Improving Screening for Externalizing Behavior Problems among Very Young Children in Primary Care: Applications of Item Response Theory

Christina Ruth Studts, Ph.D.
University of Louisville
Louisville, Kentucky

Statement of the Research Problem

Preschool-aged children exhibiting early externalizing behavior problems are at high risk of a continuing developmental pathway of antisocial behaviors, as well as low health-related quality of life, lifelong dependence on social services, and an array of other negative long-term consequences (Hann & Borek, 2001). Epidemiological data suggest that nearly 10% of very young children in the U.S. exhibit clinically significant externalizing behavior problems, but the vast majority do not receive needed services (USDHHS, 1999). Primary and secondary prevention efforts, including early identification and early intervention, have been lauded as essential strategies for alleviating this social and public health crisis (Hoagwood & Johnson, 2003). Pediatric primary care is an ideal venue for screening preschool-aged children for externalizing behavior problems (AHRQ, 2002); however, pediatricians persistently under-identify children in need of services (Costello & Edelbrock, 1985; Lavigne et al., 1993). The movement to integrate psychosocial services into primary care (e.g., Blount, 2003; Strosahl et al., 1994) presents an opportunity for social workers to remedy this deficiency.

Use of valid and reliable screening instruments may improve sensitivity in early identification efforts (Hill, Coie, Lochman, & Greenberg, 2004). Screening in primary care requires a brief, easily scored instrument which detects sub-clinical to clinical levels of behavior problems within the context of early childhood development. Existing scales are problematic due to excessive length; cost; and disparities in identification rates of female, minority, and low socioeconomic status (SES) children (Jellinek et al., 1999; Jutte, Burgos, Mendoza, Ford, & Huffman, 2003; Simpson, Jivanjee, Koroloff, Doerfler, & Garcia, 2001; Spencer, Fitch, Grogan-Kaylor, & McBeath, 2005). Thus, to accurately identify preschool-aged children in need of further evaluation and intervention, brief screening measures are needed which precisely measure behavior problems at clinical and sub-clinical levels. In addition, such instruments must perform consistently across diverse groups of very young children.

The purpose of this study was to develop a brief parent-report instrument to screen for externalizing behavior problems in the pediatric primary care setting. The

instrument was intended to meet two key criteria: (a) developmental appropriateness for the preschool age range, and (b) consistent performance across sociodemographically diverse groups of children. To accomplish this goal, modern measurement theory methods were applied to evaluate the quality (i.e., precision and utility) of measurement provided by externalizing subscale items in two existing instruments with preschool-aged children seen in pediatric primary care practices. The analytic methods also allowed the investigation of items for differential functioning between groups differing by child sex, race, and SES. Results were reviewed to identify a set of items most appropriate for screening very young children for externalizing behavior problems in diverse pediatric primary care settings.

Research Background and Hypotheses

In response to increased recognition of the negative long-term consequences associated with early emergence of externalizing behavior problems and other psychosocial issues, pediatric primary care has been identified as an ideal setting for screening and early identification efforts (AHRQ, 2002). Pediatric primary care offers additional resources beyond those offered by the educational and mental health systems to expand primary and secondary prevention practices. Such resources may be especially important for preschool-aged children who are not yet in contact with systems likely to identify older children in need of further assessment and services. While the significance of psychosocial issues in primary care settings has been recognized, primary care physicians—the de facto mental health service providers (Reiger, Goldberg, & Taube, 1978) in the U.S.—have struggled with under-identification of children in need of services (Costello, 1986; Costello & Edelbrock, 1985; 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, the assessment methods favored by most pediatric health providers are typically informal (American Academy of Pediatrics, 2000) and have low sensitivity: Pediatric primary care providers detect only 20% of children with mental health issues identified by psychologists using standardized assessment instruments (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 et al., 1998).

To improve rates of identification in pediatric primary care, standardized screening approaches using reliable and valid instruments are helpful (Halfon, Regalado, McLearn, Kuo, & Wright, 2003; Hill et al., 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, public-domain 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.

Limitations of existing instruments. 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 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 et al., 2003; Simonian & Tarnowski, 2001; Simonian, Tarnowski, Stancin, Friman, & Atkins, 1991; Spencer et al., 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 (USDHHS, 2001), bias in screening instruments can result in both over- and under-identification of children in certain groups, stymieing equitable and appropriately targeted primary and secondary prevention efforts (Spencer et al., 2005) and potentially 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 observed 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, important questions remain regarding the quality of measurement provided by these instruments with the population of interest. Alternative, modern measurement methods are available, however, and their application may overcome the limitations in instrument development inherent in CTT.

The promise of item response theory. Item response theory (IRT) is a modern statistical approach which can 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 with linear transformations, unlike the traditional psychometric indices obtained via CTT methods.

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 at specific levels of the underlying construct of interest (by convention referred to as theta; Hambleton & Swaminathan, 1985). Thus, a given item's measurement precision at various levels of theta 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 theta, but also by extraneous characteristics, such as sex, race, or SES (Teresi, 2001).

The use of IRT-based methods to evaluate the externalizing subscales of the PSC-17 and 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. Investigation of possible DIF in these scales may highlight concerns regarding health disparities and under- and over-identification of minority and low-SES children. A brief set of items could be recommended which provides the most precise and least biased measurement at desired levels of externalizing behavior problems for the target population.

Research questions and hypotheses. To accomplish the aim of this study and identify items providing the most precise and least biased measurement of externalizing behavior problems in preschool-aged children, the following research questions and hypotheses were posed:

> ***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.
> >
> > ***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).

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.

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.

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.

Methodology

A cross-sectional survey design was employed. Consistent with the requirements of IRT analyses (Reise & Yu, 1990), a large sample was selected from four pediatric primary care practices serving diverse populations of children. Eligible participants were present at non-emergency pediatric primary care appointments, were primary caregivers of children aged 3 to 5 years, were age 18 or older, and could provide informed consent and complete the battery of instruments in English. Nonrandom sampling procedures were used, in which a convenience sample of participants was recruited in the waiting rooms of the pediatric practices. Due to unique properties of IRT, this strategy did not limit generalizability of results (Hambleton & Swaminathan, 1985). Participants completed the PSC-17, the BPI, and a sociodemographic questionnaire developed by the author. Descriptive and traditional psychometric analyses were conducted to characterize the study sample and properties of the PSC-17 and BPI for comparison with previous studies.

The crux of this investigation, however, lay in the IRT analyses of item responses. First, IRT assumptions of unidimensionality, local independence, and specific trace line functions were assessed. Next, Samejima's (1969) graded response model, an IRT model for items with polytomous ordered response options, was fit using MULTILOG 7.03 software (Thissen, Chen, & Bock, 2003). Model fit was assessed using a $\chi 2/df$ ratios approach (Drasgow, Levine, Tsien, Williams, & Mead, 1995). Calibration of the 18 externalizing subscale items yielded three parameter estimates for each item: lower and upper difficulty thresholds (i.e., location), indicating the levels of externalizing behavior problems at which parents became more likely to select "sometimes" versus "never," and "often" versus "sometimes"; and a discrimination parameter (i.e., slope), reflecting the ability of the item to distinguish between children at similar levels of externalizing behavior problems. The sets of item parameter estimates facilitated comparison of the information and precision provided by each item along the continuum of externalizing behavior problems. Finally, each item was examined for DIF between groups differing by child sex, race, and SES, using two methods: the IRT-based likelihood ratio test (Thissen, 2001) and an ordinal logistic regression approach (Crane, van Belle, & Larson, 2004). Results of IRT analyses guided the identification of a set of items which (a) measured sub-clinical to clinical levels of externalizing behavior problems in preschool-aged

children most precisely, and (b) exhibited the least amount of DIF between groups of interest.

Results

Sociodemographic characteristics of the children (N = 900) were diverse: 47% were female, 50% were of minority race (predominantly African American), and 43% were of low SES. See Table 1 for detailed sample characteristics.

Table 1
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)
Child Household Composition		
Two-parent	512	(57)
Single parent	339	(38)
Caregiver other than parent	47	(5)
Child Health Insurance		
Public	634	(71)
Private	252	(28)
None	10	(1)
Socioeconomic Status (SES)		
Low	371	(43)
Medium	285	(33)
High	216	(25)
Caregiver 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)

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

Traditional psychometric analyses of the externalizing subscales of the PSC-17 and the BPI suggested similar performance to previous CTT-based investigations of distributional properties, internal consistency, and concurrent and known-groups validity (data not shown). All assumptions of IRT were met, and fit of the GRM was acceptable.

Research Question 1: Precision of measurement along the continuum of externalizing behavior problems. As hypothesized, IRT results revealed that all 18 externalizing subscale items 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. See Figure 1 for graphs of option characteristic curves (OCCs) for each item; OCCs depict the probabilities of particular response options along the continuum of externalizing behavior problems (theta), arbitrarily scaled as standard normal (i.e., from 3 standard deviations below to 3 standard deviations above the mean). Thirteen items were found to be most informative at sub-clinical to clinical levels of externalizing behavior problems (i.e., ≥ 1.5 standard deviations above the mean), while 5 items measured only low levels of externalizing behaviors. These 5 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")—were deemed undesirable for screening purposes and were eliminated from consideration for a brief screening instrument.

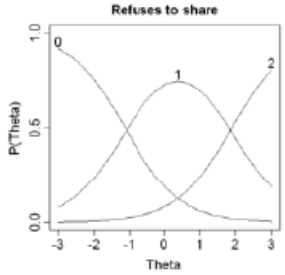
Research Question 2: Differential item functioning between groups differing by child sex, race, or SES. DIF-detection analyses identified eight items to which responses were influenced by child sex, race (controlling for SES), or SES (controlling for race), when level of externalizing behaviors was controlled. These items performed significantly differently between groups of interest, with a Bonferroni-corrected level of significance of $\alpha = .0027$. See Figures 2, 3, and 4 for depictions of the group differences in OCCs by child sex, race, and SES, respectively. Notably, within each category of DIF—by sex, race, and SES—the direction and magnitude of bias was inconsistent among items. At the scale level, various combinations of items exhibiting DIF resulted in either inflated or deflated estimates of levels of externalizing behavior problems within sociodemographic groups of interest. The eight items demonstrating DIF were eliminated from consideration for a brief screening instrument.

Integration of results: "Best" items for a brief screening instrument. 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.

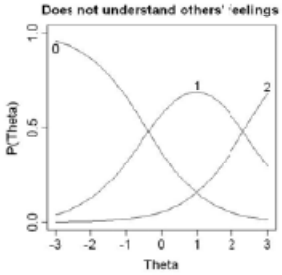
Eliminating the six items demonstrating DIF left seven items for consideration, each of which provided DIF-free measurement within the desired range of the latent

construct. Several of these seven items, however, demonstrated information peaks at identical levels of externalizing behavior problems. 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 5 depicts the relative information levels provided by the remaining seven items along the continuum of externalizing behavior problems. As illustrated in Figure 5, 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.

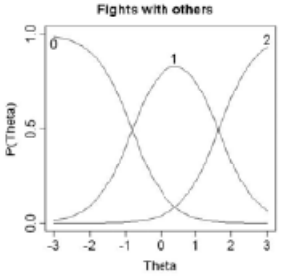
(a) PSC-17 Item 4



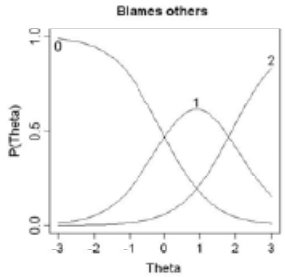
(b) PSC-17 Item 5



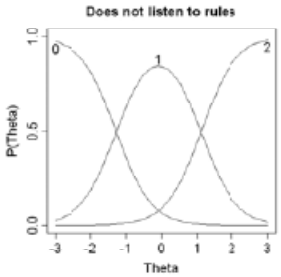
(c) PSC-17 Item 8



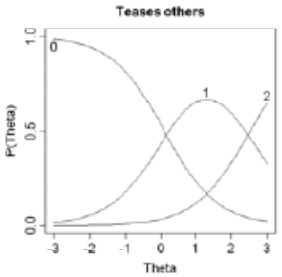
(d) PSC-17 Item 10



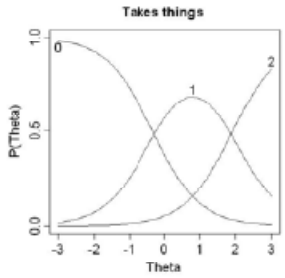
(e) PSC-17 Item 12



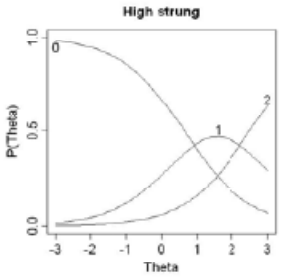
(f) PSC-17 Item 14



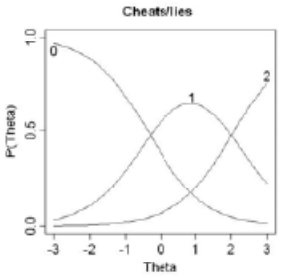
(g) PSC-17 Item 16



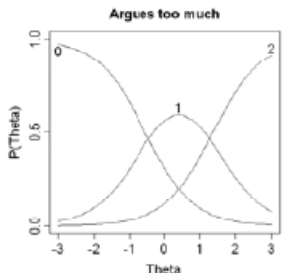
(h) BPI Item 3



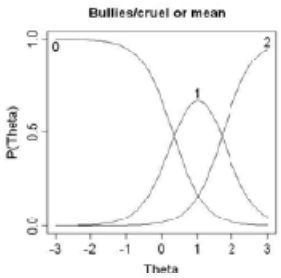
(i) BPI Item 4



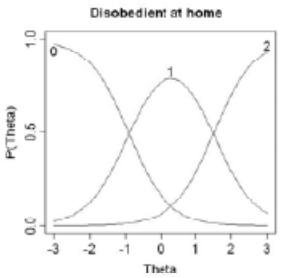
(j) BPI Item 6



(k) BPI Item 9



(l) BPI Item 10



(figure continues)

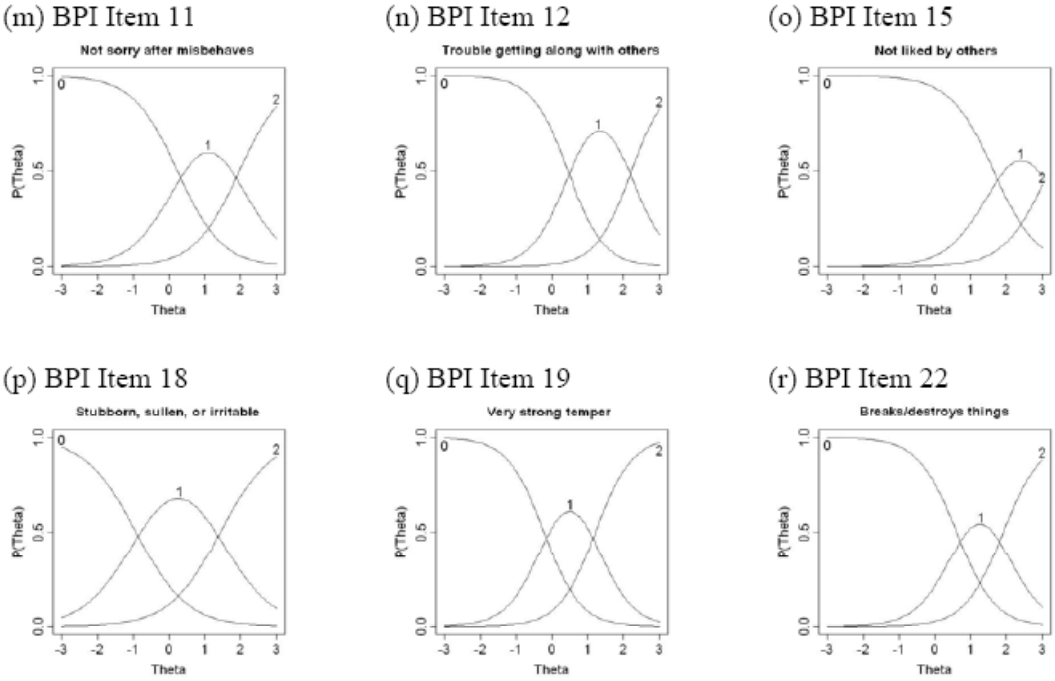


Figure 1. **Plots of graded response model option characteristic curves (OCCs) for all items in the combined externalizing subscale.**

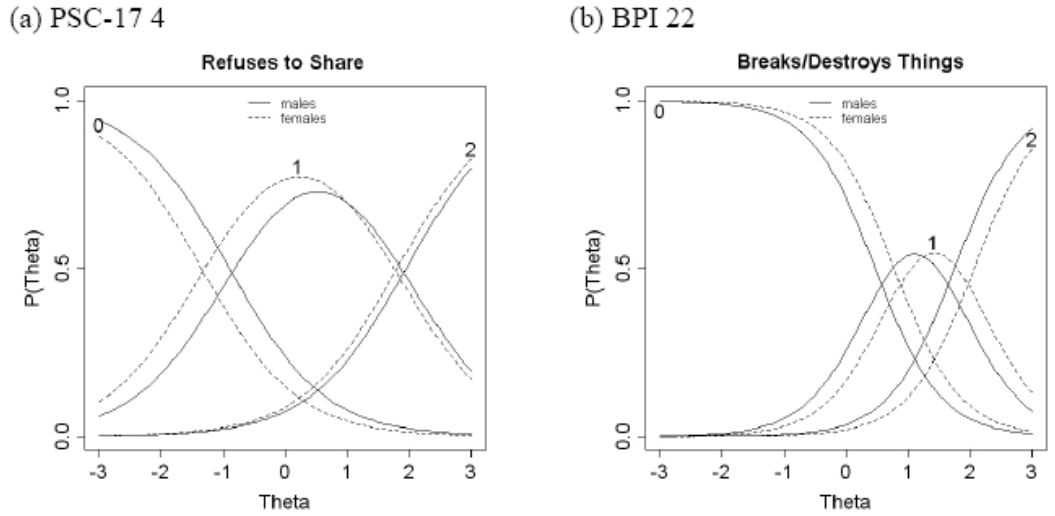


Figure 2. **Plots of graded response model option characteristic curves (OCCs) by group for items exhibiting differential item functioning by child sex.**

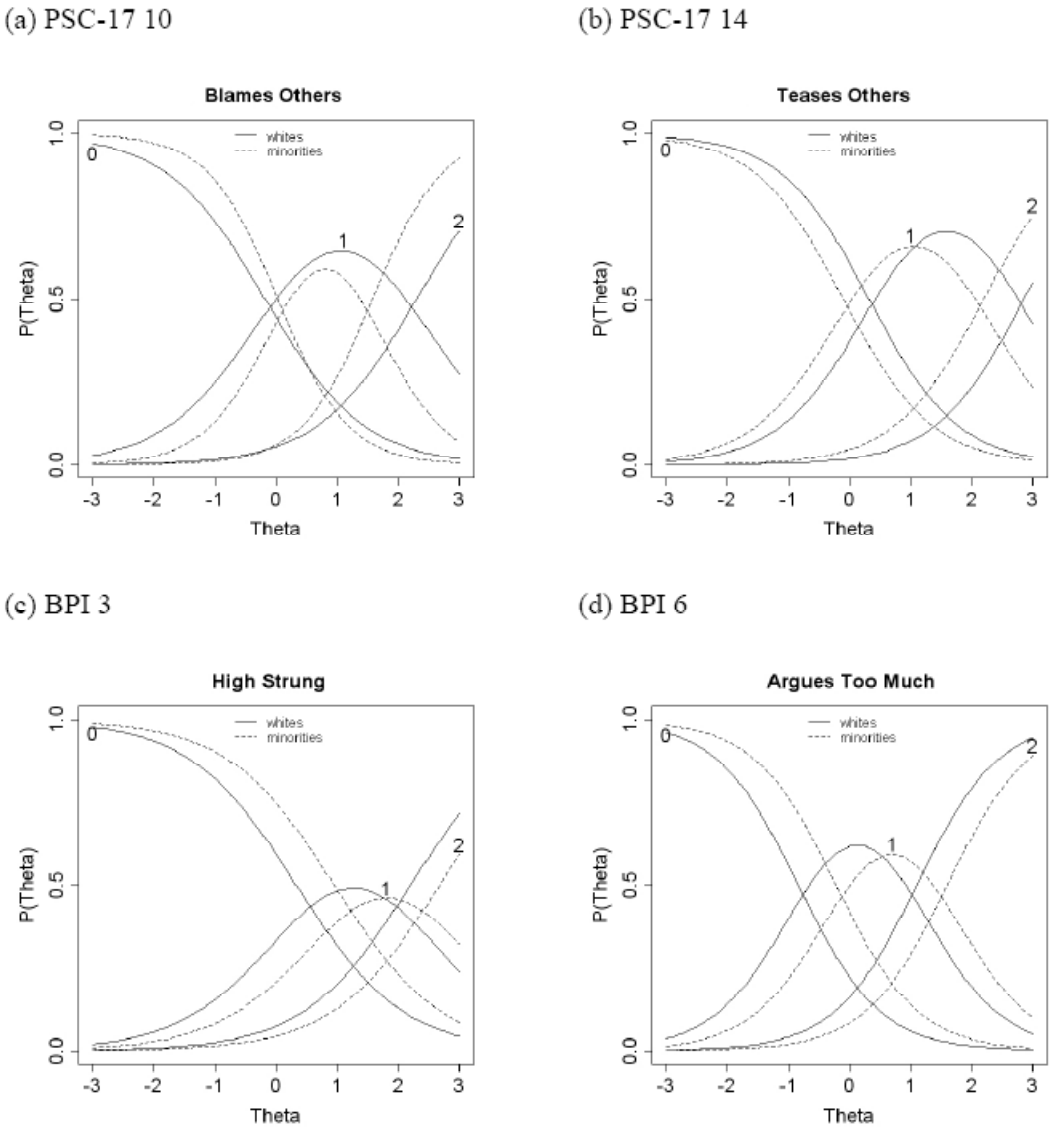
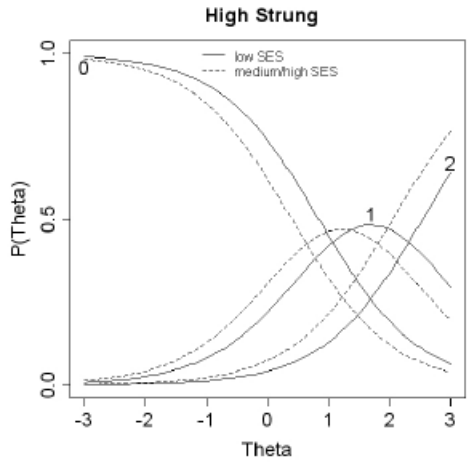
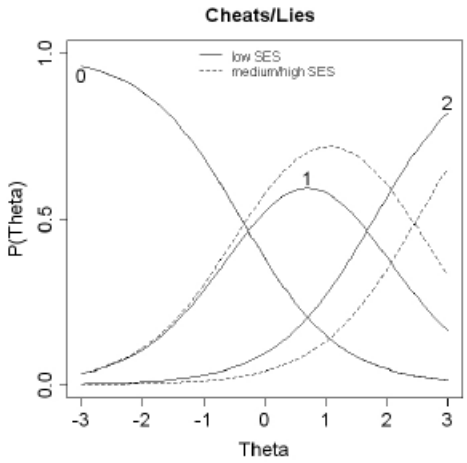


Figure 3. **Plots of graded response model option characteristic curves (OCCs) by group for items exhibiting differential item functioning by child race.**

(a) BPI 3



(b) BPI 4



(c) BPI 18

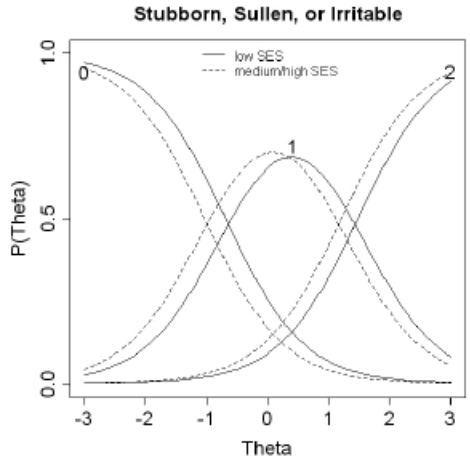


Figure 4. **Plots of graded response model option characteristic curves (OCCs) by group for items exhibiting differential item functioning by child socioeconomic status.**

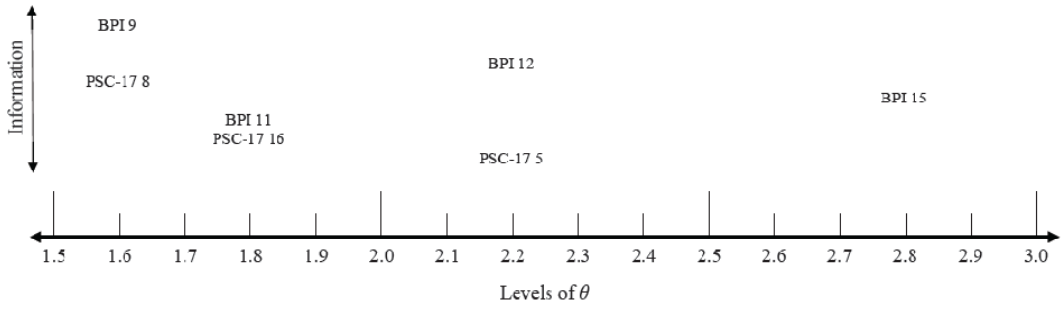


Figure 5. **Relative levels of DIF-free item information provided by items in the sub-clinical to clinical range of externalizing behavior problems.**

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.

Utility for Social Work Practice

Results highlight concerns regarding performance of the externalizing subscales of the PSC-17 and BPI with diverse preschool-aged children, due to (a) complexities of measuring behavior problems in a developmental context (Kagan, Snidman, McManis, Woodward, & Hardway, 2002), and (b) item bias (Osterlind, 1983). However, a set of four items found to be the most informative at sub-clinical to clinical levels of the latent construct, as well as the least biased between groups differing by sociodemographic characteristics, may be a promising tool for screening preschool-aged children in primary care. The content of these four items 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 (see Hann & Borek, 2001), and very young children who frequently exhibit these behaviors may benefit from further assessment and intervention by social workers or other mental health professionals.

Additional investigations of the proposed set of items are needed to assess its value in improving early identification of preschool-aged children with externalizing behavior problems in pediatric primary care. 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.

Implications. Several key implications stem from study results, including demonstrations of: (a) the limitations of traditional psychometric development and evaluation of screening instruments; (b) the added value of a modern psychometric approach in improving measurement of a serious social and public health problem; and (c) a mechanism for improvements in screening technologies for a range of psychosocial issues, potentially contributing to the reduction of health disparities exacerbated by bias in measurement instruments. The study's methodology could be replicated and applied in evaluating and improving measures of a wide range of constructs vital to social work practice and research. This translational research also demonstrates the capacity for broad dissemination of results: Findings are applicable not only to pediatric primary care, but to other settings as well, including preschools, early childcare, mental health, and the child welfare system.

Several study implications are particularly pertinent to social work education, practice, and research. Heightened attention to measurement theory in social work education could prepare social work practitioners and policy-makers to be cognizant of limitations of CTT-developed instruments commonly used for outcome evaluations at individual, program, and systems levels. This educational focus would also prepare social work researchers to further contribute to the evaluation and development of measurement tools crucial to social work practice and research via advanced measurement theory and applications (e.g., DeRoos & Allen-Meares, 1992; Nugent, 2003, 2005, 2006; Nugent & Hankins, 1992). Investigations of DIF are particularly relevant to efforts to reduce health disparities and promote social justice, key components of the social work mission (NASW, 1996, revised 2008). The social work profession calls for cultural sensitivity and competence; thus, social workers in education, research, practice, and policy settings should ensure that the instruments used within their realms of influence meet those standards.

Study 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 studies. More importantly, IRT methods 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. 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 particular 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. Other limitations to recognize include somewhat coarse categorizations of race and SES; an inability to control for caregiver race in analyses, due to small cell sizes; sole reliance on caregiver-reported data; and lack of definitive guidelines regarding sample size requirements for IRT analyses.

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. This study applied IRT to overcome limitations associated with traditional methods of scale development and evaluation, providing novel information regarding the psychometric performance of items measuring externalizing behavior problems in preschool-aged children. Results revealed several items which measured only low levels of the latent construct in very young children, as well as DIF between groups differing by child sex, race, and SES. However, a set of four informative and unbiased items 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 vital 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

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.

American Academy of Pediatrics. (2000). Fellows Survey. Elk Grove Village, IL: American Academy of Pediatrics.

Blount, A. (2003). Integrated primary care: Organizing the evidence. Families, Systems, & Health, 21, 121-133.

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.

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.

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.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. Applied Psychological Measurement, 19, 143-166.

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.

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., & 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.

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.

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.

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., & 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., Snidman, N., McManis, M., Woodward, S., & Hardway, C. (2002). One measure, one meaning: Multiple measures, clearer meaning. Development and Psychopathology, 14, 463-475.

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.

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.

NASW. (1996). Code of ethics of the National Association of Social Workers. Washington, DC: NASW Press.

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.

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.

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.

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.

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.

Reiger, D. A., Goldberg, I. D., & Taube, C. (1978). The de facto U.S. mental health service system. Archives of General Psychiatry, 35, 685-693.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. Journal of Educational Measurement, 27, 133-144.

Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 17.

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.

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.

Strosahl, K., Robinson, P., Heinrich, R., Dea, R., Del-Toro, I., Kirsh, J., et al. (1994). New dimensions in behavioral health/primary care integration. HMO Practice, 8, 176-179.

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.

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.

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.

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.