M., Jellinek, M. S., & Bishop, S. J. (1994). Screening 4- and 5-year-old children for psychosocial dysfunction: A preliminary study with the Pediatric Symptom Checklist. *Journal of Developmental & Behavioral Pediatrics, 15*, 191-197.
- Lochman, J. E., Coie, J. D., Underwood, M. K., & Terry, R. (1993). Effectiveness of social relations intervention program for aggressive and nonaggressive, rejected children. *Journal of Consulting and Clinical Psychology, 61*, 1053-1058.

- Lochman, J. E., & Wayland, K. K. (1994). Aggression, social competence, and race as predictors of negative adolescent outcomes. *Journal of the American Academy of Child and Adolescent Psychiatry, 33*, 1026-1035.
- Loeber, R. (1990). Development and risk factors of juvenile antisocial behavior and delinquency: A review. *Clinical Psychology Review, 10*, 1-41.
- Loeber, R., & Farrington, D. P. (1998). *Serious and violent juvenile offenders: Risk factors and successful interventions*. Thousand Oaks, CA: SAGE Publications.
- Loney, J., & Milich, R. (1982). Hyperactivity, inattention, and aggression in clinical practice. In M. L. Wolraich & D. Routh (Eds.), *Advances in developmental and behavioral pediatrics* (Vol. 3, pp. 113-147). Greenwich, CT: JAL.
- Lord, F. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.
- Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lord, F., & Novick, M. R. (Eds.). (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Maccoby, E. E. (1988). Gender as a social category. *Developmental Psychology, 24*(6), 755-765.
- MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*, 921-943.
- Maldonado, G., & Greenland, S. (1993). Interpreting model coefficients when the true model form is unknown. *Epidemiology, 4*, 310-318.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.
- McDonald, R. P. (1981). The dimensionality of test and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.
- Merrell, K. W. (2003). *Preschool and Kindergarten Behavior Scales, 2nd ed.* Austin, TX: PRO-ED.
- Merrell, K. W., & Walker, H. M. (2004). Deconstructing a definition: Social Maladjustment versus Emotional Disturbance and moving the EBD field forward. *Psychology in the Schools, 41*, 899-910.

- Merritt, K. A., Thompson, R. J., Keith, B. R., & Johndrow, D. A. (1993). Screening for behavioral and emotional problems in pediatric primary care. *Journal of Developmental & Behavioral Pediatrics, 14*, 340-343.
- Mezzacappa, E., Tremblay, R. E., Saul, J. P., Kindlon, D., Arseneault, L., Seguin, J., et al. (1997). Anxiety, antisocial behavior, and heart rate regulation in adolescent males. *Journal of Child Psychology and Psychiatry, 38*, 457-469.
- Miller, P. A., & Eisenberg, N. (1988). The relation of empathy to aggressive and externalizing/antisocial behavior. *Psychological Bulletin, 103*, 324-344.
- Minkovitz, C. S., Hughart, N., Strobino, D., Scharfstein, D., Grason, H., Hou, W., et al. (2003). A practice-based intervention to enhance quality of care in the first 3 years of life: The Healthy Steps for Young Children program. *JAMA: Journal of the American Medical Association, 290*, 3081-3091.
- Moffitt, T. (1994). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674-701.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.
- Murphy, J. M., Ichinose, C., Hicks, R. C., Kingdon, D., Crist-Whitzel, J., Jordan, P., et al. (1992). Utility of the Pediatric Symptom Checklist as a psychosocial screen to meet the federal Early and Periodic Screening, Diagnosis, and Treatment (EPSDT) standards: A pilot study. *The Journal of Pediatrics, 129*, 864-869.
- National Advisory Mental Health Council Workgroup on Child and Adolescent Mental Health Intervention Development and Deployment. (2001). *Blueprint for change: Research on child and adolescent mental health*. Washington, DC: National Institute of Mental Health.
- National Association of Social Workers. (2000). *Code of ethics of the National Association of Social Workers*. Washington, DC: Author.
- National Center for Health Statistics. (1989). *Current estimates from the National Health Interview Survey: United States 1988. Vital and Health Statistics*. Washington, DC: U.S. Government Printing Office.
- Navon, M., Nelson, D., Pagano, M., & Murphy, J. M. (2001). Use of the Pediatric Symptom Checklist in strategies to improve preventive behavioral health care. *Psychiatric Services, 52*, 800-804.
- New Freedom Commission on Mental Health. (2003). *Achieving the promise: Transforming mental health care in America. Final report*. Rockville, MD: DHHS Pub. No. SMA-03-3832.

- Nugent, W. R. (2003). A psychometric study of the Multi-Problem Screening Inventory depression subscale using item response and generalizability theories. *Research on Social Work Practice, 13*, 65-79.
- Nugent, W. R. (2005). The development and psychometric study of an ultra-short-form suicidal ideation measure. *Best Practice in Mental Health: An International Journal, 1*, 1-18.
- Nugent, W. R. (2006). A psychometric study of the MPSI suicidal thoughts subscale. *Stress, Trauma & Crisis: An International Journal, 9*, 1-15.
- Nugent, W. R., & Hankins, J. A. (1992). A comparison of classical, item response, and generalizability theories of measurement. *Journal of Social Service Research, 16*, 11-39.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory, 3rd ed.* New York: McGraw-Hill.
- Olds, D., Henderson, C. R. J., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., et al. (1998). Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. *JAMA: Journal of the American Medical Association, 280*, 1238-1244.
- Olfson, M., Marcus, S. C., Weissman, M. M., & Jensen, P. S. (2002). National trends in the use of psychotropic medications by children. *Journal of the American Academy of Child & Adolescent Psychiatry, 41*, 514-521.
- Oosterlaan, J., Logan, G. D., & Sergeant, J. A. (1998). Response inhibition in AD/HD, CD, comorbid AD/HD + CD, anxious, and control children: A meta-analysis of studies with the stop task. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 39*, 411-425.
- Osterlind, S. J. (1983). *Test item bias*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-030. Newbury Park, CA: Sage Publications.
- Pagano, M., & Murphy, J. M. (1996). Screening for psychosocial problems in 4-5-year-olds during routine EPSDT examinations: Validity and reliability in a Mexican-American sample. *Clinical Pediatrics, 35*, 139-146.
- Parcel, T. L., & Menaghan, E. G. (1988). *Measuring behavioral problems in a large cross sectional survey: Reliability and validity for children of the NLS youth*. Columbus, OH: Center for Human Resource Research.

- Parsons, C. K., & Hulin, C. L. (1982). An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. *Journal of Applied Psychology, 67*, 826-834.
- Parsons, T. (1951). *The social system*. New York: Free Press.
- Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (1989). A developmental perspective on antisocial behavior. *American Psychologist, 44*, 329-335.
- Patterson, G. R., Reid, J., & Dishion, T. J. (1992). *Antisocial boys*. Eugene, OR: Castalia Press.
- Pawluch, D. (1983). Transitions in pediatrics: A segmental analysis. *Social Problems, 30*, 449-465.
- Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and behavior problems in children. *Journal of Marriage and the Family, 48*, 295-307.
- Pianta, R. C., & Cox, M. J. (1999). *The transition to kindergarten. A series from the National Center for Early Development and Learning*. York, PA: Paul H. Brookes Publishing.
- Pransky, J. (1991). *Prevention: The critical need*. Springfield, MO: Burrell Foundation & Paradigm Press.
- R Development Core Team. (2007). R: A language and environment for statistical computing. Vienna, Austria.
- Raine, A., Reynolds, C., Venables, P. H., Mednick, S. A., & Farrington, D. P. (1998). Fearlessness, stimulation-seeking, and large body size at age 3 years as early predispositions to childhood aggression at age 11 years. *Archives of General Psychiatry, 55*, 745-751.
- Ramsay, J. O. (2000). *TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data*. Montreal: McGill University.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.
- Redden, S. C., Forness, S. R., Ramey, S. L., Ramey, C. T., & Brezaussek, C. M. (2003). Mental health and special education outcomes of Head Start children followed into elementary school. *NHSA Dialog: A Research-to-Practice Journal for the Early Intervention Field, 6*, 87-110.
- Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of Health-Related Quality of Life

- item banks plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(5 Suppl 1), S22-S31.
- Reid, J. B. (1993). Prevention of conduct disorder before and after school entry: Relating interventions to developmental findings. *Development and Psychopathology*, 5, 243-262.
- Regier, D. A., Goldberg, I. D., & Taube, C. (1978). The de facto U.S. mental health service system. *Archives of General Psychiatry*, 35, 685-693.
- Reijneveld, S. A., Brugman, E., Verhulst, F. C., & Verloove-Vanhorick, S. P. (2004). Identification and management of psychosocial problems among toddlers in Dutch preventive child health care. *Archives of Pediatrics & Adolescent Medicine*, 158, 811-817.
- Reijneveld, S. A., Vogels, A. G. C., Hoekstra, F., & Crone, M. R. (2006). Use of the Pediatric Symptom Checklist for the detection of psychosocial problems in preventive child healthcare. *BMC Public Health*, 6, 197-197.
- Reinke, W. M., & Herman, K. C. (2002). Creating school environments that deter antisocial behaviors in youth. *Psychology in the Schools*, 39, 549-559.
- Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, 27, 133-144.
- Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children: Manual*. Circle Pines, MN: American Guidance.
- Rubin, K. H., Moller, L., & Emptage, A. (1987). The Preschool Behavior Questionnaire: A useful index of behaviour problems in elementary school-age children? *Canadian Journal of Behavioural Science*, 19, 86-100.
- Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 17.
- Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement*, 1, 223-245.
- Sampson, R. J. (1997). Collective regulation of adolescent misbehavior: Validation from eighty Chicago neighborhoods. *Journal of Adolescent Research*, 12, 227-244.
- Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. 277, 918-924.
- Sanua, V. D. (1990). Leo Kanner (1894-1981): The man and the scientist. *Child Psychiatry and Human Development*, 21, 3-23.

- Sawyer, M. G., Whaites, L., Rey, J. M., Hazell, P. L., Graetz, B. W., & Baghurst, P. (2002). Health-related quality of life of children and adolescents with mental disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 41*, 530-537.
- Scaramella, L. V., Conger, R. D., & Simons, R. L. (1999). Parental protective influences and gender-specific increases in adolescent internalizing and externalizing problems. *Journal of Research on Adolescence 9*, 111-130.
- Schuster, M. A., Duan, N., Regalado, M., & Klein, D. J. (2000). Anticipatory guidance: What information do parents receive? What information do they want? *Archives of Pediatrics & Adolescent Medicine 154*, 1191-1198.
- Schwartz, D., Dodge, K. A., Pettit, G. S., & Bates, J. E. (1997). The early socialization of aggressive victims of bullying. *Child Development, 68*, 665-675.
- Shaffer, D., Fisher, P., & Dulcan, M. K. (1996). The NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3): Description, acceptability, prevalence rates, and performance in the MECA study. *Journal of the American Academy of Child & Adolescent Psychiatry, 35*, 865-877.
- Shaffer, D., Gould, M. S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., et al. (1983). A children's global assessment scale (CGAS). *Archives of General Psychiatry, 40*, 1228-1231.
- Shaw, D. S., Keenan, K., & Vondra, J. I. (1994). Developmental precursors of externalizing behavior: Ages 1 to 3. *Developmental Psychology, 30*, 355-364.
- Shaw, D. S., Owens, E. B., Vondra, J. I., & Winslow, E. B. (1996). Early risk factors and pathways in the development of early disruptive behavior problems. *Development and Psychopathology, 8*, 679-699.
- Shaw, D. S., Winslow, E. B., Owens, E. B., Vondra, J. I., Cohn, J. F., & Bell, R. Q. (1998). The development of early externalizing problems among children from low-income families. *Journal of Abnormal Child Psychology, 26*, 95-107.
- Simonian, S. J., & Tarnowski, K. J. (2001). Utility of the Pediatric Symptom Checklist for behavioral screening of disadvantaged children. *Child Psychiatry & Human Development, 31*, 269-278.
- Simonian, S. J., Tarnowski, K. J., Stancin, T., Friman, P. C., & Atkins, M. S. (1991). Disadvantaged children and families in pediatric primary care settings: II. Screening for behavior disturbance. *Journal of Clinical Child Psychology, 20*, 360-371.

- Simpson, J. S., Jivanjee, P., Koroloff, N., Doerfler, A., & Garcia, M. (2001). *Systems of care: Promising practices in early childhood mental health, 2001 Series, Volume III*. Washington, DC: Center for Effective Collaboration and Practice, American Institutes for Research.
- Sinclair, E., Del'Homme, M., & Gonzalez, M. (1993). Systematic screening for preschool behavioral disorders. *Behavioral Disorders, 18*, 177-188.
- Skiba, R., & Grizzle, K. (1992). Qualifications vs. logic and data: Excluding conduct disorders from the SED definition. *School Psychology Review, 20*, 580-598.
- Snodgrass, J. (1984). William Healy (1869-1963): Pioneer child psychiatrist and criminologist. *Journal of the History of the Behavioral Sciences, 20*, 332-339.
- Spencer, M. S., Fitch, D., Grogan-Kaylor, A., & McBeath, B. (2005). The equivalence of the Behavior Problem Index across U.S. ethnic groups. *Journal of Cross-Cultural Psychology, 36*, 573-589.
- SPSS. (2007). *Statistical Package for the Social Sciences, 15.0*. Chicago: SPSS Inc.
- Stark, S. (2002). *MODFIT [computer program]*. Urbana Champaign, IL: University of Illinois IRT Modeling Lab.
- Stein, R. E. K., Horwitz, S. M., Storfer-Isser, A., Heneghan, A., Olson, L., & Hoagwood, K. E. (2008). Do pediatricians think they are responsible for identification and management of child mental health problems? Results of the AAP Periodic Survey. *Ambulatory Pediatrics, 8*(1), 11-17.
- Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*, 81-97.
- Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*(4), 402-415.
- Stevenson, J., Thompson, M. J., & Sonuga-Barke, E. (1996). Mental health of preschool children and their mothers in a mixed urban/rural population. III: Latent variable models. *British Journal of Psychiatry, 168*, 26-32.
- Strong, K., Wald, N., Miller, A., & Alwan, A. (2005). Current concepts in screening for noncommunicable disease: World Health Organization Consultation Group report on methodology of noncommunicable disease screening. *Journal of Medical Screening, 12*, 12-19.
- Task Force on Research Diagnostic Criteria: Infancy and Preschool. (2003). Research diagnostic criteria for infants and preschool children: The process and empirical

- support. *Journal of the American Academy of Child & Adolescent Psychiatry*, 42, 1504-1512.
- Tay-Lim, B. S.-H., & Harwell, M. (1997). *Effects of number of items and examinees on parameter estimation in item response theory: A research synthesis*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Teresi, J. A. (2001). Statistical methods of examination of differential item functioning with applications to cross-cultural measurement of functional, physical, and mental health. *Journal of Mental Health and Aging*, 7, 31-40.
- Thissen, D. (2001). *Manual for IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning*. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.
- Thissen, D., Chen, W.-H., & Bock, R. D. (2003). *MULTILOG 7.03 [computer software]*. Lincolnwood, IL: Scientific Software International.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567-577.
- Thomasgard, M., & Metz, W. P. (2004). Promoting child social-emotional growth in primary care settings: Using a developmental approach. *Clinical Pediatrics*, 43, 119-127.
- Thun-Hohenstein, L., & Herzog, S. (2008). The predictive value of the Pediatric Symptom Checklist in 5-year-old Austrian children. *European Journal of Pediatrics*, 167(3), 323-329.
- Tremblay, R. E., Pihl, R. O., Vitaro, F., & Dobkin, P. L. (1994). Predicting early onset of male antisocial behavior from preschool behavior. *Archives of General Psychiatry*, 51, 732-739.
- Tuchman, G. (1996). Invisible differences: On the management of children in postindustrial society. *Sociological Forum*, 11, 3-23.
- Tucker, C. J., Marx, J., & Long, L. (1998). "Moving on": Residential mobility and children's school lives. *Sociology of Education*, 71, 111-129.
- U.S. Department of Agriculture Economic Research Service. (2008). County-Level Unemployment and Median Household Income for Kentucky. Retrieved January 20, 2008, from <http://www.ers.usda.gov/Data/Unemployment/RDList2.asp?ST=KY>.

- U.S. Department of Education. (2003). *25th annual report to Congress on the implementation of the Individuals with Disabilities Education Act, Vol. 1*. Washington, DC: Office of Special Education and Rehabilitative Services, Office of Special Education Programs.
- U.S. Department of Health and Human Services. (1999). *Mental health: A report of the Surgeon General*. Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health.
- U.S. Department of Health and Human Services. (2001). *Mental health: Culture, race, and ethnicity. A supplement to Mental health: A report of the Surgeon General*. Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health.
- U.S. General Accounting Office. (2003). *Medicaid and SCHIP: States use varying approaches to monitor children's access to care*. Washington, DC: U.S. General Accounting Office, GAO-03-222.
- Van Acker, R., Grant, S. H., & Henry, D. (1996). Teacher and student behavior as a function of risk for aggression. *Education & Treatment of Children, 19*, 316-334.
- van den Boom, D. C. (1994). The influence of temperament and mothering on attachment and exploration: An experimental manipulation of sensitive responsiveness among lower-class mothers with irritable infants. *Child Development, 65*, 1457-1477.
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22-29.
- Walker, H. M., Colvin, G., & Ramsey, E. (1995). *Antisocial behavior in schools: Strategies and best practices*. Pacific Grove, CA: Brooks/Cole.
- Ware, J. E. (2003). Conceptualization and measurement of health-related quality of life: Comments on an evolving field. *Archives of Physical Medical Rehabilitation, 84*(Supp. 2), S43-S51.
- Wehby, J. H., Dodge, K. A., & Valente, E. (1993). School behavior of first grade children identified as at-risk for development of conduct problems. Special issue: Behavioral disorders in young children. *Behavioral Disorders, 19*, 67-78.
- Werner, E. E. (1984). Resilient children. *Young Children, 40*, 68-72.

- Wolraich, M. L., Felice, M. E., & Drotar, D. (Eds.). (1996). *The classification of child and adolescent mental diagnoses in primary care: Diagnostic and Statistical Manual for Primary Care (DSM-PC)*. Elk Grove Village, IL: American Academy of Pediatrics.
- Woodwell, D. A. (1999). *National ambulatory medical care survey: 1997 summary. Advance data from vital and health statistics, no. 305*. Hyattsville, MD: National Center for Health Statistics.
- Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.
- Yoshikawa, H., & Knitzer, J. (1997). *Lessons from the field: Head Start mental health strategies to meet changing needs*. New York: National Center for Children in Poverty.
- Zigler, E., & Styfco, S. J. (1994). Head Start: Criticisms in a constructive context. *American Psychologist, 49*, 127-132.
- Zill, N. (1985). *Behavior problem scales developed from the 1981 Child Health Supplement to the National Health Interview Survey*. Washington, DC: Child Trends.
- Zill, N. (1990). *Behavior problem index based on parent report*. Washington, DC: Child Trends.
- Zuckerman, B., Moore, K. A., & Gleib, D. (1996). Association between child behavior problems and frequent physician visits. *Archives of Pediatrics & Adolescent Medicine, 150*, 146-153.

APPENDIX A

Diagnostic Criteria for Oppositional Defiant Disorder, Conduct Disorder, and Disruptive Behavior Disorder Not Otherwise Specified (APA, 2000)

Summary of DSM-IV criteria for diagnosis of Oppositional Defiant Disorder (ODD):

A recurrent pattern of negativistic, defiant, disobedient, and hostile behavior toward authority figures that persists for at least six months and is characterized by the frequent occurrence of at least four of the following: a) losing temper; b) arguing with adults; c) actively defying or refusing to comply with the requests or rules of adults; d) deliberately doing things to annoy others; e) blaming others for own mistakes or misbehavior; f) being touchy or easily annoyed by others; g) being angry and resentful; or h) being spiteful and vindictive. These behaviors must occur more frequently than is typically seen in individuals of comparable age and developmental level and must lead to significant impairment in social, academic, or occupational functioning. Behavioral indicators include persistent stubbornness; resistance to directions; unwillingness to compromise, give in, or negotiate with adults or peers; deliberate testing of limits, usually by ignoring orders, arguing, and failing to accept blame for misdeeds; and verbal aggression. ODD is also associated with highly reactive temperament, high motor activity, low frustration

tolerance, and frequent conflicts with others. Prevalence rates have ranged in studies from 2% to 16%, depending on population and method of assessment.

Summary of DSM-IV criteria for diagnosis of Conduct Disorder (CD): A repetitive and persistent pattern of behavior in which the basic rights of others or major age-appropriate societal norms or rules are violated. Consists of four grouping of behaviors: aggressive conduct that causes or threatens physical harm to people or animals; nonaggressive conduct causing property loss or damage; deceitfulness or theft; and serious violations of rules. Three or more of the following behaviors in the above categories must have been present in the past 12 months, and at least one in the past 6 months: a) bullying, threatening, intimidating, or starting frequent physical fights; b) use of weapons which can cause serious harm; c) being physically cruel to humans or animals; d) stealing while confronting a victim; e) forcing someone into sexual activity; f) deliberately destroying property or breaking in to property; g) frequent lying; or h) stealing items of nontrivial value without confronting the victim. Must cause clinically significant impairment in social, academic, or occupational functioning. Includes two subtypes: childhood-onset (early starters) vs. adolescent-onset (late starters). Can be specified as mild, moderate, or severe. CD is associated with lack of empathy for others, misperception of intentions of others as hostile, lack of feelings of remorse, poor frustration tolerance, irritability, temper outbursts, and recklessness. Gender differences are apparent in types of behaviors exhibited. Rates vary widely depending on population sampled and method of assessment: for males 6% to 16%, and for females 2% to 9%. CD is one of the most

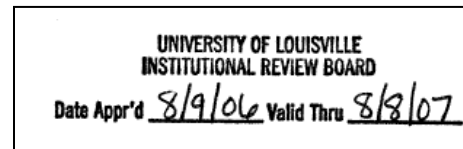
frequently diagnosed conditions in outpatient and inpatient mental health facilities for children.

Summary of DSM-IV criteria for diagnosis of Disruptive Behavior Disorder Not

Otherwise Specified: This category is reserved for oppositional defiant or conduct problems which do not meet the full diagnostic criteria for Oppositional Defiant Disorder or Conduct Disorder, yet pose clinically significant functional impairments.

APPENDIX B

Preamble Consent



■ KENT SCHOOL
OF SOCIAL WORK
Oppenheimer Hall
University of Louisville
Louisville, KY 40292

IMPROVING SCREENING FOR EXTERNALIZING BEHAVIOR PROBLEMS IN VERY YOUNG CHILDREN

September 1, 2006

Dear Parent/Caregiver:

You are being invited to participate in a research study by answering the attached survey about your child between the ages of 3 and 5. We are interested in seeing how well the questions often used to measure child behaviors problems actually work with preschool-aged children. We will not be asking you to put your name or your child's name on the questionnaire, and your answers will be kept private. Your answers will not be shared by us with your child's doctor, but you are welcome to talk about your answers with your child's doctor if you choose to. The survey will take approximately 10 to 15 minutes to complete.

There are no known risks for your being in this research study. The information collected may not benefit you directly, but it may be helpful to others. The information you provide in this survey will be used in a study focused on improving the measurement of behavior problems in young children in primary care settings. Your completed survey will be stored at the University of Louisville, in a locked office in the Carmichael Building.

As a study participant, you are invited to enter a drawing for one of five \$100 Target gift cards, with winners randomly selected from all study participants who choose to enter the drawing (expected to be about 1,000 people). The raffle will be held in summer 2007.

Individuals from the Kent School of Social Work, the Institutional Review Board (IRB), the Human Subjects Protection Program Office (HSPPO), and other regulatory agencies may inspect these records. In all other respects, however, the data will be held in confidence to the extent permitted by law. Should the data be published, your identity will not be disclosed.

Taking part in this study is voluntary. By completing this survey you agree to take part in this research study. You do not have to answer any questions that make you

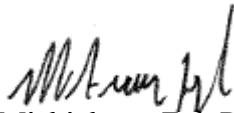
uncomfortable. You may choose not to take part at all. If you decide to be in this study you may stop taking part at any time. If you decide not to be in this study or if you stop taking part at any time, you will not lose any benefits for which you may qualify. The contact information you provide for the Target gift card drawing will not be linked with your completed survey. After the gift card drawing is completed, all contact information you provide will be destroyed by shredding.

If you have any questions, concerns, or complaints about the research study, please contact: Michiel van Zyl, Ph.D. (852-2430) or Christina Studts, M.S.W. (418-3557).

If you have any questions about your rights as a research subject, you may call the Human Subjects Protection Program Office at (502) 852-5188. You can discuss any questions about your rights as a research subject, in private, with a member of the Institutional Review Board (IRB). You may also call this number if you have other questions about the research, and you cannot reach the research staff, or want to talk to someone else. The IRB is an independent committee made up of people from the University community, staff of the institutions, as well as people from the community not connected with these institutions. The IRB has reviewed this research study.

If you have concerns or complaints about the research or research staff and you do not wish to give your name, you may call 1-877-852-1167. This is a 24 hour hot line answered by people who do not work at the University of Louisville.

Sincerely,



Michiel van Zyl, Ph.D.
Kent School of Social Work
University of Louisville



Christina R. Studts, M.S.W.
Kent School of Social Work
University of Louisville

APPENDIX C

Eligibility Checklist and Script to Invite Participation

	ELIGIBLE	NOT ELIGIBLE
1. “What are the ages of the children of whom you are the parent or primary caregiver?”	<input type="radio"/> Any ages 3-5*	<input type="radio"/> None ages 3-5
2. “Can you speak and read English?”	<input type="radio"/> Yes	<input type="radio"/> No
3. “How old are you?”	<input type="radio"/> 18+	<input type="radio"/> Under 18
4. “Is your child here for an <i>emergency</i> appointment?”	<input type="radio"/> No	<input type="radio"/> Yes
5. “Have you already participated in this study?”	<input type="radio"/> No	<input type="radio"/> Yes

TO BE ELIGIBLE, ALL FIVE RESPONSES MUST INDICATE ELIGIBILITY.

Eligible for study participation?	<input type="radio"/> Yes	<input type="radio"/> No
-----------------------------------	------------------------------	-----------------------------

*If eligible and has more than one child between the ages of 3 and 5, instruct participant to select the child in that age range who had the most recent birthday.

If a potential participant is eligible, use the following script to invite their participation:

“You are invited to participate in a research study that is looking at how well certain questions work with children ages 3 to 5 to measure behavior problems. We would like to see which questions work best with young children and are fair with children of all races and backgrounds. If you agree to participate, you will fill out two short questionnaires about your child’s behavior. We also would like for you to answer a third set of questions that will tell us about your child’s background, your background, and some additional information about how you see your child’s behavior.”

APPENDIX D

PSC-17 (Gardner et al., 1999)

PSC-17

For each item, please mark under the heading that best fits your child:	Never	Sometimes	Often
1. Fidgety, unable to sit still	O	O	O
2. Feels sad, unhappy	O	O	O
3. Daydreams too much	O	O	O
4. Refuses to share	O	O	O
5. Does not understand other people's feelings	O	O	O
6. Feels hopeless	O	O	O
7. Has trouble concentrating	O	O	O
8. Fights with other children	O	O	O
9. Is down on him or herself	O	O	O
10. Blames others for his or her troubles	O	O	O
11. Seems to be having less fun	O	O	O
12. Does not listen to rules	O	O	O
13. Acts as if driven by a motor	O	O	O
14. Teases others	O	O	O
15. Worries a lot	O	O	O
16. Takes things that do not belong to him or her	O	O	O
17. Distracted easily	O	O	O

APPENDIX E

Scoring Instructions for PSC-17 (Gardner et al., 1999)

Scoring instructions:

For each item, “never” = 0, “sometimes” = 1, and “often” = 2.

PSC17-Externalizing = Sum of scores for items 4, 5, 8, 10, 12, 14, and 16: _____

PSC17-Internalizing = Sum of scores for items 2, 6, 9, 11, and 15: _____

PSC17-Attention = Sum of scores for items 1, 3, 7, 13, and 17: _____

PSC17 Total Score = Sum of PSC17-E + PSC17-I + PSC17-A: _____

Positive scores:

PSC17-E \geq 7

PSC17-I \geq 5

PSC17-A \geq 7

PSC17 Total Score \geq 7

APPENDIX F

BPI (Peterson & Zill, 1986; Zill, 1990)

BPI

Here are some statements that describe behavior problems many children have. Please mark whether each statement is not true, sometimes true, or often true of your child during the past 3 months.

	Not True	Sometimes True	Often True
1. Has sudden changes in mood or feelings	O	O	O
2. Feels or complains that no one loves him/her	O	O	O
3. Is rather high strung, nervous, or tense	O	O	O
4. Cheats or tells lies	O	O	O
5. Is too fearful or anxious	O	O	O
6. Argues too much	O	O	O
7. Has difficulty concentrating, cannot pay attention for long	O	O	O
8. Is easily confused, seems to be in a fog	O	O	O
9. Bullies, or is cruel or mean to others	O	O	O
10. Is disobedient at home	O	O	O
11. Does not seem to be sorry after he/she misbehaves	O	O	O
12. Has trouble getting along with other children	O	O	O
13. Is impulsive, or acts without thinking	O	O	O
14. Feels worthless or inferior	O	O	O
15. Is not liked by other children	O	O	O
16. Has a lot of difficulty getting his/her mind off certain thoughts, has obsessions	O	O	O
17. Is restless or overly active, cannot sit still	O	O	O
18. Is stubborn, sullen, or irritable	O	O	O

	Not True	Sometimes True	Often True
19. Has a very strong temper and loses it easily	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
20. Is unhappy, sad, or depressed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
21. Is withdrawn, does not get involved with others	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
22. Breaks things on purpose, deliberately destroys his/her own things or others' things	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
23. Clings to adults	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
24. Cries too much	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
25. Demands a lot of attention	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
26. Is too dependent on others	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

APPENDIX G

Scoring Instructions for BPI (Peterson & Zill, 1986; Zill, 1990)

Scoring instructions:

For each item, “not true” = 0, “sometimes true” = 1, and “often true” = 2.

BPI-Antisocial = Sum of scores for items 4, 9, 11, and 22: _____

BPI-Headstrong = Sum of scores for items 3, 6, 10, 18, and 19: _____

BPI-Peer Problems = Sum of scores for items 12, 15, and 21: _____

BPI-Anxious/Depressed = Sum of scores for items 1, 2, 5, 14, and 20: _____

BPI-Dependent = Sum of scores for items 23, 24, 25, and 26: _____

BPI-Hyperactive = Sum of scores for items 7, 8, 13, 16, and 17: _____

BPI-Externalizing = BPI-Anti + BPI-H + BPI-PP – item 21: _____

BPI Total Score = BPI-Anti + BPI-H + BPI-PP + BPI-A/D + BPI-D + BPI-H: _____

Positive scores (based on 90th percentile *dichotomized* scores for children ages 4-5)¹:

BPI-Anti \geq 3

BPI-H \geq 5

BPI-PP \geq 1

BPI-A/D \geq 3

BPI-D \geq 3

BPI-H \geq 4

BPI Total Score \geq 15

¹Center for Human Resource Research (2000); dichotomized scores so that 0 = 0 and 1 or 2 = 1, so will not be applicable when items are not dichotomized.

APPENDIX H

Sociodemographic Questionnaire

DEMOGRAPHIC QUESTIONNAIRE

Thank you for being in this study. Your answers will help us learn about how these questions work with preschool-aged children. In this final part of the survey, please respond to these questions about your child and about you, to help us know more about the people in the study. Your name and your child's name will *not* be on the survey.

Questions about YOUR CHILD

Questions about YOU

Your CHILD'S age: _____ years

YOUR age: _____ years

Your CHILD'S sex: Male Female

YOUR sex: Male Female

Your CHILD'S race:

YOUR race:

- Caucasian (White) (1)
- African-American (2)
- Hispanic (3)
- Asian (4)
- Other ↴ (5)

- Caucasian (White) (1)
- African-American (2)
- Hispanic (3)
- Asian (4)
- Other ↴ (5)

Please specify: _____

Please specify: _____

Your CHILD'S primary household:

YOUR annual household income range:

- Two-parent household (1)
- Single-parent household (2)
- Caregiver other than parent (3)
- ↳ If other than parent, who? _____

- \$10,000 or less (1)
- \$10,001 - 20,000 (2)
- \$20,001 - 30,000 (3)
- \$30,001 - 40,000 (4)
- \$40,001 - 50,000 (5)
- \$50,001 - 60,000 (6)
- \$60,001 - 70,000 (7)
- \$70,001 - 80,000 (8)
- \$80,001 - 90,000 (9)
- Over \$90,000 (10)

Your CHILD'S number of siblings at home:

_____ # sisters
 _____ # brothers

Does your CHILD attend:

Preschool? YES NO # hours per week: _____
 Daycare? YES NO # hours per week: _____

—————▶ **How many years of education have YOU completed? (Circle one)** ◀—————

K 1 2 3 4 5 6 7 8 9 10 11 12 GED 13 14 15 16 17 18 19+

Your CHILD'S primary health insurance type:

- Medicaid O (1)
- K-Chip O (2)
- Private O (3)
- HMO/PPO O (4)
- None O (5)
- Other → O (6)

Please specify: _____

YOUR relationship to your child:

- Parent O (1)
- Step-parent O (2)
- Grandparent O (3)
- Foster parent O (4)
- Other → O (5)

Please specify: _____

What is the reason for your CHILD'S appointment today?

- Child's regular check-up O (1)
- Child is sick O (2)
- Appointment is for sibling O (3)
- Other O (4) → Please specify: _____

Do YOU think that your CHILD has behavior problems? O Yes O No

Has your CHILD ever been seen by a mental health professional (e.g., psychologist, clinical social worker, psychiatrist, etc.)? O Yes O No

Has your CHILD ever been prescribed medication for behavior? O Yes O No

- If YES, by whom?*
- Regular physician O (1)
 - Psychiatrist O (2)
 - Other O (3) → Please specify: _____

Have YOU ever expressed concerns to your CHILD'S primary care doctor about your CHILD'S behavior? O Yes O No

Has your CHILD'S primary care doctor ever expressed concerns to YOU about your CHILD'S behavior? O Yes O No

Has anyone else (e.g., relative, daycare provider, etc.) ever expressed concerns to YOU about your CHILD'S behavior? O Yes O No

- If YES, who?*
- Relative O (1)
 - Daycare provider O (2)
 - Other O (3) → Please specify: _____

THANK YOU for completing this survey! ☺

CURRICULUM VITAE

Christina R. Studts, M.S.W., ABD

Home Address & Contact Information: 1301 Cooper Drive
Lexington, Kentucky 40502
(859) 523-6976 (home)
(502) 418-3557 (mobile)
tina.studts@louisville.edu

Birth Date: April 16, 1971
Hometown: Los Alamos, New Mexico
Citizenship: U.S.A.
License: L.C.S.W., Kentucky (#1424, 5/2000)

EDUCATION

8/03 – present Doctoral student, Kent School of Social Work, University of
Louisville/University of Kentucky Joint Doctoral Program
Doctoral candidate, 8/05 (Anticipated completion: 5/08)
Dissertation topic: *Improving screening for externalizing behavior
problems in very young children: Applications of item response
theory to evaluate instruments in pediatric primary care*

7/05 – present Master's student (M.S. in Biostatistics), Department of Bioinformatics
and Biostatistics, School of Public Health and Information
Sciences, University of Louisville
Master's candidate, 8/07 (Anticipated completion: 12/08)

5/97 M.S.W., University of Kentucky
5/93 B.A., Psychology, University of Notre Dame

RESEARCH EXPERIENCE

7/05 – present **Graduate Research Assistant**, James Graham Brown Cancer Center,
University of Louisville
Performed data collection, database maintenance, statistical analyses,
literature reviews, and manuscript writing for three ongoing research
projects:

Lung Cancer Screening Study

Colposcopy Study

Lung Cancer Decision Making Study

Supervisor: David Hein, Ph.D., Director of Cancer Prevention and Control Program, James Graham Brown Cancer Center, University of Louisville School of Medicine

- 8/03 – present **University Fellow**, Kent School of Social Work, University of Louisville
Engaged in research activities with several faculty members during studies in the doctoral program, including:
Gerard Barber, Ph.D. and Ramona Stone, Ph.D., Kent School of Social Work: *Health and Welfare Reform*
Andy Frey, Ph.D., Kent School of Social Work: *Positive Behavioral Supports in Head Start*
Jamie L. Studts, Ph.D., James Graham Brown Cancer Center: *Colposcopy Study, Lung Cancer Decision Making Study, Lung Cancer Screening Study*
- 9/92 – 12/92 **Women and AIDS Coalition**, South Bend, Indiana
Performed data entry and conducted preliminary statistical analyses.

RESEARCH SUPPORT

1 R36 HS016940-01 Dissertation Grant, Agency for Healthcare Research and Quality (AHRQ)

Improving Screening for Externalizing Behavior Problems in Very Young Children: Applications of Item Response Theory to Evaluate Instruments in Pediatric Primary Care

This study investigates the performance of items in two commonly used pediatric behavioral screening instruments, with special attention to differences between groups categorized by sex, race, and socioeconomic status.

Role: Principal Investigator (100%)

(AWARDED BUT DECLINED DUE TO EARLY COMPLETION OF STUDY)

Kentucky Lung Cancer Research Program

Behavioral, Cognitive, and Affective Responses to Lung Cancer Screening

This study examines a range of important positive and negative sequelae of participation in a randomized controlled trial of lung cancer screening.

Role: Research Assistant (50%)

(CURRENTLY FUNDED)

Kentucky Department of Health and Human Services

The Manualization and Replication of CATS

Role: Development team member, clinical social worker/supervisor (50%)

(COMPLETED FUNDING)

PUBLICATIONS

Studts, C. R., Stone, R., & Barber, G. M. (2006). Predictors of access to health-care services among groups of TANF recipients in Kentucky. *Social Service Review*, 80(3), 527-548.

Studts, J. L., Ghate, S. R., Gill, J. L., **Studts, C. R.**, Barnes, C. N., LaJoie, A. J., Andrykowski, M. A., & LaRocca, R. V. (2006). Validity of self-reported smoking status among participants in a lung cancer screening trial. *Cancer Epidemiology, Biomarkers and Prevention*, 15(10), 1825-1828.

***Selected as a featured article among the 43 published in this issue.

MANUSCRIPTS IN PREPARATION

Turner, M. D., Barnard, P., Ballard, H., & **Studts, C. R.** (2008). *Sedation practices for pediatric cardiac catheterizations.*

Studts, C. R., Stone, R., & Barber, G. M. (2008). *Time series multilevel modeling of health status change among TANF recipients following welfare reform.*

Studts, J. L., **Studts, C. R.**, & Matera, E. L (2008). *Measuring numeracy: An item response theory analysis of the Numeracy Questionnaire.*

Studts, J. L., **Studts, C. R.**, Sephton, S., & Helm, C. W. (2008). *Comparison of measures of perceived cervical cancer risk among women undergoing colposcopy.*

Studts, J. L., **Studts, C. R.**, LaJoie, A. S., Barnes, C. N., Andrykowski, M. A., & LaRocca, R. (2008). *Predictors of change in self-reported smoking status among participants in a lung cancer screening trial.*

Studts, J. L., Barnes, C. N., LaJoie, A. S., **Studts, C. R.**, Andrykowski, M. A., & LaRocca, R. (2008). *Predictors of adherence to screening protocol among participants in a lung cancer screening trial.*

RESEARCH PRESENTATIONS

Studts, C. R., Studts, J. L., & Andrykowski, M. A. (2008). *Psychometric properties of a new Subjective Numeracy Scale: Classical and IRT analyses.* Poster presented at the 29th annual scientific sessions of the Society of Behavioral Medicine, San Diego, CA.

Studts, J. L., Barnes, C. N., **Studts, C. R.**, LaJoie, A. S., Andrykowski, M. A., & LaRocca, R. (2007). *Participant adherence in a RCT of lung cancer screening: Results from baseline to year 1.* Paper presented at the 28th annual scientific sessions of the Society of Behavioral Medicine, Washington, DC.

Ruberg, J. L., **Studts, C. R.**, Barnes, C. N., LaJoie, A. S., Cross, T., LaRocca, R. V., Andrykowski, M. A., & Studts, J. L. (2007). *Smoking cessation among participants in a RCT of lung cancer screening: Baseline to year one*. Poster presented at the 28th annual scientific sessions of the Society of Behavioral Medicine, Washington, DC.

Studts, C. R., Stone, R., & Barber, G. M. (2007). *A multilevel model of health status change among welfare recipients following welfare reform*. Paper presented at the 11th annual conference of the Society for Social Work and Research, San Francisco, CA.

Studts, C. R., Stone, R., & Barber, G. M. (2006). *Predictors of health care access among welfare recipients*. Poster presented at the 27th annual scientific sessions of the Society of Behavioral Medicine, San Francisco, CA.

Studts, C. R., Matera, E. L., & Studts, J. L. (2006). *An item response theory analysis of the Numeracy Questionnaire*. Poster presented at the 28th annual meeting of the Society of Medical Decision Making, Boston, MA.

Studts, C. R., Sephton, S., Helm, C. W., & Studts, J. L. (2006). *Perceived cervical cancer risk among women undergoing colposcopy*. Poster presented at the 27th annual scientific sessions of the Society of Behavioral Medicine, San Francisco, CA.

Barnes, C. N., **Studts, C. R.**, LaJoie, A. S., Ruberg, J. L., Cross, T., Andrykowski, M. A., LaRocca, R. V., & Studts, J. L. (2006) *Participant adherence in a RCT of lung cancer screening: Baseline to year 1*. Poster presented at Research Louisville, University of Louisville, Louisville, KY.
***2nd Place Award for Student Research

Studts, J. L., Ghate, S., Marmarato, J., Barnes, C., **Studts, C. R.**, LaJoie, A. S., & LaRocca, R. (2006). *Validity of self-reported smoking status among lung cancer screening participants*. Poster presented at the 30th annual meeting of the American Society of Preventive Oncology, Bethesda, MD.

Studts, C. R., Elliott, A., Faith, T., Royer, B., & Young, S. (2004). *Positive behavioral supports in a Head Start classroom*. Paper presented at the annual meeting of the Midwest School Social Workers, Louisville, KY.

INVITED PRESENTATIONS

Studts, C. R. (2007). *Basic biostatistics: An overview for clinicians*. Invited two-part lecture in the Spring Lecture Series of the Department of Ophthalmology, School of Medicine, University of Louisville.

Studts, C. R. (2006). *Introduction to item response theory*. Invited lecture to the Behavioral Oncology Lab, Cancer Prevention & Control Program, James Graham Brown Cancer Center, University of Louisville.

Studts, C. R., Matera, E. L., & Studts, J. L. (2006). *An item response theory analysis of the Numeracy Questionnaire*. Invited lecture in the Fall Lecture Series of the Department of Bioinformatics and Biostatistics, School of Public Health and Information Sciences, University of Louisville.

Studts, C. R. (2005, 2006). *Item response theory: Introduction and application to measurement of health outcomes*. Invited lecture to course PHCI 602, Health Services and Outcomes Research, CREST program, School of Public Health and Information Sciences, University of Louisville.

TEACHING EXPERIENCE

- 1/07 – 3/07 **Research Rotation Supervisor**, School of Medicine Med-Peds Program, University of Louisville
Supervised the research rotations of Demeka Y. Campbell, M.D., and Cynthia Bowman-Stroud, M.D., med-peds residents. Aspects of the rotations included guidance and training on literature reviews, formulating research questions and hypotheses, study design, survey design, data collection, data analysis, and scientific writing.
- 11/05 **Guest Lecturer**, CREST, School of Public Health & Information Sciences
PHCI 602, Health Services and Outcomes Research
Topic: “Measuring Depression and Anxiety”
- 12/04 Developed a masters-level social work elective course, including syllabus, readings, assignments, and examinations: *Social Work Practice in Health Care Settings*
- 9/04 – 12/04 **Tutor**
Provided private tutoring in graduate-level statistics.
- 5/04 – 7/04 **Co-Instructor**, Kent School of Social Work, Louisville, Kentucky
SW 766, Doctoral Preparation
As a teaching practicum, co-taught the summer course for incoming doctoral students, with Ruth Huber, Ph.D. (director of the doctoral program). Assisted with development of course outline and syllabus; prepared and administered the online Blackboard course site; taught portions of class sessions; provided individual and group tutoring to students; and provided feedback on homework and in-class assignments. Course content included a review of the basics of social work research,

statistics, use of SPSS, and significant past and present themes in the social work literature.

3/99 – 3/01 **Teacher**, Kaplan Educational Center, Lexington, KY and Durham, NC
Prepared and taught preparation courses for all portions of the GRE, LSAT, ACT, and SAT standardized tests to classes of up to 20 students as well as to individual students. Also taught the Verbal sections of the PCAT (Pharmacy) and DAT (Dental) standardized tests.

ADVISING ACTIVITY

Completed Advising:

Doctoral Committee (outside reader):

- Shaena Y. Gardner - Clinical Psychology (Spalding University – defended 3/06)

CLINICAL & ADMINISTRATIVE EXPERIENCE

5/04 – 9/05 **FORECAST, Kent School of Social Work**, Louisville, Kentucky

Member of Development Team, Clinician, and Clinical Supervisor

Assisted with development of a clinic designed to provide comprehensive assessments of families involved with the Jefferson County Department of Community Based Services. Contributed to development of protocols, clinic resources, and relevant literature reviews. Provided comprehensive assessments of potential foster/adoptive parents, foster/adoptive families facing possible disruption, and biological parents involved with Child Protective Services, in a clinic funded by the Kentucky Cabinet for Health and Family Services. Consulted with Cabinet staff and supervisors to provide assessments and recommendations. Provided clinical supervision to masters-level certified social workers pursuing independent licensure.

7/01 – 7/03 **Seven Counties Services, Inc.**

Principal Social Worker, School Based Services, Louisville, Kentucky

Provided mental health assessments, treatment planning, and services for elementary school children in the school setting. Consulted with school staff (teachers, guidance counselors, principals) in three Jefferson County elementary schools to provide education and recommendations regarding clients and school populations in general. Advocated for special needs of clients, such as psychoeducational assessments, classroom accommodations, and placements. Coordinated with community agencies (courts, social services) to provide appropriate and effective services. Collaborated with multidisciplinary mental health professionals as part of a treatment team.

Senior Social Worker, Bullitt County Child and Family, Shepherdsville, Kentucky

Provided mental health assessments, treatment planning, and services for children, adolescents, and families in a rural outpatient mental health agency. Coordinated with community agencies (schools, courts, social services) to provide appropriate and effective services for clients.

Collaborated with multidisciplinary mental health professionals as part of a treatment team.

7/00 – 7/01 **Duke University Medical Center**

Clinical Social Worker, Duke Children’s Primary Care, Durham, North Carolina

Provided clinical and case management social work services for a pediatric primary care clinic. Performed crisis assessment and intervention, in addition to ongoing support services. Collaborated with physicians, nurses, psychologists, and other health professionals to optimize family access to treatment and resources. Educated medical residents in the clinic setting on psychosocial/mental health issues and community resources. Coordinated efforts with multiple local agencies to improve patient and family care. Participated in on-call and coverage teams with pediatric clinical social workers throughout the medical center.

5/97 – 6/00 **Bluegrass Regional Mental Health – Mental Retardation Board, Inc.**

Program Director, R.I.S.E. Program, Harrodsburg, Kentucky

Directed a mental health and educational program for 60 children. Hired, trained, and supervised 13 mental health specialists and teachers. Coordinated efforts with local schools and agencies. Maintained administrative and direct service records. Assisted in the development and promotion of a new R.I.S.E. program in Lawrenceburg, Kentucky.

Outpatient Therapist, Harrodsburg and Stanton, Kentucky

Provided mental health assessments, treatment planning, and services for children, adults, and families in the outpatient Comprehensive Care Centers.

Mental Health Specialist, R.I.S.E. Program, Harrodsburg, Kentucky

Provided mental health and educational services to 15 children as part of a multidisciplinary team. Maintained appropriate clinical documentation of services. Assisted program director with administrative tasks and program preparation.

CONSULTING EXPERIENCE

- 2/04 – 3/04 **Kent School of Social Work**, Louisville, Kentucky
Consulted with Masters program research faculty and students to expedite the preparation and submission of MSSW research project proposals to the University of Louisville Institutional Review Board (IRB). Assisted students in preparing and revising protocols to meet ethical and scientific research standards. Collaborated with IRB staff as needed.

SERVICE EXPERIENCE

- 7/04 – present **Metro United Way Success by 6, Nurturing Young Children Action Team**
Louisville, Kentucky
Served on the Nurturing Young Children Action Team to promote collaboration of community programs and services targeting school readiness. Helped initiate a subgroup focusing on child health and safety issues.
- 8/04 – 5/05 **Kent School Doctoral Faculty, Kent School Faculty, and Kent Assembly**
Doctoral student representative to meetings; provided information to doctoral students and solicited input to present to faculty and staff.
- 8/03 – 5/05 **Kent School of Social Work Outcomes Committee**
Served on committee with focus on improving and monitoring outcome measures of the Kent School as required by accrediting bodies.
- 7/00 – 7/01 **Durham Interagency Council for Young Children with Special Needs**, Durham, North Carolina
Collaborated with community leaders toward improving local efforts to identify infants and toddlers with special needs and promoting services for this population. Served as Council Secretary.

PRACTICUM EXPERIENCE

- 1/97 – 5/97 **Domestic Violence Prevention Board**, Lexington, Kentucky
Participated in multidisciplinary and interagency strategic planning groups on state and local levels.
- 8/96 – 12/96 **Bluegrass Regional Mental Health – Mental Retardation Board, Inc.**, Winchester, Kentucky
Performed intake psychosocial assessments, determined preliminary diagnoses, and triaged client assignments to therapists under clinical supervision.

8/95 – 5/96 **Jessamine County School District**, Nicholasville, Kentucky
Developed and facilitated treatment groups in a middle school and an alternative high school under the supervision of an at-risk counselor.

VOLUNTEER EXPERIENCE

8/94 – 5/95 **Family Life Head Start Child Development Center, CAP Volunteer Program**, Mt. Vernon, Kentucky
Supervised and guided the developmental play time of 25 Head Start preschoolers. Created lesson plans and planned activities after conducting assessments of individual children's needs. Full-time volunteer.

8/93 – 8/94 **Family Life Services, CAP Volunteer Program**, Mt. Vernon, Kentucky
Provided extensive follow-up services (parenting, budgeting, problem solving, emotional support) to 30 families who completed a residential program. Assisted with day-to-day operations in the shelter through a wide variety of tasks. Full-time volunteer.

9/92 – 12/92 **University of Notre Dame Crisis Line**, Notre Dame, Indiana
Volunteer Crisis Telephone Peer Counselor

9/91 – 12/91 **St. Mary of the Angels Youth Program**, London, England
Volunteer Staff Member

9/90 – 5/91 **St. Mary's Native American Tutoring Program**, South Bend, Indiana
Volunteer Tutor for elementary school students

TRAINING AND WORKSHOP EXPERIENCE

2007 Society for Social Work and Research, *San Francisco, CA*

2006 Society of Medical Decision Making, *Boston, MA*
Evidence-Based Practice (Eileen Gambrill & Leonard Gibbs), *Louisville, KY*
Society of Behavioral Medicine, *San Francisco, CA*
Latent Class and Latent Profile Analysis: Creating Typologies via Categorical Latent Variables
Communication Skills in Statistical Consulting (Janice Derr), *Louisville, KY*

2005 Society of Behavioral Medicine, *Boston, MA*
Modern Psychometrics and Health Outcomes Assessment
Introduction to Item Response Theory: Methods and Applications
Measurement: Theory and Applications in Social Work Research (William Nugent), *Lexington, KY*

2004 Society of Behavioral Medicine, *Baltimore, MD*

- 2003 Ethics in Social Work Practice, *Louisville, KY*
Clinical Supervision Training for Kentucky Board of Social Work, *Louisville, KY*
- 2002 Kentucky Play Therapy Association Conference, *Louisville, KY*
SCERTS Interventions for Autistic Spectrum Disorders, *Indianapolis, IN*
- 2001 Safe Crisis Management, *Louisville, KY*
HIV/AIDS Awareness Training, *Louisville, KY*
- 2000 Explosive and Inflexible Children, *Lexington, KY*
Expressive Therapies with Sexually Abused Children, *Lexington, KY*
- 1999 V.I.S.I.O.N. Training (Multicultural Issues in Mental Health), *Lexington, KY*
The Canadian Play Therapy Institute, *Lexington, KY*
- 1998 The Mental Health Institute, *Louisville, KY*
- 1997 Domestic Violence Training, *Lexington, KY*
Victims Advocacy Training, *Frankfort, KY*
- 1996 The Canadian Play Therapy Institute, *Lexington, KY*
Kentucky School Social Work Conference, *Louisville, KY*
ADHD Workshop, *Lexington, KY*
- 1995 The Fall Institute: Children and Families First, *Louisville, KY*
Family Literacy: Creating a Community of Learners, *Lexington, KY*

HONORS & AWARDS

- 2007 Travel Awards: University of Louisville Graduate Student Council,
Kent School of Social Work Alumni Fund and Student Association
- 2003 – 2007 University of Louisville Graduate School Fellowship
- 1997 Alpha Delta Mu Honorary Society
- 1996 – 1997 University of Kentucky Graduate School Presidential Fellowship
- 1995 – 1996 University of Kentucky College of Social Work Scholarship
- 1989 – 1993 University of Notre Dame Orchestra
- 1989 – 1993 University of Notre Dame Dean's List

MEMBERSHIP IN PROFESSIONAL ORGANIZATIONS

- 04/06 – present American Statistical Association, Student Member
- 12/03 – present Society for Social Work and Research, Student Member
- 12/03 – present Society of Behavioral Medicine, Student/Trainee Member
- 09/95 – present National Association of Social Workers, Member