



---

## Graduate Theses, Dissertations, and Problem Reports

---

2012

# Receiver Operating Characteristics of the CAP Lie Scale and Correlates of Impression Management in Parenting Capacity Evaluations

Ryan J. Anderson  
West Virginia University

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>

---

## Recommended Citation

Anderson, Ryan J., "Receiver Operating Characteristics of the CAP Lie Scale and Correlates of Impression Management in Parenting Capacity Evaluations" (2012). *Graduate Theses, Dissertations, and Problem Reports*. 185.

<https://researchrepository.wvu.edu/etd/185>

This Dissertation is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Dissertation in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Dissertation has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact [researchrepository@mail.wvu.edu](mailto:researchrepository@mail.wvu.edu).

Receiver Operating Characteristics of the CAP Lie Scale and  
Correlates of Impression Management in Parenting Capacity Evaluations

Ryan J. Anderson, M.S.

Dissertation submitted to  
the Eberly College of Arts and Sciences at West Virginia University  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in  
Psychology

William J. Fremouw, Ph.D., Chair  
Michael Crabtree, Ph.D.  
Cheryl B. McNeil, Ph.D.  
Aaron Metzger, Ph.D.  
Tracy L. Morris, Ph.D.

Department of Psychology

Morgantown, West Virginia  
2012

Keywords: child abuse potential inventory, personality assessment inventory, child maltreatment,  
parenting evaluation, psychometrics

## ABSTRACT

### Receiver Operating Characteristics of the CAP Lie Scale and Correlates of Impression Management in Parenting Capacity Evaluations

Ryan J. Anderson

The purpose of the present study was to investigate the validity and reliability of the Child Abuse Potential L scale with a heterogeneous sample of caregivers referred for parenting capacity evaluations. One aim of the study was to assess the measurement properties of the L scale. A second aim was to evaluate the discriminative validity of the L scale by way of its receiver operating characteristics. A third aim of the study was to examine potential correlates of desirable responding on the L scale. The findings from this study provide new information about the psychometric properties of the CAP L scale and its application in clinical and forensic settings. Consistent with past investigations, caregivers produced a high rate (74.4 %) of invalid CAP profiles by way of elevated L scale scores. The L scale showed little variation across caregivers from families with different maltreatment histories. Item analyses and estimates of internal consistency showed homogeneity of the L scale, though several problem items were identified. Deletion of these items, however, produced only marginal improvements in internal consistency. The 14-item revised scale that resulted from the item deletions showed tradeoffs in sensitivity and specificity compared to the original 18-item scale. Classificatory accuracy of the 18-item scale (with emphasis on sensitivity to detect fake-good responding) was best using a cutoff score that was one to two points higher than recommendations given in the CAP manual (Milner, 1986). Last, the L scale showed inverse associations with stress and aggression. These findings suggest that caregivers perceive the context of evaluation to be coercive, pointing up the importance of procedures and pacing that increase rapport. Also, caregivers who report low levels of anger and stress produce higher L scale scores. Furthermore, findings highlight tradeoffs in L scale sensitivity and specificity that evaluators can select as a function of referral question or other relevant considerations. In sum, findings add to the scientific merit of the CAP in relation to *Daubert* criteria for testimonial admissibility.

## ACKNOWLEDGEMENTS

I wish to thank the many people who contributed to the completion of this project. Dr. William Fremouw, my committee chair and advisor, consistently provided direction and focus. His penchant for productivity and efficiency has made me a more effective clinician and researcher. I am pleased to call him my “academic parent.” The unique opportunity to evaluate and study families with maltreatment histories was made possible by Dr. Michael Crabtree. His supervision and advocacy were constant and helped me to navigate the many obstacles posed to this project and its shorter-lived antecedent. I am grateful to Drs. Aaron Metzger, Cheryl McNeil, and Tracy Morris for their insights and suggestions. Their contribution to my graduate studies and professional development extends well beyond the scope of this humble project. Dr. Metzger maintained my interest in developmental science and strengthened my understanding of parent-adolescent relationships. Dr. McNeil set in motion my interest in early intervention and provided me with skills that have proved invaluable to me as a psychologist and as a parent. Dr. Morris shaped my critical thinking skills, reinforced my commitment to behavior analysis, and brought my skepticism under stimulus control. Finally, I reserve my highest gratitude for Bethany, my wife. Her resolve, commitment, and love made all of these learning opportunities possible.

## TABLE OF CONTENTS

|   |     |
|---|-----|
| Title page.....   | i   |
| Abstract.....   | ii  |
| Acknowledgements.....   | iii |
| Introduction.....   | 1   |
| The Child Abuse Potential Inventory.....                                      | 5   |
| CAP Clinical Scales.....  | 6   |
| CAP Validity Scales.....  | 10  |
| L Scale Validity and Desirable Responding.....                                | 11  |
| The Personality Assessment Inventory.....                                     | 15  |
| Development of the PAI: Test Construction, Validity, and Reliability.....     | 16  |
| Positive Impression Management.....   | 18  |
| Risk Factors for Maltreatment and Associations with Desirable Responding..... | 22  |
| The Present Study.....  | 29  |
| Research Question 1.....  | 29  |
| Hypothesis 1.....   | 29  |
| Research Question 2.....  | 29  |
| Hypothesis 2.....   | 29  |
| Research Question 3.....  | 30  |
| Hypothesis 3.....   | 30  |
| Method.....   | 30  |

|   |    |
|---|----|
| Participants.....   | 30 |
| Procedure .....   | 31 |
| Measures .....  | 32 |
| Demographic Form .....  | 32 |
| Child Abuse Potential Inventory.....                              | 32 |
| Personality Assessment Inventory .....                            | 33 |
| Overview of Data Analyses.....                                    | 35 |
| Descriptive Analyses.....   | 35 |
| Preliminary Analyses .....  | 36 |
| Primary Analyses.....   | 37 |
| Hypothesis 1.....   | 38 |
| Hypothesis 2.....   | 39 |
| Hypothesis 3 .....  | 41 |
| Results.....  | 41 |
| Characteristics of the Analytical Sample.....                     | 41 |
| Data Screening Procedures .....                                   | 43 |
| Demographic Characteristics .....                                 | 44 |
| Descriptive and Inferential Statistics for Maltreatment Type..... | 45 |
| Descriptive Statistics for the Study Variables .....              | 46 |
| Hypothesis 1.....   | 47 |
| Bivariate Correlations for the CAP L Scale .....                  | 47 |
| Homogeneity of CAP L Scale Content .....                          | 47 |
| Full Sample CAP L Scale Revisions .....                           | 48 |

|   |     |
|---|-----|
| Subsample CAP L Scale Revisions .....                       | 48  |
| Hypothesis 2 .....  | 49  |
| Hypothesis 3 .....  | 51  |
| Regression Diagnostics .....                                | 52  |
| Model Selection Procedures .....                            | 54  |
| Final Model Interpretation .....                            | 54  |
| Discussion .....  | 55  |
| Sample Characteristics .....                                | 56  |
| Homogeneity of CAP L Scale Content .....                    | 59  |
| Receiver Operating Characteristics of the CAP L Scale ..... | 62  |
| Correlates of Desirable Responding .....                    | 64  |
| Implications for Parenting Capacity Evaluations .....       | 70  |
| Limitations and Future Research .....                       | 72  |
| References .....  | 78  |
| Footnotes .....   | 89  |
| Tables .....  | 90  |
| Figures .....   | 100 |
| Appendix .....  | 102 |

Receiver Operating Characteristics of the CAP Lie Scale and  
Correlates of Impression Management in Parenting Capacity Evaluations

Child maltreatment is a regrettably common problem in the U.S. In 2009, national and state statistics compiled by child protective services (CPS) agencies through the National Child Abuse and Neglect Data System (NCANDS) recorded 3.6 million counts of child maltreatment (U.S. Department of Health and Human Services). Investigations conducted by CPS and cooperating agencies revealed that nearly one-quarter (763,000) of the children named in the reports were in fact victims of child maltreatment.<sup>1</sup>

When investigation reveals that child maltreatment is substantiated (i.e., did occur) or indicated (i.e., is suspected), CPS agencies provide services aimed at keeping children safe and preventing future occurrence of abuse and neglect. Postinvestigation services to children and their families may include drug and alcohol counseling, mental health counseling, and even foster care placement. NCANDS showed that in 2009, just under one million (980,712) children and their families received such services. Clearly, child maltreatment is prevalent and, in many instances, leads to CPS involvement with referred families.

In addition to the immediate services necessary to keep children safe, it is estimated that victims of various forms of child maltreatment (e.g., physical abuse, sexual abuse, neglect) are at increased risk for longer-term health impairments. For example, the World Health Organization (Krug, Dahlberg, Mercy, Zwi, & Lozano, 2002) documented myriad physical, reproductive, psychological, and chronic health consequences associated with child maltreatment, ranging from traumatic brain injury to post-traumatic stress to infertility. In terms of financial detriment, “direct” expenses (e.g., hospitalization, mental healthcare, child welfare services, law

enforcement) plus “indirect” expenses (e.g., special education, delinquency, lost productivity to society) is estimated to cost the U.S. \$104 billion annually (Wang & Holton, 2007).

Given the individual and societal costs of child maltreatment, efficient and effective use of public health resources is imperative. Psychologists and other mental health professionals are in a position to help CPS agencies and the courts allocate scarce resources by way of conducting forensic evaluations of parenting competency. In general, forensic psychologists evaluate individuals involved with the legal system to inform the courts on issues where legal and psychological constructs intersect. More specific to parenting competencies: “Because legal definitions of parental fitness, abuse, and neglect are vague . . . , courts have turned to mental health professionals as expert witnesses to inform them on this topic” (Benjet, Azar, & Kuersten-Hogan, 2003, p. 239). For example, the findings of forensic evaluations with parents who have maltreated their children may be used to determine the type and intensity of postinvestigation services a family may require. Also, forensic evaluators may be called on to assess the effectiveness of services that parents and children have received or to assess the suitability of dispositions such as parent-child reunification or termination of parental rights.

From the standpoint of the forensic evaluator conducting parenting capacity evaluations, the scope of assessment is wide and requires narrowing. For example, perspectives from developmental science emphasize multiple, interactive contextual factors that broaden the range of assessment variables and levels of analysis; for example, Belsky’s (1993) ecological-developmental model or developmental psychopathology (Cicchetti & Toth, 1995). Forensic evaluators might well ask pragmatic questions such as: What are considered to be core parenting skills and how should they be measured?

Researchers interested in forensic assessment of parenting capacity have defined practice models that provide tentative answers to these practical questions (Barnum, 1997; Budd, 2001; Budd, Connell, & Clark, 2011). Prototypical models map parent learning history, cognitive functioning, and personality as well as childrearing knowledge, attitudes, and behaviors onto the legal questions under consideration (e.g., Azar, Lauretti, & Loding, 1998; Barnum, 1997; Budd, 2001). Though researchers have advanced functional-contextual models that emphasize ideographic assessment of parenting-related domains (Azar et al., 1998), use of nomothetic assessment methods such as personality inventories and parenting measures are commonplace (Budd, Poindexter, Felix, & Naik-Polan, 2001) and provide relevant sources of convergent (or divergent) data to inform legal decisions (Budd et al., 2011; Budd & Holdsworth, 1996; Grisso, 2003). The American Psychological Association (APA) Committee on Professional Practice Standards also calls for multiple methods of data gathering in parenting capacity evaluations (APA, 1999).

From the standpoint of the caregivers participating in parenting capacity evaluations, much is at stake including access to children and required remediation of deficit and problem areas. Budd (2001) has conjectured that compulsory evaluation produces a “coercive context” (p. 3) that is not conducive to obtaining reliable and valid assessment data. Furthermore, Budd and her colleagues (2011) have outlined several reasons why assessment results may be distorted, such as caregivers’ (a) selective attention to the consequences of their problem behavior, (b) interpersonal problems and untreated mental health issues, (c) fear and mistrust of CPS agencies and the systems with which they interface, as well as (d) culturally prescribed ways of interacting with professionals (p. 156). Thus, the conditions for conducting evaluations of parenting capacity are suboptimal with respect to obtaining valid and reliable assessment data.

Recent investigations support the contention that parents may distort—deliberately or unintentionally—their responding across many different methods of data collection (Bennett, Sullivan, & Lewis, 2006; Budd et al., 2011), including psychological testing (Carr, Moretti, & Cue, 2005; Ondersma, Chaffin, Mullins, & LeBreton, 2005; Stredney, Archer, & Mason, 2006). For example, Carr et al. (2005) conducted an investigation of validity problems associated with personality and parenting measures administered to a sample of respondents undergoing parenting capacity evaluation. He and his colleagues reported that nearly 20% of Personality Assessment Inventory (PAI) profiles were invalid using the conservative Positive Impression (PIM) cutoff score of 66*T*. Minnesota Multiphasic Personality Inventory–2 (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 2001) profiles were invalid at a rate of 60% using a cutoff score of 65*T* on the L scale. (The L scale or “lie scale” is interpreted as an indicator of deliberate attempts to present oneself in an unrealistically favorable manner.) Furthermore, nearly half (49%) of the Child Abuse Potential Inventory (CAP) Abuse scale scores were invalidated by way of elevations on the Faking-good index.

The CAP is a self-report measure of parenting attitudes and behaviors associated with risk for physical abuse of a child (Milner, 1986). It is the only measure of its kind (Budd, 2001; Budd et al., 2011). Furthermore, it is one of only two parenting measures determined to meet *Daubert* criteria for testimonial admissibility in legal proceedings (Medoff, 2003; Yanez & Fremouw, 2004). However, when the Faking-good index is elevated, the Abuse scale score cannot be interpreted and a valuable source of information about caregiver risk for physical abuse of a child is lost to the evaluator.

To summarize, child maltreatment is a serious and prevalent problem associated with substantial personal and societal costs. In many instances, CPS agencies and courts request that

psychologists conduct parenting capacity evaluations. Such evaluations require multiple methods of data collection, and parenting tests are a common method of data collection. The CAP, in particular, is a parenting measure that has undergone multiple validation studies with different samples, has known rates of measurement error, and has yielded many publications in peer-reviewed scientific journals. However, recent research indicates high rates of invalid CAP profiles, calling into question the utility of its validity scales, specifically, the validity of the Lie (L) scale as used to calculate the Faking-good index. Thus, the purpose of the present investigation was to conduct an item analysis of the CAP L scale and to test the receiver operating characteristics of the CAP L scale in relation to the PAI PIM scale. Also, analyses were conducted to explore clinically relevant correlates of desirable responding as measured with the CAP L scale.

In subsequent sections, the psychometric properties of the CAP are introduced followed by discussion of the psychometric properties of PAI validity scales and clinical scales relevant to the detection of desirable responding in persons undergoing parenting capacity evaluations. Finally, specific aims, research questions, and hypotheses are presented.

### **The Child Abuse Potential Inventory**

The CAP (Milner, 1986) is a 160-item self-report measure of parenting attitudes and practices. It is intended as a screening tool for the detection of physical child abuse in caregivers investigated by CPS agencies. The CAP is a well-validated measure that aids in the identification of caregivers who report parenting beliefs and behaviors similar to those observed in samples of caregivers with known histories of physical child maltreatment (Milner, 1986). More recent investigations of child maltreatment show substantial overlap among types of child maltreatment (e.g., Barnett, Manly, & Cicchetti, 1993; Sedlak & Broadhurst, 1996), and

researchers have shown that the CAP is capable of correct classifications using samples of abusers heterogeneous for maltreatment type (Ondersma et al., 2005). Thus, the CAP is an instrument with psychometric properties relevant to parenting capacity evaluations conducted in CPS settings. In the next section, the validity of the CAP is reviewed followed by a summary of its reliability.

### **CAP Clinical Scales**

The CAP is comprised of six clinical scales (Distress, Rigidity, Unhappiness, Problems with Child and Self, Problems with Family, Problems from Others). These six scales are subsumed by a superordinate Abuse scale. The 77-item Abuse scale has been the focus of extensive validation (Milner, 1994; Walker & Davies, 2010). Case-control research designs have been used to test the Abuse scale's ability to classify caregivers as (a) those with known histories of child maltreatment versus (b) those with no known histories of child maltreatment (Milner & Wimberly, 1980). For example, Milner and his colleagues (1985) reported sensitivity (i.e., ability to correctly classify known abusers) in the range of 95.5% to 100% and specificity (i.e., ability to correctly classify known nonabusers) in the range of 88.2% to 96.3% with classification rates invariant across gender and cultural background. More recent investigations, however, have shown that higher rates of false positive classifications are obtained in non-U.S. samples (Pečnik & Ajduković, 1995; Diarme, Tsiantis, & Tsitoura, 1997; Haz & Ramirez, 1998). For example, nearly 22% of nonabusers were falsely classified as abusers in Diarme and colleagues' (1997) Greek sample compared to the false positive rates of 3.7% and 11.8% reported for Milner and colleagues' (1985) U.S. samples. This finding suggests that clinicians should give conservative interpretations when applying the CAP to samples different from the CAP normative sample.

With regard to other aspects of validity, factor analyses of the CAP show evidence for six (Milner, 1986) or seven (Milner & Wimberly, 1980; Ondersma et al., 2005) empirically derived factors or scales. Forensic and clinical interpretation of the CAP is based on the six-factor solution. However, Milner (1986) has emphasized that the six Abuse factor scales (e.g., Distress, Unhappiness, Rigidity) should not be used for classification purposes. Rather, the factor scales should be used only for descriptive purpose or to generate clinical hypotheses for further evaluation. For example, a forensic evaluator might conclude that an elevated Distress scale score indicates poor frustration tolerance, sadness, persistent worry, and the perception of being socially isolated. Each area of concern would then warrant further contextual and functional assessment of the associated behaviors and their potential to impact parenting capacity. The evaluator would refrain from concluding that an elevated Distress scale score indicates parenting beliefs and behaviors similar to known samples of child abusers given that only the full Abuse scale (of which Distress is but one of six factors) has shown acceptable classification rates.

In addition to discriminative validity and content validity, the CAP Abuse scale has evidenced discriminant validity by way of low false positive classifications among medical-surgical patients (with no known histories of child maltreatment) experiencing stressful physical health problems (Milner, 1991). Likewise, Talbott (1985) showed that Abuse scale scores were not affected by stress that was experienced as distinct from parent-child interactions. Furthermore, the CAP Abuse scale has shown concurrent validity in that mothers with high scores on the CAP Abuse scale attributed greater hostile intent to child behavior and reported the use of more power-assertive parenting strategies (i.e., verbal and physical force) compared to mothers at low risk for child maltreatment (De Paúl, Asla, Pérez-Albéniz, & Torres-Gómez de Cádiz, 2006; Montes, De Paúl, & Milner, 2001).

Compared to other types of validity evidence, data supporting the predictive validity of the CAP is not as strong. For example, Milner, Gold, Ayoub, and Jacewitz (1984) followed a sample of 190 parents at risk for parenting problems for, on average, six months. The research team administered the CAP at baseline. At the conclusion of this longitudinal study, 42 parents were confirmed as having committed acts of physical abuse, neglect, or nonorganic failure-to-thrive. All 42 parents had Abuse scale scores above the CAP cutoff. However, 61 parents who committed no reported acts of abuse also had Abuse scale scores above the cutoff. Thus, the predictive validity of the CAP is of limited forensic utility given a false positive rate of 89.3%.

Comparable problems with predictive validity have been described by Mark Chaffin and his colleagues (Chaffin & Valle, 2003; Ondersma et al., 2005). For example, Chaffin and Valle documented temporal decreases in CAP scores that did not correspond to decreases in the occurrence of future instances of child maltreatment. A plausible hypothesis may be that the CAP is not sensitive to changes in observable and meaningful risk factors impinging on caregiver-child dyads over the course of the dyads' involvement with CPS agencies. This is consistent with Chaffin and Valle who proposed that the CAP may be most sensitive to "superficial" (p. 476) markers of risk. This conclusion does not undermine use of the CAP as a baseline screening instrument. It does, however, call into question use of the CAP as an instrument capable of detecting the types of behavioral change associated with reduction in the risk for future child maltreatment.

With regard to reliability, Milner (1986) has reported data on the internal consistency and temporal stability of the CAP and its constituent scales. Internal consistency estimates by way of Kuder-Richardson correlations (KR-20) for the Abuse scale across physical abuse ( $N = 152$ ) and neglect ( $N = 218$ ) samples was .95 and .93, respectively. Furthermore, when the internal

consistency of the Abuse scale was assessed as a function of gender, age, education, and cultural background; estimates were entirely above .87. The CAP Distress scale showed comparable internal consistency (.87 to .96). However, internal consistency estimates for Rigidity, Problems from Others, Unhappiness, Problems with Child and Self, and Problems with Family were lower with values in the range of .80 to .30 and greater variability across subgroups (e.g., gender, age). To some extent, discrepancy in estimates of internal consistency across scales is likely to be a function of scale length given the construction of the KR-20 formula (Nunnally & Bernstein, 1994). For example, the Problems with Child and Self scale subsumes six items, whereas the Abuse scale subsumes 77 items. In sum, the internal consistency of the Abuse scale is consistent with its use as a screening tool, whereas the internal consistency of the other scales is consistent with their use for descriptive purposes only.

The temporal stability of the CAP was assessed for control subjects across one-day, one-week, one-month, and three-month intervals. The Abuse scale showed a systematic, albeit slight, decrease in reliability as a function of time with the greatest change occurring at the three-month interval,  $r = .75$ . The Rigidity scale, as might be expected, showed the greatest temporal stability, and the three problems scales showed the least temporal stability. It is worth noting that at the three-month interval, the Abuse scale showed almost no change in the mean score, 85.24 versus 85.91, but showed change in the standard deviation, 74.78 versus 63.59. It is likely that this change is due to the temporal stability of the factors that comprise the Abuse scale. Some factors (e.g., Rigidity) appear less sensitive to temporal effects, whereas other factors (e.g., Problems with Family) appear more sensitive to temporal effects and may be affected by cases that are outliers for recent problems such as family conflict or others stressors that may fluctuate over relatively brief time periods. Again, this points to the value of the CAP as a screening tool,

but indicates that its factor scales should be interpreted with caution and for descriptive purposes only. In fact, changes in factor scales across time may be of greater interest to evaluators (and researchers) than the presence of high versus low scores on factor scales at any single point in time. Investigations in this area could provide valuable information about assessment of static versus dynamic risk factors for child maltreatment.

### **CAP Validity Scales**

In addition to the six clinical scales, three validity scales have been developed for the CAP: Inconsistency (IC), Random Response (RR), and Lie (L). IC is comprised of 20 similarly worded item pairs. Three item pairs (six items total) are shared with the RR scale and the L scale. The remaining 17 IC pairs overlap with Abuse scale items. The RR and L scales are comprised of 18 items each, and no items from either scale overlap with Abuse scale items. Thus, the CAP is comprised of 77 abuse-related items and 36 validity-specific items. The remaining 47 items of the 160 total CAP items are exploratory.

The RR scale was developed to detect haphazard responding to CAP items; that is, to detect the practice of responding to CAP items without attending to item content (Milner, 1986). In RR scale validation studies, items were chosen on the basis of having low endorsement rates and low correlations with the Abuse scale. The resultant 18-item RR scale yielded misclassification rates below 5% (Milner & Robertson, 1985).

The IC scale was developed to complement the RR scale by way of a Random-response index (Milner, 1986). Development of the IC scale and Random-response index was undertaken subsequent to the findings that (a) the RR scale was sensitive to fake-bad responding (Robertson & Milner, 1985) and (b) random responding was a low base-rate phenomenon, occurring in less than 5% of respondents in the RR scale validation samples (Milner & Robertson, 1985). The IC

scale is thus comprised of item pairs that received a high percentage of similar responses in the development sample (Milner, 1986). Scores from the IC scale and the RR scale are used in tandem to make decisions about random responding such that when both scales are elevated, random responding is said to be present. The Faking-bad index also makes use of the IC scale and the RR scale in tandem. This Index capitalizes on the finding that the RR scale is sensitive to fake-bad responding. Thus, fake-bad responding is indicated when the RR scale is elevated and the IC scale is not elevated.

The L scale was developed “in an effort to eliminate individuals who attempt to distort their responses in a socially desirable manner” (Milner, 1986, p. 30). Five validation studies are reported in Milner (1982) and in the second edition of the *Child Abuse Potential Inventory Manual* (Milner, 1986). The general strategy used in the development of the L scale was fourfold. First, items with content describing socially desirable, but rarely attainable, personal and interpersonal qualities (e.g., “I always do what is right,” “I never listen to gossip”) were generated. Second, items that yielded an 85%–15% split on the “agree-disagree” response format were identified where the 15% portion corresponded to endorsement of the socially desirable quality. Third, L scale items with low ( $-.08$  to  $.08$ ), statistically nonsignificant ( $p > .05$ ) correlations with the Abuse scale were selected. Fourth, the Faking-good index was developed and makes use of the L scale and the RR scale in tandem to ensure that respondents’ socially desirable presentations were not due to the effects of random responding. The effect of random responding is an important consideration given that the L scale subsumes nine “agree” items and nine “disagree” items.

**L scale validity and desirable responding.** Of the six CAP validity scales and indexes, the L scale is central to the present investigation. Forensic evaluators report perennial frustration

at the high rate of invalid CAP profiles due to elevations on the L scale and Faking-good index. Investigators have confirmed this practice-level frustration by showing that rates of invalid profiles may range from approximately 30% (Ondersma et al., 2005) to nearly 50% (Carr et al., 2005). Invalid protocols provide little substantive information about the respondent other than the fact that he or she endorsed L scale items that were uncommonly endorsed in the normative sample. This manner of responding may or may not have affected the Abuse scale score. Thus, it is worthwhile to consider a program of research to investigate the measurement properties and classificatory ability of the L scale. For instance, by evaluating item discrimination, content homogeneity, and various L scale cut scores; it may be possible to optimize the classificatory function of the L scale. There are several reasons to think that the classificatory function of the L scale can be improved. These reasons derive from past research on the measurement and structural properties of the scale.

With regard to item discrimination and content homogeneity, Ondersma and colleagues (2005) reported use of a shortened six-item L scale constructed for a brief 33-item version of the CAP. The six-item L scale produced an overall classification rate comparable to the 18-item full version of the L scale. Thus, 12 items were found to yield insignificant incremental validity and were dropped. It is probable that dropped items measured content unrelated to the domain of interest. Nunnally and Bernstein (1994) suggest that heterogeneity of item content is due to random measurement error or item content that is too diverse. The presence and extent of these problems is evaluated by way of item analysis (i.e., item discrimination) and estimation of internal consistency.

Additional indicators of L scale heterogeneity such as low correlations with other measures of desirable responding have been reported. For example, Robertson and Milner

(1983, 1985) reported that the CAP L scale shared less than 10% variance with the Marlowe-Crowne Social Desirability Scale (M-CSDS) and approximately 25% variance with the MMPI L scale, leading them to conclude that “the majority of the variance remains to be explained” (p. 428). Furthermore, Robertson and Milner (1985) administered the CAP L scale and the M-CSDS under three different instructional sets: “be honest,” “respond in a socially desirable manner,” and “respond in a socially undesirable manner.” Results showed that the L scale was elevated more often than expected compared to the M-CSDS under the “be honest” condition. Also, the L scale was sensitive to the “socially undesirable” instruction set, indicating that it was sensitive to fake-bad response sets as well as fake-good response sets. Collectively, these findings also indicate heterogeneity of L scale content.

The L scale and RR scale when used in tandem to form the Faking-Good Index provide a rational or inferential means of removing the effects of fake-bad responding. The logic of this procedure, paraphrased from Milner (1986), is as follows: The L scale measures fake-good responding, fake-bad responding, and random responding. The RR scale measures fake-bad responding and random responding. Therefore, an elevated L scale and depressed RR scale indicate a fake good response set. Yet, there is no empirical evidence that the variance in fake-bad responding measured by the L scale and RR scale is shared. Thus, a heterogeneous L scale—one sensitive to both “fake good” and “fake bad” responding—is problematic in that CAP profiles could be determined invalid on the basis of responding (e.g., fake-bad) that would not be expected to produce artificially low Abuse scale scores. Under such conditions, the L scale would be expected to produce excessive false positive classifications for fake-good responding.

How, then, should an elevated L scale be interpreted? This is a broad, yet legitimate question in response to concerns about the psychometric properties of the CAP L scale.

Targeted, preliminary investigation of its measurement properties (i.e., item discrimination, internal consistency) and discriminative validity will be the focus of the present study. Prior to proceeding with review of the criterion measures (for tests of discriminative and convergent validity) and prior to discussing study procedures, basic definitions from research on desirable responding are provided. This overview is intentionally brief as the present study was not designed to provide a thoroughgoing analysis of the construct validity of the L scale in relation to self-presentation and its variants.

D. L. Paulhus has established a program of research on desirable responding. Several of his definitions are relevant to the present study. For instance, Paulhus (2003) defines *self-presentation* as “the generic term for the tendency to describe oneself in a self-serving fashion” (p. 858). Self-presentation, then, subsumes specific types of desirable responding such as *self-deception* and *impression management*, where the former is defined as “narcissistic” and “overconfident” (Paulhus, 2002, p. 64), and the latter as “deliberate exaggeration, faking, and lying” to accommodate situational demands (Paulhus & Vazire, 2007, p. 229).

Nunnally and Bernstein (1994) have been critical of Paulhus’ (1984) theorizing on self-enhancement on the basis of his distinguishing between unconscious (e.g., self-deception) versus conscious (e.g., impression management) processes. They point up psychologists’ persistent failures in differentiating levels of consciousness. Furthermore, Nunnally and Bernstein have been skeptical of the trait definitions implicit in self-deception (e.g., narcissism), noting that situation-specific presentations and presentations that generalize across many situations are not mutually exclusive categories (p. 384). Thus, their emphasis is on the contextual determinants of self-presentation, which is consistent with the definition of *impression management* given by Paulhus and Vazire (2007).

Impression management is relevant to parenting capacity evaluations. Based on conceptualizations from Paulhus (2003) and Nunnally and Bernstein (1994), it seems that many referred parents responding to questions about child behaviors or parenting practices are likely to adjust their responses to reflect socially prescribed functioning. More specifically, parents may offer exaggerations such as, “I never have any trouble using timeout effectively” or unrealistically positive reactions to child behavior such as, “He soiled his new clothing, but I was really pleased to see him having so much fun.” In this manner, impression management—and other nonmutually exclusive self-presentations—show high relevance to the evaluation of parenting capacity.

To summarize, parenting capacity evaluation can be enhanced by the use of well-validated and reliable instruments such as the CAP. Yet, even instruments backed by extensive research show areas for improvement. The L scale of the CAP is a reasonable and important target for further investigation and refinement. In the next section, the PAI (Morey, 2007) is reviewed as a criterion measure relevant to the investigation of desirable responding and the psychometric characteristics of the CAP L scale.

### **The Personality Assessment Inventory**

The Personality Assessment Inventory (PAI; Morey, 2007) is a 344-item self-report, objective test of personality and psychopathology developed to assist clinicians in screening, diagnosing, and treatment planning. The PAI is comprised of 11 clinical scales (e.g., Depression, Alcohol Problems), 5 treatment scales (e.g., Aggression, Nonsupport), 2 interpersonal scales (i.e., Dominance, Warmth), and 4 validity scales (e.g., Positive Impression Management, Negative Impression Management). All scales subsume unique items; that is, scales do not overlap.

A number of sources indicate that the PAI has increased in popularity across the two decades since its introduction. Over ten years ago, Piotrowski (2000) summarized survey data from American Psychological Association accredited clinical training programs and internship sites that ranked the PAI within the top four most frequently used objective personality inventories. Edens, Cruise, and Buffington-Vollum (2001) described the PAI as a popular forensic instrument with an evidence base sufficient to support its testimonial admissibility in legal proceedings (see also Mullen & Edens, 2008 for a case law survey of the PAI).

Studies published in a special issue of the *Journal of Personality Assessment* (Kurtz & Blais, 2007) lend further support to the popularity of the PAI by way of its validation with a variety of populations and clinical problems; for example, male batterers (Chambers & Wilson, 2007), trauma brain injury (Kurtz, Shealy, & Putnam, 2007), Borderline Personality Disorder (Jacobo, Blais, Baity, & Harley, 2007; Stein, Pinsker-Aspen, & Hilsenroth, 2007), aggression in veterans diagnosed with combat-related trauma (Crawford, Calhoun, Braxton, & Beckham, 2007), and prediction of violence among incarcerated males (Walters, 2007) and institutional misconduct among incarcerated females (Skopp, Edens, & Ruiz, 2007). Validation studies have persisted in this vein as evidenced by results from a PsycINFO keyword search that returned 119 investigations of the PAI published in English language, peer-reviewed journals from the time of the aforementioned special issue to the present time (August 2011). In sum, the PAI appears to be a measure of wide acceptance with an evidence base relevant to multiple clinical and forensic problems and populations.

### **Development of the PAI: Test Construction, Validity, and Reliability**

Morey (2007) has described the development of the PAI as based on “both rational and empirical methods of scale development” (p. 1). The classical test theory (e.g., Chronbach &

Meehl, 1955) approach to convergent validity and discriminant validity figured prominently into the development of the PAI. However, test construction also followed rationale from item response theory (IRT) to ensure *breadth* and *depth* in sampling content domains (Morey, 2007). For example, the PAI response format is on a four-point Likert scale (*False, Not At All True to Very True*) intended to measure the *severity* and *intensity* of clinical problems.

Two early versions of the PAI were administered to four different samples, and items and scales were subjected to further empirical refinement. Inspection of item means and item standard deviations was undertaken to ensure depth of content sampling. Inspection of scale internal consistencies was undertaken to ensure breadth of content sampling. The resultant 344-item PAI was standardized using (a) a U.S. Census-matched community sample ( $N = 1,000$ ), (b) a clinical sample ( $N = 1,265$ ), and (c) combined college samples ( $N = 1,051$ ). When interpreting profiles from individual administrations, the respondent's scores are compared to the census-matched community sample.

The three standardization samples showed median internal consistency estimates of .81, .86, and .82, respectively (Morey, 2007). The census-matched sample showed internal consistency estimates of .72 for the Negative Impression Management (NIM) scale and .71 for the Positive Impression Management (PIM) scale. Internal consistencies of the clinical scales for the census-matched sample ranged from .90 to .74, and the internal consistencies of the treatment scales for the census-matched sample ranged from .85 to .72. Estimates showed little variability across age, gender, or race/ethnicity.

Test-retest reliability of the PAI was examined using a combined sample of community adults and college students ( $N = 155$ ) (Morey, 2007). Test administrations were approximately one month apart, and mean *T*-scores showed, on average, changes of approximately two *T*-score

points or less across administrations. Also, the pattern of scale elevations and high point scale elevations showed moderate configural stability (Morey, 2007). For example, among respondents with a clinically significant high point at the first administration, approximately 77% had the same clinically significant high point at the second administration.

In addition to a strong conceptual rationale, large and nationally representative standardization sample, and good internal consistency and test-retest reliability; researchers have amassed a large body of evidence for the convergent and discriminant validity of the PAI. Morey (2007) has reviewed these validity studies in chapter 9 of the second edition of the *PAI Professional Manual*. Further critical analysis of these studies is beyond the scope of the present review. However, findings for scales relevant to the research questions and hypotheses of the present study are reviewed in subsequent sections below (i.e., scales associated with impression management and risk for maltreatment). Scales of interest to the present study were: PIM, Depression, Stress, Nonsupport, and Aggression. PIM is reviewed in the next section.

### **Positive Impression Management**

The Positive Impression Management (PIM) scale is a nine-item measure of self-presentation. It is one of four primary scales used to evaluate the validity of PAI profiles. As with the CAP L scale, PIM is defined by endorsement of unrealistically positive beliefs and behaviors that are not commonly endorsed in the general population (Morey, 2007). For instance, sample PIM items read: “There have been times when I could have been more thoughtful than I was” and “Sometimes I let little things bother me too much.” Thus, when a respondent produces a PIM scale elevation, the resultant PAI profile is likely to be distorted in a positive direction; that is, the profile is likely to show few indicators of mental health concerns, environmental stressors, or interpersonal problems. Morey’s rationale for construction of the

PIM scale indicates a broad definition of impression management and self-enhancement consistent with Paulhus and Vazire (2007; Paulhus, 2002). Furthermore, Peebles and Moore (1998) reported a correlation of .71 between the PIM scale and the Impression Management scale of the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984) and .75 between the PIM scale and the Self-Deceptive Enhancement scale of the BIDR. However, consistent with Nunnally and Bernstein (1994), Morey does not distinguish self-presentation as conscious versus unconscious or personality trait versus situational demand.

A number of findings, including the ones reported by Peebles and Moore (1998), provide evidence for the convergent and discriminant validity of the PIM scale. For example, in a study evaluating positive self-presentations among parents undergoing parenting capacity evaluations; Carr and colleagues (2005) reported a .60 convergent correlation for the PIM scale and the MMPI-2 Lie (L) scale. Carr and his colleagues also reported a -.50 divergent correlation for the PIM scale and the MMPI-2 Infrequency (F) scale.

Other researchers have conducted validation studies using criterion-group strategies to evaluate the classificatory ability of the PIM scale. These studies typically involve random assignment of participants to instructional sets such as “fake-good,” “fake-bad,” or “standard responding” (e.g., Morey & Lanier, 1998). Some investigators have evaluated instructional sets with specific populations such as persons with substance use disorders (e.g., Fals-Stewart, 1996; Fals-Stewart & Lucente, 1997). In most studies, fake-good respondents are treated as “naïve.” That is, they are not given strategies to achieve a fake-good presentation. In one study, however, Baer and Wetter (1997) “coached” fake-good respondents by making them aware of the presence and function of the PIM scale. Still other researchers have provided incentives (e.g., monetary)

or stimuli (e.g., role-play scenarios such as involvement in child custody litigation) to improve the ecological validity of the analogue situation.

Collectively, findings from these investigations indicate several recommendations for effective use of the PIM scale. First, researchers have reported that invalid profiles occur at a much higher rate among populations who have much at stake with regard to the outcome of assessment. For example, Carr and colleagues (2005) reported invalid PAI profiles at a rate of approximately 18% using the liberal 68*T* cutoff score. This rate is nearly six times higher than the rate of approximately 3% observed in the community standardization sample (Morey, 2007). Likewise, Fals-Stewart (1996; Fals-Stewart & Lucente, 1997) reported higher rates of invalid profiles among respondents with substance use disorders who were referred for forensic evaluation by the legal system. These findings suggest that base rates for desirable responding are likely to approach 50% using more conservative cutoff scores (i.e.,  $\leq 57T$ ).

Second, and directly related to the use of more conservative cutoff scores; Cashel, Rogers, Sewell, and Martin-Cannici (1995), Morey and Lanier (1998), and Peebles and Moore (1998) reported findings that indicate a PIM cutoff score of 57*T* as providing an optimal balance between sensitivity and specificity. Morey and Lanier conducted receiver operating characteristic curve analyses with different PIM cutoff scores and different base rates for desirable responding. Their table of rate estimates indicates that even lower cutoff scores such as 54*T* or 52*T* may be appropriate in situations where high rates of sensitivity are desirable (even at the expense of false positive classification); for instance, when screening for positive response distortion in parenting capacity evaluations.

Third, findings show that use of multivariable algorithms provide little incremental validity in classification accuracy (Morey & Lanier, 1998; Peebles & Moore, 1998) and are

difficult to replicate in cross-validation studies (Fals-Stewart & Lucente, 1997). With regard to the former concern, the Defensiveness Index (DEF)—which takes into account eight aspects of defensive responding in PAI profiles, including elevations on the PIM scale—did not account for significant variance beyond that of PIM in differentiating between fake-good versus standard response instructions (Morey & Lanier). Related to this finding, Baer and Wetter (1997) and Peebles and Moore found that DEF produced less accurate classification rates compared to PIM. Furthermore, Morey and Lanier reported that another multivariable estimate of response distortion, the Cashel Discriminant Function (CDF; Cashel et al., 1995), correlated with measures of positive *and* negative impression. In sum, these findings suggest that among the three possible measures of positive response distortion available for the PAI, the PIM scale is the best criterion. In fact, in one study (Peebles & Moore, 1998), the nine-item PIM scale outperformed the 40-item BIDR (Paulhus, 1984).

Examination of evidence for criterion validity, discriminant validity, and convergent validity support the conclusion that the PIM scale is a valid measure of desirable responding. Consistent with the goal of demonstrating the classification rates of the CAP L scale, the PIM scale is considered an appropriate criterion measure. Kurtz and Blais (2007) have rightly emphasized that “one does not wish for validity scales that are unduly sensitive and unnecessarily limit the assessor in drawing on as much potentially useful information as possible in making clinical inferences and decisions” (p. 1). Indeed, this consideration is relevant to the present investigation. Given invalidity rates as high as 49% (Carr et al., 2005), it is possible that the recommended CAP L scale cutoff provides too conservative an estimate of profile validity. Alternatively, it is possible that the recommended cutoff may be appropriately sensitive given the CAP’s function as a screening instrument and its application in high-stakes evaluations. In order

to evaluate these possibilities, the PAI PIM scale will serve as the criterion measure in a test of the classificatory performance of the CAP L scale across several cutoff scores and base rate estimates. The chief goal, then, will be to maximize CAP L scale sensitivity while ensuring the greatest number of interpretable CAP Abuse scale profiles.

### **Risk Factors for Maltreatment and Associations with Desirable Responding**

Contiguous with the principal goal of evaluating the CAP L scale, a secondary goal is to evaluate possible correlates of desirable responding as relevant to parenting capacity evaluations. Given the recommendation that parenting capacity evaluation is to be multifaceted (Budd et al., 2011), identification of factors associated with desirable response sets is considered a worthwhile endeavor. Identification of such factors may provide evaluators with (a) target areas for further assessment, (b) a set of conditions that could be altered to improve rates of valid responding, or (c) variables that given further study and refinement could be used to develop more sophisticated classificatory algorithms. Focal points for this endeavor may include parent mental health (e.g., depression) and coping ability (e.g., managing anger and aggression) as well as broader contextual factors (e.g., environmental stressors, social support). These areas have shown reliable associations with risk for child maltreatment and may also affect desirable responding in parenting capacity evaluations. Parental depression and aggression are reviewed first. Review of demographic characteristics, environmental stress, and social support follows. Finally, the method of measuring these five domains for the present study is detailed.

General mental health distress has been shown to be a statistically significant correlate of child abuse potential (Jellinek et al., 1992; Rinehart et al., 2005). More specifically, depression and aggression are well-documented, proximal risk factors for child maltreatment. For example, depression has shown consistent associations with risk for child abuse, accounting for unique

variance in addition to other robust variables such as cognitive disability, substance use, and cumulative trauma (Cohen, Hien, & Batchelder, 2008) and after controlling for socioeconomic status and family cohesion (Mammen, Kolko, & Pilkonis, 2002). More complex models have tested aggression as a mediator of the association between depression and maltreatment risk (Hien, Cohen, Caldeira, Flom, & Wasserman, 2010; Shay & Knutson, 2008). For example, Hien and her colleagues investigated the associations of depression as measured by structured clinical interview, anger arousal and reactivity as measured by the Novaco Anger Scale and Provocation Inventory (NAS-PI; Novaco, 2003), and risk for child maltreatment as measured by the CAP. Among study participants, 20% met current diagnostic criteria for a depressive episode and 54% reported a lifetime history of depression. The average CAP Abuse scale score for depressed participants was greater than 166. A test of mediation showed that anger arousal and reactivity partially mediated the association between diagnostic group and abuse potential. That is, depressed participants compared to nondepressed participants tended to experience greater anger reactivity, which in turn was associated with higher CAP scores.

Depression and aggression are most commonly observed among individuals. Yet, both depression and aggression influence relationships. Indeed, this is the basis for their investigation in studies of child maltreatment. Thus, there is reason to believe that depression and aggression are likely to affect parents' interactions with CPS, and more specifically, with professionals who conduct parenting capacity evaluations. However, the magnitude and direction of the effect may vary as a function of the severity of depressed mood or the pervasiveness of impairment in managing anger. For example, an individual showing mild to moderate symptoms of depression may attempt to reduce the aversive quality of the parenting evaluation by providing socially acceptable responses, even when such responses do not reflect her or his true experiences. An

individual showing moderate to severe depression, on the other hand, may engage in more candid or even exaggerated responding in order to elicit help and alleviate distress. Comparable arguments can be made for aggression and anger reactivity. In sum, depression and aggression are identified as potential individual-level correlates of desirable responding in parenting capacity evaluations.

Many of the aforementioned studies of depression and aggression investigated samples that were homogenous with regard to socioeconomic status and other demographic factors. For example, the urban sample recruited by Hien and colleagues (2010) was uniformly poor, urban, 100% female, and over 70% African-American. Within more heterogeneous samples, demographic factors have shown important associations with risk for child maltreatment. For example, Wu and colleagues (2004) identified five risk factors associated with substantiated child maltreatment in a birth cohort of nearly 200,000 Florida infants. The five risk factors were maternal tobacco use during pregnancy, having three or more children, being a Medicaid beneficiary, being single, and having an infant with low birth weight. Thirteen percent of cohort mothers showed three or more of these risks and accounted for 50% of all CPS cases for the birth cohort. With regard to other salient demographic characteristics, Bennett and colleagues (2006) documented young parental age, low parental educational attainment, and low parental occupational status as risk factors that discriminated mothers with histories of child maltreatment from mothers without known histories of child maltreatment (see also Sedlak & Broadhurst, 1996).

With regard to demographic factors, it is interesting to note that Bennett and colleagues (2006) were interested in mothers who concealed their history of child maltreatment. They found that mothers who concealed their status (28% of the sample) versus mothers who did not

conceal their status did not differ on demographic characteristics. Thus, demographic variables such as employment status and educational attainment can be misleading variables in the context of child maltreatment studies. They may mask other unseen or confounding contextual factors (Azar & Cote, 2002). Contextual factors such as environmental stress and social support are reviewed next.

Whereas parental depression, aggression, and certain demographic characteristics are risk factors observed at the level of the individual parent; other risk factors such as parental stress and social support are observed at the level of the family and the surrounding community and afford a broader, contextual analysis of risk for child maltreatment. Indeed, a number of studies in the child maltreatment literature have moved beyond singular focus on individual-level risk factors in keeping with scholarly and theoretical accounts that emphasize functional and contextual assessment (e.g., Azar et al., 1998; Cicchetti & Lynch, 1993).

A number of investigators have studied contextual factors that show direct and indirect effects on parents' risk for child maltreatment. For example, Crouch and Behl (2001) reported that parents' beliefs in corporal punishment were associated with increased risk for maltreatment measured with the CAP. This association, however, was moderated by parenting stress. In fact, this group of highly stressed parents produced an *average* CAP score of 212, three points below the cutoff recommended by Milner (1986). Other investigators have reported similar effects for parenting stress (Rodriguez & Green, 1997) and neighborhood stress (Guterman, Lee, Taylor, & Rathouz, 2009) as related to risk for child maltreatment. In the latter study, parents' negative perceptions of neighborhood processes were associated with increases in parenting stress, placing them at increased risk for physical or psychological abuse of a child as well as increased risk for child neglect. Collectively, these findings show that family-level stressors and

community-level stressors affect parenting practices. It is possible that parents who find themselves under high levels of stress may be less attuned to the appropriateness of their parenting strategies. Also, their day-to-day concerns about parenting and beliefs about child welfare systems are likely to be much different compared to parents under less stress. Thus it is conceivable that parents experiencing multiple stressors may adopt a more pragmatic approach to answering an evaluator's questions about parenting capacity.

Other investigators have pointed up the important but complex effects of social support as related to child maltreatment. Lyons, Henly, and Schuerman (2005) reported that low levels of social support (i.e., instrumental, emotional) in conjunction with depression produced decreases in positive parenting (e.g., using social praise, assigning household chores). Yet, increases in social support produced little change in positive parenting and actually increased negative parenting (e.g., losing one's temper, using physical punishment). These mixed results can be explained, in part, by descriptive findings on neglectful mothers. For example, Coohy (1995) found that neglectful mothers rated *quality* and *source* of social support differently compared to mothers with no history of neglect. The two groups of mothers showed discrepant perceptions of social support that varied according to type of support (e.g., instrumental versus emotional) and source of support (e.g., one's mother versus one's partner). Also, Crouch, Milner, and Thomsen (2001) found that abused children who received early support were more likely to accept support as adults. Receiving support as adults, in turn, was associated with reduced risk for child maltreatment.

In sum, these results indicate that social support has different effects on risk factors associated with child maltreatment and that parents have different perceptions of social support depending on life experiences in receiving help from others. Those at greatest risk for child

maltreatment may show the greatest skepticism and greatest variability in their preferences for receiving support. Thus, interactions with CPS—and parenting capacity evaluations in particular—may further affect these parents’ perceptions of social support and occasion mistrust and impression management.

Depression, aggression, stress, and social support are factors that affect risk for child maltreatment. They may also be associated with desirable responding in parenting capacity evaluations. The PAI, as reviewed above, is a well-validated and reliable measure of personality and psychopathology (Morey, 2007). The PAI subsumes multiple clinical scales that are well-suited to measuring individual risk factors such as depression. The Depression scale (DEP) was developed to assess the primary symptoms of depression and the range of severity of the symptoms across affect, cognition, and physiology. DEP has shown good criterion validity with diagnoses of depressive disorders made by way of structured clinical interview. Furthermore, DEP has shown good convergent validity with other self-report measures of depression such as the MMPI–2 and the Beck Depression Inventory (BDI; Beck, Steer, & Brown, 1996). Also, DEP has shown good discriminant validity with measures of psychological health and well-being.

The PAI Aggression scale (AGG) is not a clinical scale. Morey (2007) has categorized AGG as a “treatment consideration scale” (p. 235). As such, AGG is thought to provide information about barriers to treatment not assessed by the clinical scales. Specifically, AGG was developed as a measure of anger, aggression, and the ability to cope with or manage anger and aggression. AGG has been used to discriminate persons with histories of violence from persons without histories of violence (Douglas, Hart, & Kropp, 2001). AGG has shown a convergent pattern of correlations with the State-Trait Anger Inventory (STAXI; Spielberger,

1999), especially with STAXI subscales of Trait Anger (i.e., angry, reactive disposition) and Anger Control (Morey, 2007). Also, AGG has shown convergent correlations with the Conflict Tactics Scale (CTS; Straus, 1979); for example, Psychological Aggression and Physical Violence.

In addition to clinical scales and treatment consideration scales, the PAI subsumes two measures of respondents' perceptions of their environmental circumstances, the Stress (STR) and Nonsupport (NON) scales. Morey (2007) has defined STR as a measure of "predictability, organization, and structure of the person's surroundings" (p. 250). NON is defined as a measure of the "availability and quality of supports in the environment that can potentially help in managing . . . stressors" (p. 250). Respondents who show elevations on PAI clinical scales tend to produce elevations on the STR and NON scales. STR has shown patterns of convergent correlations with measures of tension and strain, whereas NON has shown patterns of convergent correlations with resentment and hostility (Morey, 2007).

These four PAI scales (i.e., DEP, AGG, STR, NON) were investigated as possible correlates of desirable responding. Whereas these factors have shown reliable associations with risk for child maltreatment, their impact on the veracity of parental disclosures during parenting capacity evaluations has not been studied. One advantage of selecting this set of factors for exploratory analysis is with regard to multilevel coverage of risk factors. In particular, STR and NON afford a wider assessment of environmental factors that may affect parents' response sets on self-report measures such as the CAP. Collectively, these factors may provide useful information on ways to improve rates of valid responding during parenting capacity evaluations.

### **The Present Study**

The purpose of the present study was to investigate the validity and reliability of the CAP L scale with a heterogeneous sample of caregivers referred for parenting capacity evaluations. One aim of the study was to assess the measurement properties (i.e., homogeneity of content, internal consistency) of the L scale. A second aim was to evaluate the discriminative validity of the L scale by way of its receiver operating characteristics (ROC). A third aim of the study was to examine potential correlates of desirable responding on the L scale. In accord with the purpose and aims of this study, the following set of research questions and hypotheses were specified:

#### **Research Question 1**

What are the measurement properties of the CAP L scale with regard to homogeneity of content and measurement error?

**Hypothesis 1.** Given data from a heterogeneous sample of parents with diverse maltreatment histories, item discrimination with the CAP L scale would show items with poor ( $< .30$ ) item-total correlations; although, internal consistency would be in the acceptable (.71 to .80) or good (.81 to .90) range.

#### **Research Question 2**

Given any improvements in measurement properties derived from the tests of hypothesis-1, what is the classificatory performance of the CAP L scale?

**Hypothesis 2.** Using the PAI PIM scale as the criterion measure for desirable responding, the revised CAP L scale would show improvements in classification rates compared to classification rates reported by Milner and Crouch (1997) and would yield more information such as base rates of desirable responding and area under the curve (AUC) to aid clinicians in

their classificatory decisions. The greatest improvements in classification would be observed for specificity, or the ability to correctly identify valid profiles. (In the present study, sensitivity was defined as the correct classification of invalid profiles.)

### **Research Question 3**

Is depression, aggression, stress, or inadequate social support related to desirable responding on the CAP L scale after controlling for the effects of participants' demographic characteristics?

**Hypothesis 3.** This hypothesis was exploratory. It was expected that at least one of the variables would show statistically significant associations with desirable responding measured by the CAP L scale. The expected direction of the association (i.e., positive, negative) was unspecified. Given that this hypothesis was exploratory, results are interpreted with caution and with the caveat that cross-validation of statistically significant results is necessary.

## **Method**

### **Participants**

Participants were 138 male and female caregivers who underwent parenting capacity evaluations at Washington County Children and Youth Services (WCCYS) in Washington, Pennsylvania between July 2006 and June 2011. Masters-level practicum students enrolled in an APA-accredited doctoral program in clinical psychology conducted the evaluations under the direct supervision of a licensed clinical psychologist.

WCCYS is a county government agency that responds to the community's concerns regarding the safety of children. Such concerns include physical abuse of a child, sexual abuse of a child, and child neglect. Agency involvement is available to any child or family in

Washington County based on need as determined by assessment and investigation by WCCYS staff.

### **Procedure**

University institutional review board approval was obtained for the use (in the present study) of data collected for routine clinical purposes. Records from the parenting capacity evaluations are the property of the supervising clinical psychologist and are stored alphabetically by last name in a locked filing cabinet in a locked room dedicated to the storage of the supervising psychologist's clinical records.

Records from all parenting capacity evaluations conducted between July 2006 and June 2011 were reviewed. To be considered for inclusion in the present study, caregivers completed the CAP and the PAI as part of their parenting capacity evaluations. In addition to test data, caregivers' records were reviewed for demographic data such as age, education, employment status, relationship status, and type or types of child maltreatment indicated. Caregiver was defined broadly as anyone 18 years of age or older whom WCCYS staff or the courts determined to be a current or prospective custodian of a child or children receiving services from WCCYS. Caregivers were excluded from the present study if they were under 18 years of age at the time their parenting capacity evaluation was conducted, or if they obtained a score of 75 or less on a standardized test of intelligence. Also, caregivers were excluded if their scores exceeded the recommended cut-off scores on the CAP Faking-bad index ( $RR \geq 6$  and  $IC \leq 5$ ) or the CAP Random-response index ( $RR \geq 6$  and  $IC \geq 6$ ) (Milner, 1986). Likewise, caregivers were excluded if their scores exceeded the recommended cut-off scores on the PAI Negative Impression Managements scale ( $NIM; \geq 92T$ ), the PAI Infrequency scale ( $INF; \geq 86T$ ), or the PAI Inconsistency scale ( $ICN; \geq 73T$ ) (Morey, 2007). Participants were not excluded on the

basis of elevated PAI PIM scores or CAP L scale scores as such elevations were the focus of the present investigation.

## Measures

**Demographic form.** (Appendix A). The demographic form was developed for this study and included information on type of maltreatment, caregiver relationship to the child, caregiver age, caregiver gender, caregiver relationship status, and caregiver educational attainment. These variables have shown associations with risk for child maltreatment (Bennett et al., 2006; Sedlak & Broadhurst, 1996; Wu et al., 2004).

**Child Abuse Potential Inventory (CAP).** The CAP (Milner, 1986) is a 160-item self-report measure of parenting attitudes and practices intended as a screening tool for the detection of physical child abuse in caregivers investigated by CPS agencies. CAP items are worded as statements and responses are marked “Agree” or “Disagree.” The CAP is a well-validated measure that aids in the identification of caregivers who report parenting beliefs and behaviors similar to those observed in samples of caregivers with known histories of physical child maltreatment. Researchers have shown that the CAP also yields correct classifications with samples of abusers heterogeneous for maltreatment type (Ondersma et al., 2005).

The CAP is comprised of six clinical scales (Distress, Rigidity, Unhappiness, Problems with Child and Self, Problems with Family, Problems from Others) that are subsumed by a superordinate Abuse scale. Sample Abuse scale items read: “Sometimes I fear that I will lose control of myself” and “Children should never cause any trouble.” The 77-item Abuse scale has been the focus of extensive validation (Milner, 1994; Walker & Davies, 2010). It has shown sensitivity in the range of 95.5% to 100% and specificity in the range of 88.2% to 96.3% (Milner

et al., 1985). Internal consistency estimates for the Abuse scale across physical abuse and neglect samples was .95 and .93, respectively (Milner, 1986).

In addition to the six clinical scales, three validity scales have been developed for the CAP: Inconsistency (IC), Random Response (RR), and Lie (L). The L scale is of primary interest to the present study. It is an 18-item scale developed to detect socially desirable responding. Sample L scale items read: "I am always a good person" and "I sometimes fail to keep all of my promises." According to Robertson and Milner (1983, 1985), the L scale has shown modest correlations with other measures of desirable responding such as the Marlowe-Crowne Social Desirability Scale and the MMPI Lie scale. Internal consistency estimates for the L scale were in the .60 to .84 range with some differences in estimates as a function of caregiver sex and physical abuse versus neglect (Milner, 1986).

**Personality Assessment Inventory (PAI).** The Personality Assessment Inventory (PAI; Morey, 2007) is a 344-item self-report, objective test of personality and psychopathology. PAI items are worded as statements and responses are made on a four-point scale with anchors ranging from "False, not at all true" to "Very true." The PAI is comprised of 11 clinical scales (e.g., Depression, Alcohol Problems), 5 treatment scales (e.g., Aggression, Nonsupport), 2 interpersonal scales (i.e., Dominance, Warmth), and 4 validity scales (e.g., Positive Impression Management, Negative Impression Management). All scales subsume unique items. When interpreting profiles from individual administrations, the respondent's scores are compared to a census-matched community sample of 1,000 respondents. Internal consistency estimates for the clinical scales for the census-matched sample ranged from .90 to .74

Five PAI scales are of interest to the present study: Positive Impression Management (PIM), Depression (DEP), Aggression (AGG), Stress (STR), and Nonsupport (NON). The PIM

scale is a nine-item measure of self-presentation as defined by endorsement of unrealistically positive beliefs and behaviors that are not commonly endorsed in the general population.

Sample PIM items read: “There have been times when I could have been more thoughtful than I was” and “Sometimes I let little things bother me too much.” Peebles and Moore (1998)

reported correlations of .71 and .75 for PIM and the Impression Management and Self-Deceptive Enhancement scales of the Balanced Inventory of Desirable Responding (Paulhus, 1984). Carr and colleagues (2005) reported a .60 convergent correlation for the PIM scale and the MMPI-2 L scale. Morey and Lanier (1998) reported on the receiver operating characteristics of PIM and sensitivity estimates were in the range of .82 to .98, and specificity estimates across the same cut scores were in the range of .93 to .67. The census matched PAI normative sample showed internal consistency estimates of .71 for PIM.

DEP was developed to measure the severity of depression symptoms across affect, cognition, and physiology. Sample DEP items read: “Much of the time I am sad for no real reason” and “Sometimes I think I am worthless.” According to Morey (2007), DEP has shown good criterion validity with diagnoses of depressive disorders made by way of structured clinical interview and has shown good convergent validity with other self-report measures of depression such as the MMPI-2 and the Beck Depression Inventory (Morey, 2007). The census matched PAI normative sample showed an internal consistency estimate of .87 for DEP.

AGG was developed to provide information about barriers to treatment not assessed by the clinical scales. Specifically, AGG was developed as a measure of anger, aggression, and the ability to cope with or manage anger and aggression. Sample AGG items read: “I tell people off when they deserve it” and “People are afraid of my temper.” AGG has been used to discriminate persons with histories of violence from persons without histories of violence (Douglas et al.,

2001). According to Morey, AGG has shown a convergent pattern of correlations with the State-Trait Anger Inventory and the Conflict Tactics Scale. The census matched PAI normative sample showed an internal consistency estimate of .85 for AGG.

STR and NON were developed as measures of respondents' perceptions of their environmental circumstances. Morey (2007) has defined STR as a measure of "predictability, organization, and structure of the person's surroundings" (p. 250), whereas NON is defined as a measure of the "availability and quality of supports in the environment that can potentially help in managing . . . stressors" (p. 250). Sample STR items read: "My life is very unpredictable" and "Things are not going well in my family." Sample NON items read: "My friends are available if I need them" and "If I'm having problems, I have people I can talk to." According to Morey, STR has shown patterns of convergent correlations with measures of tension and strain, whereas NON has shown patterns of convergent correlations with resentment and hostility. The census matched PAI normative sample showed an internal consistency estimate of .76 for STR and .72 for NON.

## **Overview of Data Analyses**

### **Descriptive Analyses**

Descriptive findings were reported for the participants, including type of maltreatment indicated, caregiver relationship to the child, caregiver gender, caregiver race or cultural background, caregiver relationship status, caregiver employment status, child gender, and child age. When more than one child was named in the report, the age and gender of the youngest child was reported. Also, mean caregiver age and mean caregiver educational attainment was reported.

Descriptive statistics such as the mean, standard deviation, skew, and kurtosis for the CAP Abuse scale, the CAP validity scales, the PAI validity scales, DEP, AGG, STR, and NON are reported. The number of participants excluded due to elevations on the PAI NIM or INF scales or the CAP Faking-bad or Random-response indexes were reported.

### **Preliminary Analyses**

Prior to conducting the primary analyses, the study variables were analyzed for accuracy of data entry, missing values, outliers, and fit with the assumptions of the statistical models used to describe the data.

First, univariate descriptive statistics (e.g., mean, standard deviation) for the study variables were examined for out-of-range values to evaluate accuracy of data entry. Boxplots and Cleveland dotplots were examined to assess for the presence of univariate outliers. Outlying values were adjusted to fall on the whisker of the boxplot, where the whisker is the 75<sup>th</sup> percentile plus 1.5 times the interquartile range.

Second, the amount and pattern of missing data at the item-level was assessed. Tabachnick and Fidell (2007) have noted that 5% of data missing listwise and at random is an acceptable justification for deleting cases with missing values. However, to maintain statistical power for the primary analyses, it was proposed that any missing values at the item-level would be estimated by way of multiple imputation.

Third, data were assessed according to the assumptions of the statistical models used to test the study hypotheses. In general, pairwise plots of study variables (i.e., CAP L scale, DEP, AGG, STR, NON) were screened for nonlinearity and heteroscedasticity. The same variables were screened for nonnormal distributions of residuals. Given the presence of nonnormal distributions, square root, logarithmic, and inverse transformations were considered. Additional

diagnostic procedures were performed to assess for the presence of influential cases, or those cases with a combination of scores on study variables that are discrepant from all other cases in the sample (Tabachnick & Fidell, p. 74). To reduce the impact of multivariate outliers, it was proposed that scores for such cases would be adjusted, or the cases would be deleted from the primary analyses with rationale for the chosen solution given in the Results section.

For ROC analyses, the two CAP L scale distributions (obtained from splitting the sample on the criterion) were screened for equal variances (McFall & Treat, 1999; Nunnally & Bernstein, 1994). Given unequal variances, it was proposed that nonparametric estimation methods would be used.

### **Primary Analyses**

Of the 138 male and female caregivers who underwent parenting capacity evaluations at WCCYS between July 2006 and June 2011, it was expected that nearly half of the caregivers were, in some way, related. Examples of this type of relationship include married biological parents, separated biological parents, stepparents married to biological parents, stepparents not married to biological parents. In some cases, there were three or more caregivers related in this manner who underwent parenting capacity evaluations such as when a biological father and biological mother separate and then form new partnerships. When these situations occur, it is probable that data obtained from related caregivers will correlate more strongly than data obtained from unrelated caregivers. This nesting of caregivers within families violates the assumption of independence that is central to the statistical analyses to be used in this study.

Two methods were used to remove dependency from the data. For tests of hypothesis-1 and hypothesis-2, a subsample of unrelated participants was used. The subsample was drawn by first identifying individuals who currently share or have previously shared an identifiable

relationship such as husband-wife, girlfriend-boyfriend, ex-husband-ex-wife, etc. Once these related caregivers (two or more) were identified, one of the individuals was drawn at random for inclusion in the subsample. Statistical analyses—item-total correlations, estimates of internal consistency, and ROC curve analyses—were conducted on the subsample, which included 95 participants. For tests of hypothesis-3, the full sample of participants was used and mixed effects modeling was used to account for multiple caregivers nested within the same families.

**Hypotheses 1.** Nunnally and Bernstein (1994) emphasized low measurement error and homogeneity of content as important measurement properties of tests. In part, examination of item-total correlations and estimation of internal consistency values can be used to evaluate homogeneity of test content.

To evaluate homogeneity of CAP L scale content, corrected item-total correlations were used. Corrected item-total correlations are obtained by removing the sum of squares for the item under consideration from the total sum of squares. Because CAP items are dichotomous, point-biserial correlations were used. According to Nunnally and Bernstein (1994), item-total correlations tend to range from .00 to .40 with .30 being a reasonable cutoff for defining a discriminating item. In the present study, items that produced negative values were dropped. Items at or below .30 were considered for removal pending two additional considerations: (1) The effect that removal of the item would have on the balance of items keyed “Agree” versus “Disagree” and (2) the performance of the same item in reliability analyses.

To test the internal consistency and measurement error of the CAP L scale, the Kuder-Richardson formula 20 (KR-20) was used. In addition to estimating the internal consistency of the total scale, values of internal consistency with each item deleted was inspected. If deleting a given item improved internal consistency, then that item was considered for deletion. To reduce

subjectivity in making this determination, such an item was deleted only if it also showed a weak ( $\leq .30$ ) corrected item-total correlation. As the final step in the decision to drop items, only *pairs* of items were dropped so that the count of items keyed “Agree” and “Disagree” on the L scale remained balanced. This balance is considered important in order to maintain the utility of the Faking-good index.

Subsequent to deletion of any items, the corrected item-total correlations and KR-20 values were recalculated. This process was iterative until stable estimates were achieved.

**Hypothesis 2.** The ROC curve for the CAP L scale was plotted for a range of cutoff scores. The ROC curve is a graphical representation of the hit rate (i.e., true positive) of a scale as a function of its false alarm rate (i.e., false positive) over a range of cutting scores (McFall & Treat, 1999). Sensitivity is graphed on the abscissa, and 1 - specificity (i.e., false alarm rate) is graphed on the ordinate. The line of chance is a straight line that bisects abscissa and ordinate at the origin. At the bottom left corner no false positives are obtained and no true positives are obtained, whereas at the top right corner all true positives are identified at the cost of a 100% error rate for false positives (Pintea & Moldovan, 2009). The top left corner, on the other hand, reflects all true positives with no false positives or false negatives. The closer the ROC curve passes to the upper left corner of the graph, the better the classification rates of the scale and the larger the total area under the curve (AUC).

One advantage of the AUC statistic is that it is independent of the base rates of desirable responding and independent of cutoff scores on the L scale. AUC is interpreted as the probability of correctly classifying two randomly drawn observations taken from the two underlying distributions corresponding to presence versus absence of the phenomenon under consideration (McFall & Treat, 1999).

In the present study, the phenomenon of interest was the presence versus the absence of desirable responding as measured with the CAP L scale. The PAI PIM scale was the criterion against which the L scale was evaluated. That is, the PIM scale was used to divide the sample according to participants engaging in positive impression management versus participants not engaging in positive impression management. Morey (2007), in the *PAI Professional Manual*, recommended that PIM scores of 57T and higher are likely to affect accurate interpretation of the PAI clinical scales and treatment consideration scales. Furthermore, Morey and Lanier (1998) showed that PIM scores in the range of 54T to 61T yielded optimal sensitivity and specificity. For example, the 57T cutoff yielded sensitivity of .93 and specificity of .78. Consistent with these findings and the recommendations in the *PAI Professional Manual*, a cutoff score of 57T was used in the present study.

ROC curve analysis was conducted using the 18- item, original CAP L scale as well as the revised CAP L scale. The revised CAP L scale resulted from tests of hypothesis 1. The AUC statistic was reported as well as sensitivity and specificity rates across multiple cutoff scores. Positive predictive power and negative predictive power as a function of base rates also was reported. Results were summarized in table format to aid clinicians and forensic evaluators in selecting optimal L scale cut scores for their intended use of the CAP.

**Hypothesis 3.** Linear mixed effects models were used to examine correlates of desirable responding on the CAP L scale. The question under consideration was whether there was a relation between desirable responding, demographic variables (i.e., caregiver relationship to the child, maltreatment type), depression, aggression, stress, and nonsupport. Linear regression was considered an inappropriate statistical model given that CAP L scale scores showed stronger associations for caregivers from the same families compared to caregivers from different

families. Thus, desirable responding was modeled as a function of its hypothesized correlates, and the intercept for each family was allowed to vary at random. This is an instance of the random intercept model with family specified as a random effect (Zuur, Ieno, Walker, Saveliev, & Smith, 2009).

A multistep protocol outlined by Zuur and colleagues (2009) was used to test the random intercept model using R and its associated packages (R Development Core Team, 2008). The protocol involves screening procedures to ensure the mixed effect term for family is, in fact, warranted. Once screening procedures were complete, the model was evaluated in terms of the explanatory variables: caregiver relationship to the child, maltreatment type, DEP, AGG, STR, and NON. The variable obtaining the lowest statistically significance  $p$ -value ( $p < .01$ ) was dropped and the reduced model (minus that variable) was respecified. This process was iterative until an optimal model with statistically significant explanatory variables was obtained.

As noted previously, tests of hypothesis-3 were exploratory and results are interpreted as preliminary according to the following caveat: Statistically significant findings should be used solely for the purpose of (a) testing more refined hypotheses or (b) conducting cross-validation studies; Results of hypothesis-3 are not intended for direct clinical applications.

## **Results**

### **Characteristics of the Analytical Sample**

One hundred thirty-seven adults underwent parenting capacity evaluations. Twelve of these adults were excluded from the sample prior to estimating demographic information for the sample. First, participants were excluded if they had never before been a caregiver to the child named in the CPS report. For example, one of the persons evaluated was the fiancé of the child's biological father, and this woman had never before been a caregiver to the child. In sum, three

such cases were identified and excluded. Second, participants were excluded if they failed to complete sufficient PAI items or CAP items to produce valid test protocols. In sum, three such cases were excluded: One caregiver failed to complete four CAP IC scale items, one caregiver failed to complete 31 PAI items and 45 CAP items, and one caregiver failed to complete 52 PAI items. Third, participants were excluded if they obtained out-of-range values (see Morey, 2007) on the following PAI validity scales: ICN, INF, or NIM. Four participants showed scores  $\geq 73$  on the ICN scale and one participant showed a score  $\geq 92$  on the NIM scale. No participants showed scores  $\geq 86$  on the INF scale. Fourth, participants were excluded if they obtained out-of-range values (see Milner, 1986) on the following CAP validity indexes: Fake Bad and Random Response. One participant showed an elevated Fake Bad index defined as a Random Response scale score  $\geq 6$  and an Inconsistency scale score  $\leq 5$ . No participants showed elevated Random Response indexes defined as a Random Response scale score  $\geq 6$  and an Inconsistency scale score  $\geq 6$ . Fifth, for cases where results of intelligence testing was available ( $n = 6$ ), no participants were obtained full scale intelligence quotients  $\leq 75$  and, thus, none were excluded.

A total of 125 participants were included in the study. This core group of participants will be referred to as the full sample. Within the full sample, 57 participants within 27 families were found to have shared relationships through a given child; for example, two biological parents or a biological parent and a step-parent. To account for these types of shared relationships that violate the assumption of independence central to many of the statistical analyses conducted in this study, a subsample of unrelated participants was drawn from the full sample. In 24 of 27 families, there were two adults with a shared relationship to a given child. In 3 of 27 families, there were three adults with a shared relationship to a given child. Thus, one adult was drawn at random from each dyad or triad. This resulted in deletion of 30 caregivers

for a subsample of 95 participants representing 95 families. The necessity of this procedure was supported by linear regression diagnostics and the fitting of mixed effects models with a random intercept term for *family* as detailed under the Results subsection titled, “Regression diagnostics.” Mixed effects models were fit with the full sample of 125 participants. For some analyses (e.g., bivariate correlations, reliability analyses), results are reported for both samples. ROC curves were fit with the subsample of 95 unrelated participants.

**Data screening procedures.** Subsequent to excluding the twelve cases (detailed in the preceding section) and prior to calculating descriptive or inferential statistics, study variables were examined for data entry errors, missing values, and univariate outliers. Examination of all study variables at the item level and at the scale level showed plausible ranges, means, and standard deviations. There were no cases with missing data at the item level or at the scale level for any variables used to investigate the three primary hypotheses of this study.

With regard to univariate outliers, boxplots for the CAP L scale and PAI AGG, DEP, NON, and STR scales were inspected. Values for the CAP L scale were entered into the dataset as raw scores. Values for PAI scales were entered into the dataset as *T*-scores, which are standardized scores with  $M = 50$  and  $SD = 10$ . The boxplot of the CAP L scale showed no outlying cases. The boxplot for the AGG scale showed four outlying cases with values  $\geq 80T$ . These four cases were adjusted to fall at the whisker of the boxplot using the formula: [3rd quartile + 1.5(interquartile range)]. Using this formula, the adjusted value for AGG outliers was 75*T*. Re-examination of the new AGG boxplot showed that 75*T* was still an outlying value. Further adjustments were not conducted given that the 95th percentile for AGG was 73.4*T* and values above the 95th percentile in a sample of  $N = 125$  are expected. The boxplot for the DEP scale showed two outlying cases with values  $\geq 85T$ . These two cases were adjusted to fall at the

whisker of the boxplot. The adjusted value for DEP outliers was 83T. Re-examination of the new DEP boxplot showed no outlying cases. The boxplot for the NON scale showed one outlying case with a value of 102T. This case was adjusted to fall at the whisker of the boxplot. The adjusted value for the NON outlier was 81.5T. Re-examination of the new NON boxplot showed no outlying cases. The boxplot for the STR scale showed no outlying cases.

**Demographic characteristics.** Results for the full sample and subsample are shown in Table 1. Female caregivers constituted 52% of the full sample and approximately 53% of the subsample. Female and male participants were between 18 and 55 years of age. Mean female caregiver age for the full sample and subsample was 29.52 years ( $SD = 7.67$ ) and 30.74 years ( $SD = 7.48$ ), respectively. Mean male caregiver age for the full sample and subsample was 34.00 years ( $SD = 9.35$ ) and 34.80 years ( $SD = 9.83$ ), respectively. Just under 90% of full-sample and subsample participants were biological mothers or fathers, and just under 10% of full-sample and subsample participants were other male or female caregivers. One study participant was awaiting the results of paternity testing.

Children of study participants were between the ages of 1 and 192 months. Mean female child age for the full sample and subsample was 62.12 months ( $SD = 65.59$ ) and 71.68 months ( $SD = 68.13$ ), respectively. Mean male child age for the full sample and subsample was 53.89 months ( $SD = 61.43$ ) and 51.19 months ( $SD = 58.58$ ), respectively. Thus, greater variability was present in *child ages* across the full sample and the subsample compared to variability in *caregiver ages* across the full sample and the subsample.

Overall, demographic data from the full sample of caregivers was similar to demographic data from the subsample of caregivers. From each sample, approximately 86% of participants were White, approximately 11% were Black or African American, and approximately 3% were

biracial. For each sample, approximately two-thirds of participants had at least a high school diploma or GED. Approximately half of each sample was employed. Approximately 70% of each sample was in a relationship. More fine-grained distinctions are shown in Table 1, including missing values for caregiver education, caregiver employment status, and caregiver relationship status.

**Descriptive and inferential statistics for maltreatment type.** In addition to demographic variables such as age, sex, and employment status; participants were classified according to the following maltreatment types: physical only, sexual only, neglect only, emotional only, truancy/other legal only, and more than one type of abuse (see Table 1 for classification rates by sex). The most common type of child maltreatment observed for the full sample was neglect (37.6%) followed by multiple types (28.8%), physical (25.6%), truancy/other legal (3.2%), emotional (2.4%), and sexual (2.4%). Similar rates across maltreatment types were observed in the subsample.

A one-way analysis of variance (ANOVA) for CAP L scale scores by maltreatment type was conducted. Cell frequencies for truancy/other legal maltreatment, emotional maltreatment, and sexual maltreatment were low; therefore, these classifications were not included in the analyses. To meet the assumption of independence required for ANOVA, analyses were conducted on the subsample ( $n = 95$ ). With low-frequency maltreatment types excluded from the analyses, the subsample was further reduced to  $n = 87$ . Levene's test was significant at  $p = .017$ , indicating heterogeneity of Lie scale variance across maltreatment types. Heterogeneity of variance was further supported by visual comparison of boxplots of each maltreatment type. Howell (2007) has recommended Welch's procedure for evaluating main effects and the Games-Howell procedure for post hoc comparisons when heterogeneity of variance is present.

ANOVA results (reported as Welch's  $F$ ) showed no statistically significant difference across maltreatment types, [ $F(2, 48.70) = 0.82, p = .447$ ], and none of the pairwise comparisons computed by way of the Games-Howell procedure showed statistically significant differences. In sum, L scale scores were comparable across the three maltreatment types: neglect only, physical only, and multiple types.

**Descriptive statistics for the study variables.** The CAP L scale was central to the hypotheses of this study. Values of the L scale in the full sample ranged from 0 to 18 ( $M = 9.19, SD = 4.00$ ). Values for the L scale in the subsample ranged from 1 to 18 ( $M = 8.93, SD = 3.92$ ). The L scale cutoff score recommended by Milner (1986) is 7. Thus, the sample for this study was, on average, above the recommended L scale cutoff score. The CAP Abuse scale was not the focus of any of the hypotheses investigated in this study; however, it is worth noting that the mean CAP Abuse score in the full sample and subsample was 105.96 ( $SD = 91.81$ ) and 110.55 ( $SD = 92.78$ ), respectively. This is substantially lower than the range of mean Abuse scale scores (170.0 to 308.2) across parents with confirmed histories of physical maltreatment, sexual maltreatment, and neglect reported by Milner (1986).

Five PAI scales were investigated in hypotheses two and three of this study. Descriptive statistics for PAI scales are reported as  $T$ -scores. The PAI PIM scale was used as the criterion for desirable responding in hypothesis two. It ranged from 15 $T$  to 77 $T$  ( $M = 55.51, SD = 12.19$ ) in the full sample and 22 $T$  to 75 $T$  ( $M = 54.66, SD = 11.45$ ) in the subsample. The PIM scale cutoff score recommended by Morey (2007; Morey & Lanier, 1998) is 57 $T$ . The PAI AGG, DEP, NON, and STR scales were investigated in hypothesis three. The ranges, means, and standard deviations for these scales for the full sample and the subsample are shown in Table 2.

Also shown in Table 2 are values for CAP validity scales, PAI validity scales, and values for skew and kurtosis for all study variables for the full sample and the subsample.

### **Hypothesis 1**

Under hypothesis 1, homogeneity of CAP L scale content was investigated. It was hypothesized that the psychometric properties of the CAP L scale could be improved by examining corrected item-total correlations and internal consistency estimates and subsequently deleting scale items producing unsatisfactory values.

Values for the CAP L scale items prior to conducting any tests associated with hypothesis 1 are shown in Table 3. Estimates for the full sample and subsample are given and the estimates of internal consistency for the 18-item L scale were .813 and .800, respectively.

**Bivariate correlations for the CAP L scale.** Phi coefficients for all pairs of items for the 18-item L scale are shown in Table 4. Full sample coefficients are shown below the diagonal and subsample coefficients are shown above the diagonal. Inspection of the coefficients showed six L scale items that did not correlate well with other L scale items across both samples: 3, 5, 10, 11, 15, and 17. These six items showed few statistically significant correlations, and the majority of coefficients for these six items were in the range .00 to .22. For the full sample, item 2 also showed a pattern of low, statistically nonsignificant coefficients. For the subsample, item 4 and item 14 also showed a pattern of low, statistically nonsignificant coefficients.

**Homogeneity of CAP L scale content.** The purpose of this set of analyses was to identify and eliminate CAP L scale items (a) that showed corrected item total correlations  $< .30$  and (b) that produced any improvement in scale reliability if deleted. However, items that violated both criteria were not eliminated from the CAP L scale if doing so resulted in an upset in the balance of items keyed *Agree* and *Disagree*. This set of analyses was performed for the

subsample and the full sample. The subsample results were used in tests of hypothesis 2 of this study, and the full sample results were used in tests of hypothesis 3 of this study. Results are shown in Table 5.

***Full sample CAP L scale revisions.*** Reliability analyses were conducted in six iterations. In the first iteration, three items showed corrected total correlations  $< .30$  and improvements in internal consistency if deleted. The worst values were obtained for the item, “I never listen to gossip.” This item was deleted. A second iteration of reliability analyses was conducted, and two items showed problem values for the two criteria. The worst values were obtained for the item, “I sometimes act silly.” This item was deleted. A third iteration was conducted, and one item showed problem values: “I sometimes think of myself before others.” This item was deleted. A fourth iteration was conducted, and one item showed problem values: “I sometimes think of myself first.” This item was deleted. A fifth iteration was conducted, and all items showed acceptable values; however, there was not a balance of items keyed *Agree* and *Disagree*. To restore the balance, the last deleted item keyed *Agree*, “I sometimes think of myself before others,” was reinstated for the sixth iteration. The sixth iteration was conducted with a 14-item CAP L scale. Thirteen items showed acceptable values for the two criteria. The item, “I sometimes think of myself before others,” showed problem values for the two criteria; however, it was retained in order to maintain the balance of items keyed *Agree* and *Disagree*. The final revised scale for the full sample contained 14 items and yielded a full-scale reliability coefficient of .828.

***Subsample CAP L scale revisions.*** Reliability analyses were conducted in seven iterations. In the first iteration, four items showed corrected total correlations  $< .30$  and improvements in internal consistency if deleted. The worst values were obtained for the item, “I

never worry about my health.” This item was deleted. A second iteration of reliability analyses was conducted, and three items showed problem values for the two criteria. The worst values were obtained for the item, “I sometimes think of myself before others.” This item was deleted. A third iteration was conducted, and three items showed problem values. The worst values were obtained for the item, “I sometimes think of myself first.” This item was deleted. A fourth iteration was conducted, and two items showed problem values. The worst values were obtained for the item, “I sometimes act silly.” This item was deleted. A fifth iteration was conducted, and one item showed problem values: “I never listen to gossip.” A sixth iteration was conducted, and all 13 items showed acceptable values for the two criteria; however, there was not a balance of items keyed *Agree* and *Disagree*. To restore the balance, the last deleted item keyed *Agree*, “I sometimes act silly,” was reinstated for the seventh iteration. The seventh iteration was conducted with a 14-item CAP L scale. Thirteen items showed acceptable values for the two criteria. The item, “I sometimes act silly,” showed problem values for the two criteria; however, it was retained in order to maintain the balance of items keyed *Agree* and *Disagree*. The final revised scale for the subsample contained 14 items and yielded a full-scale reliability coefficient of .817.

## **Hypothesis 2**

Under hypothesis 2, ROC curves for the 18-item CAP L scale and the revised 14-item CAP L scale were estimated. It was hypothesized that the ROC curve for the revised 14-item CAP L scale would yield a higher value for the area under the curve (AUC) statistic and superior rates of sensitivity and specificity compared to the 18-item CAP L scale. All tests of hypothesis two were conducted on the subsample ( $n = 95$ ).

The PAI PIM scale was the criterion for determining group membership. Participants with PIM scores  $\geq 57T$  were classified as engaging in positive impression management or “fake-good” responding and were coded “1” in the dataset. Participants with PIM scores  $< 57T$  were classified as engaging in honest responding and were coded “0” in the dataset.

Histograms for the 18-item L scale and the 14-item L scale approximated a normal distribution for the  $n = 95$  participants. Values for skew and kurtosis (see Table 2) supported this finding. After splitting the sample using the PIM criterion of  $57T$ , 51 participants were classified as fake-good responders and 44 participants were classified as honest responders. Histograms for the 18-item L scale and the 14-item L scale by response type (i.e., honest, fake good) are shown in Figure 1. Compared to the L and revised-L histograms for the  $n = 95$  respondents, the four histograms in Figure 1 show less visual evidence for normality.

Given the distributions observed in the four histograms and given unequal subsamples of honest responders versus fake-good responders, the AUC statistic was estimated in SPSS version 17.0 using nonparametric or distribution-free methods. Values for AUC range from 0.00 to 1.00, with 0.50 describing a model that is no better than chance (McFall & Treat, 1999). The AUC statistic for the 18-item L scale performed at a level that was significantly better than chance,  $AUC = .924, p < .0001, 95\% \text{ CI } [.875, .973]$ . The resultant AUC value is defined as a 92.4% rate of correctly classifying any two randomly drawn observations where one observation is drawn from each of the two underlying distributions of fake good responding versus honest responding (McFall & Treat, 1999). The AUC statistic for the 14-item L scale also performed at a level that was significantly better than chance,  $AUC = .928, p < .0001, 95\% \text{ CI } [.881, .975]$ . The resultant AUC value is defined as a 92.8% rate of correctly classifying any two randomly drawn observations from each of the two underlying distributions. Thus, AUC is an estimate of

overall performance and indicates comparable rates of classification for the 18-item L scale and the 14-item L scale.

In order to establish the optimal cutoff point for each scale, the sensitivity and specificity for each possible cutoff point was calculated and is shown in Table 6. The best performance for the 18-item L scale with an emphasis on sensitivity (i.e., correct classification of fake-good responding) is obtained at a cutoff score of 8. The best performance for the 14-item L scale is obtained at a cutoff score of 6. This result is shown visually in Figure 2, which graphs the ROC curve for each scale. Optimal sensitivity/specificity is obtained from the points on the ROC curves nearest to the upper left corner of the graph area. Thus, the optimal values for the 18-item L scale were sensitivity = .961 and specificity = .705 and for the 14-item L scale were sensitivity = .882 and specificity = .750.

A practical test of the utility of these results was conducted by using the L scale cutoff score of 8 and the revised L scale cutoff score of 6 to reclassify study participants ( $n = 95$ ). Results are shown in contingency table format in Table 7. Comparison of the two scales at their respective optimal cutoff points indicates several tradeoffs. For example, the L scale shows a greater total correct classification rate and lower false negative rate compared to the revised L scale. The revised L scale shows a lower false positive rate compared to the L scale.

### **Hypothesis 3**

Hypothesis 3 was exploratory. Its aim was to investigate correlates of desirable responding. Desirable responding was the outcome variable and was measured by way of the 18-item CAP L scale and the 14-item revised CAP L scale. Separate models were fit for the two L scales. Hypothesized correlates included caregiver relationship to the child, maltreatment type, caregiver aggression, caregiver depression, caregiver stress, and caregiver perceived lack of

social support. The latter four variables were taken from the following PAI scales: AGG, DEP, NON, and STR. All tests of hypothesis 3 were conducted on the full sample ( $N = 125$ ). All analyses for hypothesis 3 were conducted using the open-source statistical software R (R Development Core Team, 2008) and the following R packages: lattice graphics (Sarkar, 2008) and linear and nonlinear mixed effects models (Pinheiro, Bates, DebRoy, Sarkar, & the R Core Team, 2008).

**Regression diagnostics.** Linear regression diagnostics were conducted for models fitted with the 18-item L scale and the 14-item revised L scale. First, univariate outliers were adjusted for AGG, DEP, and NON as described above in the subsection titled, “Data screening procedures.” Second, histograms of these four variables were examined for normal distributions. Visual inspection of the histograms showed positive skew. Visual inspection of pairwise scatterplots showed linear associations with the L scale and revised L scale. A logarithmic base 10 transformation reduced positive skew. However, the variables were not transformed in order to retain the original metric (i.e., *T*-scores) of the explanatory variables. In further support of this decision, parameter estimates were nearly identical for separate models fit with transformed versus untransformed explanatory variables. Third, variables were evaluated for collinearity. Correlations between the six explanatory variables were in the range | .02 to .60 |, indicating that collinearity was not a problem. Bivariate correlations are shown in Table 8. Fourth, two linear regression models were fitted corresponding to the two outcome variables, the L scale and the revised L scale. The six explanatory variables were the same for each model. For both models, there was no discernable pattern in the scatterplot of residual values versus fitted values, indicating homogeneity of variance. Furthermore residuals for both models were normally distributed. Fifth, influence statistics were calculated to determine if any cases were multivariate

outliers. For both models, no cases showed standardized DFBetas  $> 1.00$ , indicating that no cases were exerting undue influence on model parameters (Tabachnick & Fidell, 2007).

Likewise, no cases showed values for Cook's distance  $> 1.00$  (Cook & Weisberg, 1982).

Finally, to evaluate the assumption of independence, residuals for both models were plotted against the nominal variable, *family*. As described above, family had 95 levels reflecting 95 distinct study families. Examination of the boxplots of the standardized model residuals showed that the spread of residuals was different across families, indicating that a random term for family was warranted. This was further supported by comparison of the fixed effects model (i.e., linear regression) to the mixed effects model where family was specified as a random effect and the six explanatory variables were specified as fixed effects. The linear regression models were fitted using the generalized least squares (GLS) method. The GLS method with no variance covariates specified produces the same linear regression model as the ordinary least squares method, but with the added ability to conduct model comparisons by way of the likelihood ratio test (*LR*; Zuur et al., 2009). First, model comparisons were conducted with the 18-item L scale set as the outcome variable. Results showed that the model including the random intercept for family was a significantly better fit to the data,  $LR(1, N = 125) = 1.88, p < .01$ . Second, model comparisons were conducted with the 14-item revised L scale set as the outcome variable. Again, results showed that the model including the random intercept for family was a significantly better fit to the data,  $LR(1, N = 125) = 3.65, p < .005$ . For both random intercept models, there was no discernable pattern in the scatterplot of residual values versus fitted values, indicating homogeneity of variance. The intraclass correlation coefficient for the model using the 18-item L scale as the outcome variable was .25, and the model using the 14-item revised L scale as the outcome variable was .26.

**Model selection procedures.** To select the optimal fixed structure in terms of explanatory variables, all six variables were entered simultaneously into the mixed effects models. The least significant (i.e., lowest  $p$ -value  $> .05$ ) fixed effect was then removed from the model and a new model minus that least significant fixed effect was specified. The significance of the removed parameter (i.e., fixed effect) was then estimated by comparing the original model to the nested model (i.e., original model minus the least significant parameter) by way of the likelihood ratio test. This process was iterative until only variables with statistically significant  $p$ -values entering models with significant likelihood ratio tests remained.

For the model with the 18-item CAP L scale as the outcome variable, nonsignificant explanatory variables (i.e., fixed effects) were removed in the following order: NON, maltreatment type, caregiver relation, and DEP. The final model included a random effect for family and fixed effects for AGG and STR. Model summary statistics are displayed in the top half of Table 9. For the model with the 14-item revised CAP L scale as the outcome variable, the order of removal proved to be the same. Thus, the final included the same parameters: a random effect for family and fixed effects for AGG and STR. Model summary statistics are displayed in the bottom half of Table 9.

**Final model interpretation.** Interpretation of both models is essentially the same. The fixed effects portion of the models showed that at the individual level, caregivers reporting lower Aggression and Stress scores on the PAI showed higher scores or increased tendency to “fake good” on the CAP L scale (or revised CAP L scale). The models were linear; thus, the alternative interpretation also is valid: Caregivers reporting higher Aggression (AGG) and Stress (STR) scores on the PAI showed lower scores or decreased tendency to “fake good” on the CAP L scale (or revised CAP L scale). AGG and STR was coded in the dataset as  $T$ -scores, and the

CAP L scale and revised CAP L scale was coded in the dataset as raw scores. Thus, the model coefficients for the L scale model indicate that as AGG decreases by one *T*-score point, the L scale increases by .132. Stated in broader terms, a 10-point (i.e., one *SD*) decrease on the AGG scale (holding STR constant) is associated with a 1.32-point increase on the CAP L scale. The same interpretation holds true for STR and for the effects of AGG and STR on the revised L scale. The random intercept portion of the models showed that fixed model parameters (i.e., AGG, STR) vary by family. However, slopes for families were permitted to vary at random about the slope fitted for the fixed components. Thus, no specific results can be defined for the effect of family.

### **Discussion**

This study investigated the internal consistency and discriminant validity of the CAP L scale and correlates of desirable responding in a heterogeneous sample of 125 male and female caregivers referred for parenting capacity evaluations. The findings from this study provide new information about the psychometric properties of the CAP L scale and its application in clinical and forensic settings. Consistent with past findings (e.g., Carr et al., 2005; Ondersma et al., 2005), caregivers produced a high rate (74.4 %) of invalid CAP profiles by way of elevated L scale scores. The L scale showed little variation across caregivers from families with different maltreatment histories. Item analyses and estimates of internal consistency showed homogeneity of the L scale, though several problem items were identified. Deletion of these items, however, produced only marginal improvements in internal consistency. The 14-item revised scale that resulted from the item deletions showed tradeoffs in sensitivity and specificity compared to the original 18-item scale. Classificatory accuracy of the 18-item scale (with emphasis on sensitivity to detect fake-good responding) was best using a cutoff score that was one to two points higher

than recommendations given in the CAP manual (Milner, 1986). Last, the L scale showed inverse associations with stress and aggression, indicating plausible barriers to accurate forensic evaluation of parenting capacity when child maltreatment is indicated or suspected.

In subsequent sections, findings per study hypothesis are discussed with regard to forensic implications along with targeted assessment of study limitations and possibilities for future research with the CAP L scale.

### **Sample Characteristics**

Compared to the county in which the research was conducted (U.S. Census Bureau, 2012) fewer caregivers in the present study were white (85.6% versus 94.1%) and more caregivers were black or African American (11.2% versus 3.3%) and multiracial (3.2% versus 1.5%). These statistics should not be interpreted as suggesting a direct association for race and child maltreatment. Any greater representation of a particular racial group is likely due to a multiplicity of social, contextual (e.g., neighborhood), and economic factors (Azar & Cote, 2002). However, this was a less racially diverse sample than that studied by Ondersma et al. (2005) and that reported in the CAP manual (Milner, 1986). Mean caregiver age, range of caregiver ages, and caregiver education was similar to that reported in other studies (Carr et al., 2005; Milner, 1986; Ondersma et al., 2005). In sum, this sample was comparable to other samples in terms of several key demographics, though it was somewhat less diverse in terms of race compared to the CAP normative sample.

The relative heterogeneity of the present sample also was apparent in terms of maltreatment type. The predominant type of child maltreatment was neglect followed by multiple types of abuse and physical abuse respectively. Sexual, emotional, and other types of maltreatment were relatively uncommon. Mean comparisons of the neglect group, the multi-type

group, and the physical abuse group showed no statistically significant differences for L scale scores. This finding indicates that caregivers with disparate maltreatment histories do not differ in their propensity to engage in desirable responding. That is, there seems to be nothing inherent to maltreatment type to increase or decrease the likelihood that a caregiver will endorse unrealistic and highly virtuous personal and interpersonal qualities. For instance, a biological father indicated for physical abuse would be expected to show no greater or lesser tendency to respond desirably than a biological father indicated for neglect. Broadly interpreted, this finding indicates that descriptive or topographically defined features, compared to interpersonal and contextual factors, have less influence on desirable responding in parenting capacity evaluations.

Desirable responding was a substantial concern when interpreting clinical test data obtained from the present sample of caregivers. Consider that nearly 75% of caregivers produced invalid CAP profiles due to elevated Faking-good indexes, and the mean L scale score for the full sample was nearly two points higher than the recommended cutoff of seven (Milner, 1986). Furthermore, over half (52.8%) of the study participants produced questionable PAI profiles due to elevated ( $\geq 57T$ ) PIM scores. This is comparable to the rate of invalid profiles (49%) observed in the sample of caregivers studied by Carr and colleagues (2005) and substantially higher than the rate of invalid profiles (30.3%) observed in the samples of caregivers studied by Ondersma and colleagues (2005). The former sample was comparable to the sample studied in the present investigation in that study participants were caregivers referred for court- or CPS-ordered parenting evaluations, whereas the latter samples were different in that study participants were caregivers referred for prevention services or treatment. Thus, rates of invalid profiles appear to increase as the stakes of evaluation increase and as caregivers are evaluated for forensic versus clinical purposes.

To further examine demographic factors associated with the rates of invalid profiles across studies (i.e., Carr et al., 2005; Ondersma et al., 2005; the present study), broad comparison of rates of education and employment was made. There was no discernible pattern across the studies to suggest that less education (or employment) was associated with higher rates of desirable responding. In fact, the present study showed higher rates of caregiver education *and* higher rates of invalid profiles. This finding can be contrasted with data reported in the CAP manual that identified “a modest but consistent relationship between lie scale scores and education” and led Milner (1986, p. 11) to recommend an L scale cutoff score of eight for caregivers with less than a twelfth-grade education. The present study findings suggest, however, that the presence versus absence of a high school education may not be a sufficiently sensitive indicator of desirable responding. This conjecture is supported by findings reported by Budd, Heilman, and Kane (2000) where caregivers with elevated CAP Faking-good indexes were, on average, nearly one standard deviation lower in reading achievement scores compared to a control group of low abuse-risk caregivers with valid CAP profiles. There was, however, no statistically significant difference for years of education between the two groups. Thus, it can be hypothesized that education attainment may be, at best, a stand-in measure of reading ability. At worst, education attainment may be a confounding variable.

Last with regard to sample characteristics, it is worth noting that the mean CAP Abuse scale score for the present sample was approximately 106, whereas the mean CAP Abuse scale score for the normative sample was 91. The present sample was comprised of caregivers who were indicated or suspected child abusers; however, the CAP normative sample contained a preponderance of caregivers with no known history of child abuse. Examining only valid CAP profiles in the present study, the mean Abuse score was 155.41 ( $SD = 106.00$ ), which is

comparable to the mean Abuse score of 142.89 reported by Carr and colleagues (2005) for participants with valid CAP profiles. The mean Abuse scores across clinical samples studied by Ondersma and colleagues (2005) ranged from 143.8 to 183.2. Collectively, these findings indicate that Abuse scale scores vary inversely with L scale scores and may be further affected by the context of evaluation, though the latter factor has not been studied directly.

### **Homogeneity of CAP L Scale Content**

Discussion in this section corresponds to hypothesis-1 of the present study and examination of corrected item-total correlations and internal consistency estimates of the CAP L scale. As hypothesized, item analyses showed a number of items with poor discrimination. However, removal of these items had little effect on estimates of error variance for the scale.

Milner's (1986, 1982) construction strategy for the L scale involved selection of items designed to measure highly desirable social qualities that only 15% of normative samples with no known histories of child maltreatment endorsed. The goal, then, was to detect individuals who consistently endorsed multiple unrealistic, highly desirable social qualities. Examination of endorsement rates for the 18-item L scale in the present study are given in Table 3. As shown, caregivers endorsed socially desirable qualities at high rates. For 12 of the items, the rates of endorsing the desirable quality were greater than 50%. Several items (i.e., 9, 12, 15, 16), however, showed much the opposite effect, and the rates of endorsing the desirable quality were less than 25%. Content of the four items subsumes displays of basic emotion such as happiness and anger. Perhaps widespread cultural acceptance of anger, cursing, and silliness has increased since the items first were written. Regardless of the reason for higher null endorsement rates, caregivers responded to these items in the direction opposite of that originally intended by Milner.

It is worth noting, however, that three of these four items showed seven or more statistically significant correlations each with other Lie scale items (see Table 4). The item, “I sometimes act silly,” was the sole exception and showed no statistically significant correlations with other Lie scale items. This item also performed poorly in calculations of corrected item-total correlations, producing correlations below the .30 cutoff. (However, it was retained in the subsample version of the revised CAP L scale to preserve the balance of items keyed Agree-Disagree.)

There were five additional items that did not correlate well with other Lie scale items. These included: “I never worry about my health,” “I sometimes think of myself first,” “People sometimes take advantage of me,” “I never listen to gossip,” and “I sometimes think of myself before others.” The obvious and most basic interpretation of the findings is that these items measure a construct different from the construct or constructs subsumed by the CAP L scale. This interpretation also is supported by findings for corrected item-total correlations and internal consistency as shown in Table 5.<sup>2</sup> Examination of the 13 CAP L scale items that obtained adequate values on these two metrics showed content consistent with overly positive, unrealistic self descriptions. Closer examination of item content showed items that describe overt, observable behaviors such as “I never get mad at others” and “I never do anything that is bad for my health” as well as items that describe more covert, socially prescribed behaviors such as “I am always happy with what I have” and socially proscribed items such as “Sometimes I have bad thoughts.” Paulhus (2002; 1998) has attempted to distinguish facets of the desirable responding construct; for example, impression management and self-deceptive enhancement/denial with key distinctions of the facets based on conscious versus unconscious processes. No attempt was made in the present study to evaluate the extent to which a person is aware (i.e., attending)

versus unaware (i.e., not attending) of desirable responding. In fact, rather than attributing desirable responding to individual characteristics such as narcissism, self-deception, or unconscious processes; it seems more plausible that the coercive context (Budd, 2001) of the forensic parenting capacity evaluation affected participants' desirable responding.

Last with regard to homogeneity of L scale content, only minor improvements in internal consistency were observed when comparing the full, 18-item L scale to the revised 14-item L scale. Internal consistency between the two measures differed by only one-hundredth of a point. By many standards, reliability of .70 or better is desirable. However, Nunnally and Bernstein (1994, p. 265) cautioned that reliability of .80 may not be sufficient when making important decisions about the fate of individual, as in the present study, versus detecting group differences, as in research settings. Nunnally and Bernstein noted that decisions such as whether or not an individual has engaged in socially desirable versus honest responding often are based on small differences in test scores; for example, obtaining a seven versus an eight on the CAP L scale. Thus, the standard error of measurement (SEM) must be minimized. For example, the CAP L scale standard deviation for the study sample ( $N = 125$ ) was 4.00, and the internal consistency was .828. Using the formula  $SEM = SD \sqrt{1.0 - \alpha}$ ,  $SEM = 1.66$ . Thus, a participant with a score of 6 on the L scale would yield a 95% confidence interval of  $6 \pm 3.32$ . Interpreted another way, on 95 of 100 hypothetical administrations of the L scale, this person's observed score would fall between 2.68 and 9.32. This range of observed scores spans the conventional cutoffs for the L scale. Clearly this is not a narrow confidence band and points up limitations in the psychometric properties of the L scale and, more importantly, its ability to correctly classify caregivers as fake-good versus honest responders.

### **Receiver Operating Characteristics of the CAP L Scale**

Discussion in this section corresponds to hypothesis-2 of the present study and examination of the discriminative validity of the CAP L scale. As hypothesized, the revised 14-item L scale showed a superior AUC value; however, the magnitude of the difference was only two-hundredths of a point. Comparison of the ROC curves for the 18-item versus 14-item L scale yielded tradeoffs in sensitivity and specificity as shown numerically in Table 6 and visually in Figure 2. These tradeoffs were not sufficiently advantageous to recommend the revised scale over the original scale; thus, the majority of findings are discussed with regard to the original, 18-item L scale.

Discriminative validity is defined as the degree to which a measure differentiates individuals in groups formed by way of an independent criterion (Haynes & O'Brien, 2000). In the present investigation of the discriminative validity of the CAP L scale, the sample ( $n = 95$ ) was split almost evenly (46.32% honest) on the PIM criterion of *57T*. Using the L scale to sort caregivers into honest versus fake-good groups, optimal sensitivity and specificity was obtained with a cutoff score of eight for the 18-item L scale. At this cutoff, sensitivity—defined as correct classification of fake-good responding—was .961 and specificity—defined as correct classification of honest responding—was .705. The 14-item revised L scale yielded an optimal cutoff score of six with sensitivity of .882 and specificity of .750. The tradeoff between the two scales at these cutoff scores is in terms of higher sensitivity for the 18-item scale versus higher specificity for the 14-item scale. One goal of the present study was to determine whether or not CAP L scale specificity could be improved by undertaking revisions informed by estimates of homogeneity and internal consistency of scale content (i.e., hypothesis 1). This goal was not achieved in the present investigation as improvements in specificity were at the expense of

decrements in sensitivity. With emphasis correctly placed on detection of possible child abusers, L scale sensitivity is prioritized over specificity as the former determines detection of caregivers whose Abuse scale scores are likely to be affected by desirable responding. Thus, the 18-item L scale showed superior performance in the present study and produced a superior balance of sensitivity and specificity compared to the 14-item L scale. This conclusion gains further support when considering the fact that the rate of false positive classifications (i.e., persons incorrectly classified as fake-good respondents) differed by only 2.1% (i.e., 2 study participants) from the 18-item scale to the 14-item scale.

One finding from the ROC curve analyses conducted in this study with direct implications for CAP interpretation was the cutoff score of eight for the 18-item L scale. The cutoff score of seven recommended by Milner (1986) showed perfect sensitivity (1.00), but poor specificity (.568). This finding emphasizes the problem of obtaining high rates of invalid profiles or inflated false positive classifications. The solution to this problem, however, is not clear cut. First, Milner has prescribed a cut score of 8 for participants with less than a twelfth-grade education. Yet, this recommendation is complicated by the finding that this sample was better educated than the CAP normative samples and engaged in higher rates of fake-good responding, which suggests the opposite association for education and desirable responding. As previously discussed, reading ability may be a more sensitive indicator of desirable responding and warrants further investigation. Second, any decrement in sensitivity below 1.00 will inflate the rate of false negatives (i.e., missed occurrences of fake-good responding), which begs the question of whether or not it is appropriate to relax the detection of fake-good respondents in order to correctly identify more honest respondents and, thus, increase the number of interpretable CAP profiles. As posed, this question may be more amenable to legal, ethical, or

moral considerations; however, empirical analogues may be possible such as investigation of the conditions and factors that give rise to successful faking, where successful faking is defined as high abuse risk, low Abuse score, and low L score. In sum, it is recommended that clinicians and forensic evaluators approach data reported in Table 6 with caution: There is no substitute for establishing one's own local norms and standards.

### **Correlates of Desirable Responding**

Discussion in this section corresponds to hypothesis-3 of the present study and examination of the effects of depression, aggression, and contextual problems such as perceived stressors and lack of social support. This hypothesis was exploratory, as only one other study to date (Budd et al., 2000) has endeavored to identify factors associated with invalid CAP profiles. Budd and colleagues found that (a) adolescent mothers with invalid CAP profiles differed from (b) adolescent mothers with valid elevated CAP profiles and (c) adolescent mothers with valid normal CAP profiles. The group of mothers with invalid profiles showed statistically lower reading achievement and higher global emotional distress compared to the other two groups. The present study aimed to extend findings in the area of emotional distress by examining the effects of depression, aggression, stress, and social support on fake-good responding on the CAP L scale. The effects of caregiver relationship to the child and type of maltreatment also were examined.

Depression, aggression, stress, and nonsupport measured by corresponding PAI scales DEP, AGG, STR, and NON showed large, statistically significant bivariate correlations. This is consistent with findings reported by Morey (2007) with the exception of the relation between DEP and AGG, which showed lower magnitude of association in the PAI clinical (.28) and normative (.35) samples compared to the magnitude of association in the present sample (.60).

Caregiver relationship to the child and maltreatment type showed small, statistically nonsignificant bivariate correlations. Caregiver relationship has not been studied in previous investigations, but was included in the present study under the premise that a caregiver with a biological relationship to a given child might show a stronger investment and thus a greater propensity to conceal shortcomings compared to nonbiological parents. With regard to maltreatment type, Bennett and colleagues (2006) differentiated mothers who concealed a history of child abuse from mothers who acknowledged a history of child abuse on the basis of the type and intensity of maltreatment perpetrated. More specifically, mothers who acknowledged a history of child abuse had higher scores on measures of child maltreatment compared to mothers who concealed a history of child abuse. Bennett and colleagues used measures of maltreatment type and intensity, whereas the present study used a measure of maltreatment type only and did not measure intensity. In sum, the bivariate correlations estimated for hypothesis-3 of the present study showed a plausible pattern of associations.

Mixed effects modeling was used to identify a multivariable model that best described the data collected from the heterogeneous group of caregivers who participated in this study. More specifically, a random intercept model was chosen in order to account for the fact that some caregivers were from the same families and shared relationships with the same child or children. This type of statistical model was necessary and preferable to ordinary least squares regression given that caregivers who were from the same family tended to show a different pattern of associations for study variables compared to caregivers who were not from the same family. For instance, two related caregivers tended to show higher stress scores than any two unrelated caregivers. These procedures yielded a model that showed statistically significant, inverse associations for aggression and stress with desirable responding. That is, as PAI AGG

and STR increased, CAP fake-good responding decreased. Given the continuous nature of the AGG and STR scales and parameters for interpretation given by Morey (2007), the opposite relation also holds: as PAI AGG and STR decreased, CAP fake-good responding increased. Depression, social support, type of maltreatment, and caregiver relationship to the child did not show statistical significance in multivariable analyses.

As previously discussed, maltreatment type may have been an insignificant factor due to the manner in which it was measured—according to type versus intensity (Bennett et al., 2006). Caregiver relationship as related to desirable responding has not been studied previously. The null finding from the present data suggests that biological relationship to a child has little to do with desirable responding. In certain respects, this is a plausible result. For instance, the legal consequences to the parent indicated for child maltreatment are the same regardless of his or her blood relation to the child. In other respects, this is a surprising result. For instance, it can be speculated that the biological parent indicated for child maltreatment might experience greater subjective feelings of guilt or shame compared to nonbiological parents who may have varying degrees of investment in the child, such as a stepparent who only recently became a part of the child's life.

The findings that depression and lack of social support were not significantly associated with desirable responding are interpreted from two vantage points. The first interpretation given is with regard to examination of the mean PAI scale (i.e., DEP, NON) scores that were used to measure depression and lack of social support. DEP and NON were in the average range, meaning that caregivers in the present sample did not report high levels of depression and nonsupport (see Table 2). The DEP scale measures diagnostic features and severity of depression (Morey, 2007). Average scores for DEP indicate few to no problems and low

likelihood that depression is interfering with social functioning. A similar interpretation is given for NON: Caregivers from this sample were generally satisfied with the availability and quality of social support from acquaintances, friends, and family members. The second interpretation given is with regard to examination of multivariable models tested in previous investigations. For example, Coohey (1995) reported complex effects for social support that varied as a function of the quality of the support offered and the specific person providing the support. For depression, a number of studies have identified variables with more proximal or direct effects on caregivers' risk for child maltreatment (e.g., Crouch & Behl, 2001; Hien et al., 2010; Rodriguez & Green, 1997; Shay & Knutson, 2008). It should be noted, however, that the statistical models tested in these studies posited different dependent variables than the one tested in the present study; namely, child abuse potential (i.e., CAP Abuse scale) versus desirable responding (i.e., CAP L scale).

The two variables with statistically significant associations with desirable responding in the multivariable model were aggression and stress. In the context of the PAI, Morey (2007) developed AGG and STR as "treatment consideration scales" that do not relate directly to specific criteria for psychiatric diagnoses. These scales identify environmental factors and associated behaviors likely to affect interpersonal functioning, especially as related to psychological treatment or assessment. AGG is a measure of aggressive attitudes, verbal aggression, and physical aggression. Low scores ( $< 40T$ ) describe "meek and unassertive" persons; average scores ( $< 60T$ ) describe persons who show "reasonable control over the expression of anger and hostility;" and higher scores ( $> 70T$ ) describe persons who are "chronically angry" and who show "potential for aggression" (Morey, p. 45). Sample items include: "Sometimes my temper explodes and I completely lose control," "People would be

surprised if I yelled at someone,” and “My temper never gets me into trouble.” STR is a measure of “predictability, organization, and structure of the person’s surroundings” (p. 250). It is a proximal measure of stress in the sense that it evaluates *current* or *recent* stressors. Low scores ( $< 60T$ ) describe persons who perceive their lives to be “stable, predictable, and uneventful;” elevated scores ( $\geq 70T$ ) describe persons experiencing stressors that are frequent sources of “rumination, worry, and unhappiness;” and very high scores ( $\geq 85T$ ) describe persons “surrounded by crises” (p. 46). Sample items include: “There isn’t much stability at home” and “Things are not going well in my family.”

It is not surprising that stress and aggression are relevant to the study of caregivers undergoing evaluation for suspected or indicated child maltreatment. A number of studies have identified anger and stress as risk factors for child maltreatment. For example, Crouch and Behl (2001) reported on the interaction of parenting stress with beliefs in the acceptability of corporal punishment as related to increased child abuse potential. Rodriguez and Green (1997) found that the combination of anger expression and parenting stress was associated with child abuse potential. Guterman et al. (2009) showed that parenting stress had direct effects on caregivers’ risk for abuse and neglect, whereas more distal stressors such as impoverished and violent neighborhoods had weaker, indirect effects on caregivers’ risk for abuse and neglect by way of parenting stress. Caregiver depression is known to impact risk for maltreatment; however, its effects tend to be mediated by way of aggression (Hien et al., 2010; Shay & Knutson, 2008). Shay and Knutson and Hien and colleagues concluded that maternal depression may function to lower the threshold for angry, coercive parenting. Collectively, these findings show that anger and parenting stress are proximal factors with direct bearing on the risk of highly problematic parent-child interactions.

The question remains, however, as to why these same factors were found to be associated with desirable responding in the present study. One possibility is that anger and stress are ubiquitous and experienced by all people to some extent, and the behavior that defines the coping response can be deemed maladaptive or adaptive only in relation to the context in which it occurs. The context for the present study was forensic evaluation of parenting capacity (versus parent-child interactions as in the studies reviewed above). Desirable responding is a likely outcome of the high-stakes evaluative context, and is, perhaps, an adaptive response. For instance, it is reasonable to expect that a caregiver facing legal charges, removal of a child from his or her home, or parenting remediation classes might withhold or minimize his or her problems. However, this tendency appears to vary as a function of stress and anger. It is hypothesized that at lower levels of anger, hostility, and distress; persons undergoing parenting capacity evaluations perceive that little benefit can be obtained from a straightforward presentation. Such persons may perceive that they have the coping resources and self-restraint to manage their own affairs. At higher levels of anger, hostility, and distress; persons undergoing parenting capacity evaluations (a) may find it more difficult to manage their presentation and/or (b) may find it necessary to reach out to others for assistance. These interpretations also need to be considered in terms of the proximity of stressors and anger-provoking stimuli. For instance, persons undergoing parenting capacity evaluations may have high levels of stress and poor anger coping skills independent of the evaluation. Alternatively, the evaluation itself may occasion stress and poor anger coping, or, perhaps, there is an additive or multiplicative effect for anger and stress stemming from the evaluation *and* other contexts. More targeted investigation is required in order to differentiate the contexts within and the processes through which anger and stress achieve their effects on desirable responding.

**Implications for Parenting Capacity Evaluations**

Findings from the present study have several implications for professionals who conduct forensic evaluations of parenting capacity. First, the high number of invalid CAP profiles due to fake-good responding supports the contention that caregivers tend to perceive the evaluation context as coercive (Budd, 2001). To reduce negative effects of this perception, it is essential that evaluators take the time to establish rapport with caregivers. Practically speaking, this can be accomplished by way of informed consent and adopting a comfortable pace for evaluations.

Forensic informed consent is different from clinical informed consent due to the fact that the evaluator's client is the agency making the referral and, ultimately, the court (Budd et al., 2011). Thus, in stark contrast to the typical psychologist-client relationship, there is virtually no limit to the information that will be communicated to third parties. This must be clear to the caregiver at the outset of the evaluation, but with several caveats. For example, limitations of confidentiality can be communicated within the context of the specific referral question. Some caregivers are unclear as to why they are being evaluated by a psychologist. Many caregivers immediately and often incorrectly assume that the purpose of the evaluation is "to take my kids away from me." Thus, informing the caregiver that the purpose of the evaluation is (a) to determine whether or not he or she would benefit from a targeted parenting intervention versus (b) to gather information to inform a termination of parental rights hearing are two distinct evaluation scenarios that are likely to differentially affect caregiver behavior during the evaluation.

When the stakes of the evaluation are high as in scenario-b, the pacing of the evaluation may be especially important. The evaluation should be conducted over multiple sessions to reduce fatigue and to establish as much trust as possible. As a best-practice guideline, multiple

sources of data should be obtained, including—whenever it is not contraindicated—direct observation of the parent interacting with the child (APA, 1999; Azar et al., 1998; Budd, 2001; Budd et al., 2011). A number of caregivers involved in the present study verbalized relief and increased trust of the forensic evaluators when parent-child interactional assessments were conducted in addition to the standard procedures of testing and interview.

Related to the need to improve rapport, a second implication of the present study is that caregivers experiencing low levels of stress and anger tend to engage in desirable responding. The mechanisms responsible for this association are not yet clear; thus, it is not advisable to correct CAP L scale scores using the regression coefficients for AGG and STR reported in Table 9. However, the general inverse association of aggression and stress with desirable responding can be used to improve rapport with caregivers who report low stress and good anger management. For instance, caregivers who fit this profile may respond well to a strength-based approach to evaluation. Clearly, evaluators should pursue a balanced profile of strengths and weakness for all caregiver; however, the breadth and depth of focus may be increased to improve rapport with those caregivers most susceptible to desirable responding. According to data from the present study, caregivers who are hostile, aggressive, and stressed show lower levels of desirable responding consistent with a presentation characterized by “what you see is what you get.”

A third implication of the present study is with regard to the ROC characteristics of the present sample. The values for sensitivity and specificity of the CAP L scale reported in Table 6 may provide a helpful reference tool for evaluators who use the CAP. The values for sensitivity and specificity show a number of tradeoffs that may be useful for answering different types of referral questions. For instance, given a referral question to establish type and level of parenting

intervention for a caregiver suspected of physical neglect; the evaluator might tradeoff sensitivity for specificity to increase interpretable data from the CAP Abuse scale and factors. On the other hand, given a referral question to establish disposition; the evaluator would likely tradeoff specificity for sensitivity to increase confidence in the validity of the CAP clinical data. Thus, Table 6 and other empirical findings from this study add to the scientific merit of the CAP and further demonstrate its utility in light of *Daubert* criteria for testimonial admissibility in legal proceedings (Yanez & Fremouw, 2004). However, evaluators should exercise caution in using findings from the present study when the caregivers they evaluate differ from the present sample on relevant demographic characteristics.

A fourth and final implication is that the revised 14-item CAP L scale showed few to no empirical advantages to support its use over the original 18-item CAP L scale. Item analyses pointed up a number of potential problems with the 18-item scale such as whether or not its error variance is sufficiently small for forensic decision making. Also, a number of L scale items showed poor item-total correlations, and the dimensionality of the L scale remains unknown. Despite these concerns, the present attempt to identify a more reliable, homogeneous scale with very high internal consistency and improved discriminant validity was not achieved. Thus, the 18-item scale should be retained until more comprehensive and empirically convincing revisions are available.

### **Limitations and Future Research**

The findings and conclusions of the current study were affected by several limitations. First, the sample of caregivers studied in the present investigation was less racially diverse compared to other studies (e.g., Ondersma et al., 2005) and compared to the CAP normative sample (Milner, 1986). This limitation is of practical significance and, thus, can be minimized or

limited as long as practitioners who apply findings from this study ensure that the caregivers they evaluate are comparable with regard to relevant demographic factors.

Second and also with regard to sample characteristics, the present study did not measure reading achievement and there were missing data for education attainment. Due to the amount of missing data, education attainment was not used in any of the study analyses including ROC curve analyses to study the discriminant validity of the CAP L scale at different cutoff points. Milner (1986) recommended using a CAP L scale cutoff score of seven when caregivers have greater than a 12<sup>th</sup> grade education and a CAP L scale cutoff score of eight when caregivers have less than a 12<sup>th</sup> grade education. This recommendation, however, was complicated by several factors: (a) Caregivers in the present study had higher educational attainment than caregivers in the CAP normative sample, but higher mean CAP L scale scores than caregivers in the CAP normative sample; and (b) results reported by Budd and colleagues (2000) showed that reading ability was a more sensitive measure than educational attainment as related to desirable or fake-good responding. The broad pattern of results for education and desirable responding across the present study, the CAP normative sample, and the study conducted by Budd and colleagues indicate a more pervasive concern that is not isolated to the present study. Specifically, the education- and achievement-related correlates of desirable responding on the CAP L scale are poorly understood.

Third and also with regard to sample characteristics, the present study was conducted with a sample of caregivers who had varying levels of legal and CPS involvement and, thus, varying degrees of certainty surrounding allegations of child maltreatment. Not all caregivers were founded perpetrators of child maltreatment. That is, not all cases involved judicial adjudication of child maltreatment. Furthermore, multiple types of maltreatment were present in

the sample. These limitations are of concern when comparisons are made with the CAP normative sample, which included only caregivers with physical maltreatment histories. However, comparison with other studies (e.g., Carr et al., 2005; Ondersma et al., 2005) and with common parameters of clinical and forensic practice suggests that the CAP is often used with samples that are heterogeneous for type of maltreatment and level of involvement with the judicial system and CPS agencies. That is, few practitioners or researchers are interested solely in physical child maltreatment. Thus, heterogeneity of this variety is not necessarily problematic as it provides new evidence with regard to how the CAP L scale can be expected to perform under such conditions. Again, practitioners are cautioned to attend to these features of the present sample before applying any of the findings.

Fourth, the criterion measure of desirable responding was the PAI PIM scale, which has its own psychometric limitations, and thus it could be argued that the PIM scale is not an adequate gold standard for desirable responding. Studies on the psychometric properties of the PIM scale help to mitigate this problem. For example, the PIM scale has known rates of discriminative validity: PIM scores of *57T* corresponded to sensitivity of .933 and specificity of .778 (Morey & Lanier, 1998). Furthermore, the PIM scale shows a pattern of strong (.60 to .75) convergent correlations with other well-validated and reliable measures of desirable responding (Carr et al., 2005; Peebles & Moore, 1998). As an alternative to the PIM scale, a simulation study could have been conducted with random assignment of caregivers to fake good, standard instruction, and no instruction response sets; or a within-subjects design with counterbalanced instruction sets could have been employed. However, simulation procedures also are subject to limitations and there are certain practical and ethical concerns with regard to their application in child maltreatment settings. For example, desirable responding may be so high as to produce

little variability across groups or instruction sets, and simulation procedures would reduce the forensic utility of the evaluation.

Clearly, further research on the psychometric properties of the CAP L scale is needed to resolve limitations with the present study and to replicate and extend findings. Four areas for future investigation are identified. These include examination of the effects of the evaluation context in relation to the broader psychosocial contexts of the caregiver, examination of the extent to which desirable responding is situation-specific, examination of the link between the CAP L scale and the CAP Abuse scale, and examination of the psychometric properties of the CAP L scale.

With regard to contextual factors, results of the present study showed that family, financial, work, and relationship stress was associated with desirable responding. These stressors likely affect stress that is specific to CPS involvement as well as stress that is specific to the parenting capacity evaluation itself. The manner in which these multiple levels of stress converge on desirable responding within the parenting capacity evaluation is poorly understood. It is reasonable to propose, however, that the most proximal stressors would show the largest effects on desirable responding (e.g., Belsky, 1993; Bronfenbrenner & Morris, 2006). Drawing from ecological theories, it is hypothesized that these concentric levels of stress would be multiplicative. A relevant question is whether broader psychosocial stress predisposes caregivers to experience more or less stress during the parenting capacity evaluation. The present data suggest that broader experiences of stress actually reduce desirable responding, suggesting that the evaluation itself may be perceived as less stressful and, perhaps, as an opportunity to “come clean” or to receive assistance.

The extent to which desirable responding is situation-specific versus pervasive is of limited utility in improving the psychometric properties of the CAP L scale. However, this information could prove valuable in improving rapport and obtaining greater percentages of valid test profiles. Simple investigation of situation X maltreatment status could provide information as to whether caregivers with and without maltreatment histories show differences in situation-specific desirable responding. Further studies might be conducted to identify procedures to reduce defensiveness and desirable responding. Study manipulations might involve providing more comprehensive discussions of informed consent or psychological and legal education about CPS involvement.

At present, the link between elevated CAP L scale scores and low CAP Abuse scale scores is inferred. There currently is no quantitative basis for an association between high L scale scores and low Abuse scale scores. Thus, it is possible that a respondent could produce a high L scale score and a low Abuse scale score and truly have a low risk for child maltreatment. Likewise it is possible that a respondent could produce a low L scale score and a low Abuse scale score and truly have a high risk for child maltreatment. These possibilities highlight the importance of conducting further investigation of the correlates of desirable responding in parenting capacity evaluations. Identification of correlates that also show associations with risk for child maltreatment may be candidates for tests of mediation. Models of mediated effects and models of moderated mediation provide reasonable methods of detecting and describing the processes and conditions whereby desirable responding (i.e., L scale) affects risk (i.e., Abuse scale).

Finally, and relevant to each of the areas for future investigation previously discussed, the CAP L scale may benefit from refinement to enhance its psychometric properties. This was the

primary aim of the present investigation and though valuable information regarding homogeneity of scale content and sensitivity and specificity to differentiate desirable versus honest responding was obtained; no improvements were made to the L scale. For most purposes, the CAP L scale is considered a very good measure, with internal consistency greater than .80 and strong evidence for discriminant validity. Yet, findings from the present study point up areas for improvement: Five L scale items showed problems that indicate heterogeneity of scale content, internal consistency is not as high as that recommended for high-stakes testing, and specificity is lacking. With regard to heterogeneity of content, more thoroughgoing analysis of construct validity is warranted, including multitrait-multimethod evaluation of convergent and divergent associations and factor analytic studies. Concurrent with these endeavors, the L scale may require content revisions; for example, rewriting problem items and writing new items to ensure cultural and contextual relevance.

## References

- American Psychological Association Committee on Professional Practice and Standards (1999). Guidelines for psychological evaluations in child protection matters. *American Psychologist*, 54, 586-593.
- Azar, S. T., & Cote, L. R. (2002). Sociocultural issues in the evaluation of the needs of children in custody decision making: What do our current frameworks for evaluating parenting practices have to offer? *International Journal of Law and Psychiatry*, 25, 193-217.
- Azar, S. T., Lauretti, A. F., & Loding, B. V. (1998). The evaluation of parental fitness in termination of parental rights cases: A functional-contextual perspective. *Clinical Child and Family Psychology Review*, 1, 77-100.
- Baer, R. A., & Wetter, M. W. (1997). Effects of information about validity scales on underreporting of symptoms on the Personality Assessment Inventory. *Journal of Personality Assessment*, 68, 402-413.
- Barnett, D., Manly, J. T., & Cicchetti, D. (1993). Defining child maltreatment: The interface between policy and research. In D. Cicchetti & S. L. Toth (Eds.), *Child abuse, child development, and social policy* (pp. 7-73). Norwood, NJ: Ablex.
- Barnum, R. (1997). A suggested framework for forensic consultation cases of child abuse and neglect. *Journal of the American Academy of Psychiatry and Law*, 25, 581-593.
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Pearson Assessments.
- Belsky, J. (1993). Etiology of child maltreatment: A developmental-ecological analysis. *Psychological Bulletin*, 114, 413-434.

- Benjet, C., Azar, S. T., & Kuersten-Hogan, R. (2003). Evaluating the parental fitness of psychiatrically diagnosed individuals: Advocating a functional-contextual analysis of parenting. *Journal of Family Psychology, 17*, 238-251.
- Bennett, D. S., Sullivan, M. W., & Lewis, M. (2006). Relations of parental report and observation of parenting to maltreatment history. *Child Maltreatment, 11*, 63-75. doi: 10.1177/1077559505283589.
- Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner & W. Damon (Eds.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (6th ed., pp. 793 – 828). Hoboken, NJ: Wiley.
- Budd, K. S. (2001). Assessing parenting competence in child protection cases: A clinical practice model. *Clinical Child and Family Psychology Review, 4*, 1-18.
- Budd, K. S. (2005). Assessing parenting capacity in a child welfare context. *Children and Youth Services Review, 27*, 429-444.
- Budd, K. S., Connell, M., & Clark, J. R. (2011). *Evaluation of parenting capacity in child protection*. New York: Oxford.
- Budd, K. S., Heilman, N. E., & Kane, D. (2000). Psychosocial correlates of child abuse potential in multiply disadvantaged adolescent mothers. *Child Abuse and Neglect, 24*, 611-625.
- Budd, K. S., & Holdsworth, M. J. (1996). Issues in clinical assessment of minimal parenting competence. *Journal of Clinical Child Psychology, 25*, 2-14.
- Budd, K. S., Poindexter, L. M., Felix, E. D., & Naik-Polan (2001). Clinical assessment of parents in child protection cases: An empirical analysis. *Law and Human Behavior, 25*, 93-108.
- Cashel, M. L., Rogers, R., Sewell, K., & Martin-Cannici, C. (1995). The Personality Assessment Inventory (PAI) and the detection of defensiveness. *Assessment, 2*, 333-342.

- Chaffin, M., & Valle, L. A. (2003). Dynamic prediction characteristics of the Child Abuse Potential Inventory. *Child Abuse & Neglect*, 27, 463-481.
- Chambers, A. L., & Wilson, M. N. (2007). Assessing male batterers with the Personality Assessment Inventory. *Journal of Personality Assessment*, 88, 58-66.
- Cicchetti, D., & Lynch, M. (1993). Toward an ecological/transactional model of community violence and child maltreatment: Consequences for children's development. *Psychiatry*, 56, 96-118.
- Cohen, L. R., Hien, D. A., & Barchelder, S. (2008). The impact of cumulative maternal trauma and diagnosis on parenting behavior. *Child Maltreatment*, 13, 27-38.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.
- Coohey, C. (1995). Neglectful mothers, their mothers, and partners: The significance of mutual aid. *Child Abuse and Neglect*, 19, 885-895.
- Crawford, E. F., Calhoun, P. S., Braxton, L. E., & Beckham, J. C. (2007). Validity of the Personality Assessment Inventory Aggression scales and Violence Potential index in veterans with PTSD. *Journal of Personality Assessment*, 88, 91-99.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Crouch, J. L., & Behl, L. E. (2001). Relationships among parental beliefs in corporal punishment, reported stress, and physical child abuse potential. *Child Abuse and Neglect*, 25, 413-49.

- Crouch, J. L., Milner, J. S., & Thomsen, C. (2001). Childhood physical abuse, early social support, and risk for maltreatment: Current social support as a mediator of risk for child physical abuse. *Child Abuse and Neglect*, 25, 93-107.
- de Paúl, J., Asla, N., Pérez-Albénez, A., & Torres-Gómez de Cádiz, B. (2006). Impact of stress and mitigating information on evaluations, attributions, affect, disciplinary choices, and expectations of compliance in mothers at high and low risk for child physical abuse. *Journal of Interpersonal Violence*, 21, 1018-1045.
- Diareme, s., Tsiantis, J., & Tsitoura, S. (1997). Cross-cultural validation of the Child Abuse Potential Inventory in Greece: A preliminary study. *Child Abuse and Neglect*, 21, 1067-1079.
- Douglas, K. S., Hart, S. D., & Kropp, P. R. (2001). Validity of the Personality Assessment Inventory for forensic assessments. *International Journal of Offender Therapy and Comparative Criminology*, 45, 183-197.
- Edens, J. F., Cruise, K. R., & Buffington-Vollum, J. K. (2001). Forensic and correctional applications of the Personality Assessment Inventory. *Behavioral Sciences and the Law*, 19, 519-543.
- Fals-Stewart, W. (1996). The ability of individuals with psychoactive substance use disorders to escape detection by the Personality Assessment Inventory. *Psychological Assessment*, 8, 60-68.
- Fals-Stewart, W., & Lucente, S. (1997). Identifying positive dissimulation by substance-abusing individuals on the Personality Assessment Inventory: A cross-validation study. *Journal of Personality Assessment*, 68, 455-469.

- Guterman, N. B., Lee, Shawna, S. J., Taylor, C. A., & Rathouz, P. J. (2009). Parental perceptions of neighborhood processes, stress, personal control, and risk for physical child abuse and neglect. *Child Abuse and Neglect*, 33, 897-906.
- Haynes, S. N., & O'Brien, W. H. (2000). *Principles and practice of behavioral assessment*. Plenum: New York.
- Haz, A. M., & Ramirez, V. (1998). Preliminary validation of the Child Abuse Potential Inventory in Chile. *Child Abuse and Neglect*, 22, 869-879.
- Hien, D., Cohen, L. R., Caldeira, N. A., Flom, P., & Wasserman, G. (2010). Depression and anger as risk factors underlying the relationship between maternal substance involvement and child abuse potential. *Child Abuse and Neglect*, 34, 105-113.
- Jacobo, M. C., Blais, M. A., Baity, M. R., & Harley, R. (2007). Concurrent validity of the Personality Assessment Inventory Borderline scales in patients seeking dialectical behavioral therapy. *Journal of Personality Assessment*, 88, 75-81.
- Jellinek, M. S., Murphy, J. M., Poitras, F., Quinn, D., Bishop, S. J., & Goshko, M. (1992). Serious child mistreatment in Massachusetts: The course of 206 children through the courts. *Child Abuse and Neglect*, 16, 179-185.
- Krug, E. G., Dahlberg, L. L., Mercy, J. A., Zwi, A. B., & Lozano, R. (Eds.). (2002) World report on violence and health. Geneva: World Health Organization. Retrieved June 23, 2010, from [http://whqlibdoc.who.int/publications/2002/9241545615\\_eng.pdf](http://whqlibdoc.who.int/publications/2002/9241545615_eng.pdf)
- Kurtz, J. E., & Blais, M. A. (2007). Introduction to the special issue on the Personality Assessment Inventory. *Journal of Personality Assessment*, 88, 1-4.

- Kurtz, J. E., Shealy, S. E., & Putnam, S. H. (2007). Another look at paradoxical severity effects in head injury with the Personality Assessment Inventory. *Journal of Personality Assessment, 88*, 67-74.
- Lyons, S. J., Henly, J. R., & Schuerman, J. R. (2005). Informal support in maltreating families: Its effect on parenting practices. *Children and Youth Services Review, 27*, 21-38.
- Mammen, O. K., Kolko, D. J., & Pilkonis, (2002). P. A. Negative affect and parental aggression in child physical abuse. *Child Abuse and Neglect, 26*, 407-424.
- McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*, 215-241.
- Milner, J. S. (1982). Development of a lie scale for the Child Abuse Potential Inventory. *Psychological Reports, 50*, 871-874.
- Milner, J. S. (1986). *The Child Abuse Potential Inventory manual* (2nd ed.). Dekalb, IL: Psytec Inc.
- Milner, J. S. (1991). Medical Conditions and Child Abuse Potential Inventory specificity. *Psychological Assessment, 3*, 208-212.
- Milner, J. S. (1994). Assessing physical child abuse risk: The Child Abuse Potential Inventory. *Clinical Psychology Review, 14*, 547-583.
- Milner, J. S., Gold, R. G., Ayoub, C., & Jacewitz, M. M. (1984). Predictive validity of the Child Abuse Potential Inventory. *Journal of Consulting and Clinical Psychology, 52*, 879-884.
- Milner, J. S., & Robertson, K. R. (1985). Development of a random response scale for the Child Abuse Potential Inventory. *Journal of Clinical Psychology, 41*, 639-643.
- Milner, J. S., & Wimberly, R. C. (1980). Prediction and explanation of child abuse. *Journal of Clinical Psychology, 36*, 875-884.

- Montes, M. P., de Paúl, J., & Milner, J. S. (2001). Evaluations, attributions, affect, and disciplinary choices in mothers at high and low risk for child physical abuse. *Child Abuse and Neglect*, 25, 1015-1036.
- Morey, L. C. (2007). *Personality Assessment Inventory professional manual* (2nd ed.). Lutz, FL: PAR Inc.
- Morey, L. C., & Lanier, V. W. (1998). Operating characteristics of six response distortion indicators for the Personality Assessment Inventory. *Assessment*, 5, 203-214.
- Mullen, K. L., & Edens, J. F. (2008). A case law summary of the Personality Assessment Inventory: Examining its role in civil and criminal trials. *Journal of Personality Assessment*, 90, 300-303.
- Novaco, R. W. (2003). *The Novaco Anger Scale and Provocation Inventory Manual*. Torrance, CA: Western Psychological Services.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3<sup>rd</sup> ed.). New York: McGraw-Hill.
- Ondersma, S. J., Chaffin, M. J., Mullins, S. M., & LeBreton, J. M. (2005). A brief form of the Child Abuse Potential Inventory: Development and Validation. *Journal of Clinical Child and Adolescent Psychology*, 34, 301-311.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598-609.
- Paulhus, D. L. (1998). *Paulhus Scales User's Manual*. North Tonawanda, NY: Multi-Health Systems Inc.

- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Brown, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Erlbaum.
- Paulhus, D. L. (2003). Self-presentation measurement. In R. Fernandez-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 858-861). Thousand Oaks, CA: Sage.
- Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robbins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224-239). New York: Guilford.
- Peebles, J., & Morre, R. J. (1998). Detecting socially desirable responding with the Personality Assessment Inventory: The Positive Impression Management scale and the Defensiveness Index. *Journal of Clinical Psychology*, 54, 621-628.
- Pečnik, N. & Ajduković, M. (1995). The Child Abuse Potential Inventory: Cross validation in Croatia. *Psychological Reports*, 76, 979-985.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & the R Core Team (2008). nlme: Linear and nonlinear mixed effects models. R package version 3.1-103.
- Pintea, S., & Moldovan, R. (2009). The receiver-operating characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychotherapies*, 9, 49-66.
- Piotrowski, C. (2000). How popular is the Personality Assessment Inventory in practice and training? *Psychological Reports*, 86, 65-66.
- R Development Core Team (2008). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>

Rinehart, D. J., Becker, M. A., Buckley, P. R., Dailey, K., Reichardt, C. S., Graeber, C., . . .

Brown, E. (2005). The relationship between mothers' child abuse potential and current mental health symptoms. *Journal of Behavioral Health Services & Research*, 32, 155-166.

Robertson, K. R., & Milner, J. S. (1983). Construct validity of the Child Abuse Potential Inventory. *Journal of Clinical Psychology*, 39, 426-429.

Robertson, K. R., & Milner, J. S. (1985). Convergent and discriminant validity of the Child Abuse Potential Inventory. *Journal of Personality Assessment*, 49, 86-88.

Rodriguez, C. M., & Green, A. J. (1997). Parenting stress and anger expression as predictors of child abuse potential. *Child Abuse and Neglect*, 21, 367-377.

Sarkar, D. (2008). Lattice: Lattice graphics. R package version 0.20-0.

Sedlak, A. J., & Broadhurst, D. D. (1996). *Third National Incidence Study on child abuse and neglect*. Washington, DC: U.S. Department of Health and Human Services.

Shay, N. L., & Knutson, J. F. (2008). Maternal depression and trait anger as risk factors for escalated physical discipline. *Child Maltreatment*, 13, 39-49.

Skopp, N. A., Edens, J. F., & Ruiz, M. A. (2007). Risk factors for institutional misconduct among incarcerated women: An examination of the criterion-related validity of the Personality Assessment Inventory. *Journal of Personality Assessment*, 88, 107-118.

Spielberger, C. D. (1999). *State-Trait Anger Expression Inventory-2 test manual*. Port Huron, MI: Sigma Assessment Systems.

Stein, M. B., Pinsker-Aspen, J. H., & Hilsenroth, M. J. (2007). Borderline pathology and the Personality Assessment Inventory (PAI): An evaluation of criterion and concurrent validity. *Journal of Personality Assessment*, 88, 82-90.

- Straus, M. A. (1979). Measuring interfamily conflict and violence: The Conflict Tactics Scale (CTS). *Journal of Marriage and the Family*, 41, 75-88.
- Stredney, R. V., Archer, R. P., & Mason, J. A. (2006). MMPI-2 and MCMI-III characteristics of parental competency examinees. *Journal of Personality Assessment*, 87, 113-115.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York: Pearson.
- U.S. Department of Health and Human Services Administration on Children, Youth, and Families (2010). *Child maltreatment 2009*. Washington, D.C.: U.S. Government Printing Office.
- U.S. Census Bureau (2012). State and county quickfacts. [Data file]. Retrieved from <http://quickfacts.census.gov/qfd/states/>
- Walker, C. A., & Davies, J. (2010). A critical review of the psychometric evidence base of the Child Abuse Potential Inventory. *Journal of Family Violence*, 25, 215-227.
- Wang, C., & Holton, J. (2007). Total estimated cost of child abuse and neglect in the United States. Chicago, IL: Prevent Child Abuse America. Retrieved June 28, 2010, from [http://www.preventchildabuse.org/about\\_us/media\\_releases/pcaa\\_pew\\_economic\\_impact\\_study\\_final.pdf](http://www.preventchildabuse.org/about_us/media_releases/pcaa_pew_economic_impact_study_final.pdf)
- Wu, S. S., Ma, C., Carter, R. L., Ariet, M., Feaver, E. A., Resnick, M. B., & Roth, J. (2004). Risk factors for infant maltreatment: a population-based study. *Child Abuse and Neglect*, 28, 1253-1264.
- Yanez, Y. T., & Fremouw, W. (2004). The application of the *Daubert* standard to parental capacity measures. *American Journal of Forensic Psychology*, 22, 5-29.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York: Springer.

## Footnotes

<sup>1</sup>NCANDS is a voluntary system, and data from non-CPS sources indicate that these numbers may underestimate the prevalence of child maltreatment. For example, Finkelhor, Turner, Ormrod, and Hamby (2009) estimated one-year and lifetime prevalence rates of maltreatment from a nationally representative sample of children. More than 1 of 10 (10.2%) children reported past year maltreatment, and nearly 1 of 5 (18.6%) children reported lifetime maltreatment.

<sup>2</sup>Note that using the criteria of (a) corrected item-total correlations  $> .30$  and (b) improvement to internal consistency if a given item were deleted, item 5 and item 15 would have been deleted as both items yielded insufficient values for each metric. Item 5 and item 15 were retained, however, in order to maintain the balance of items keyed Agree-Disagree. This was necessary to preserve the functionality of the Faking-good index. Interested readers are referred to Milner (1986, p. 34) for information about the construction of and rationale for the Faking-good index.

Table 1

*Descriptive Statistics for the Demographic Characteristics of the Full Sample (N = 125) and the Subsample (n = 95)*

| Characteristic                  | Female              | Male                | Missing Values |
|---------------------------------|---------------------|---------------------|----------------|
| Sex of caregiver                | 65 (50)             | 60 (45)             | None           |
| M caregiver age in years        | 29.52 (30.74)       | 34.00 (34.80)       | None           |
| Standard deviation              | 7.67 (7.48)         | 9.35 (9.83)         |                |
| Range                           | 18 to 53 (20 to 53) | 20 to 55 (20 to 55) |                |
| Caregiver race                  |                     |                     | None           |
| Black/African American          | 6 (4)               | 8 (6)               |                |
| Biracial                        | 2 (2)               | 2 (1)               |                |
| White/Caucasian                 | 57 (44)             | 50 (38)             |                |
| Caregiver education             |                     |                     | 18 (15)        |
| < High school diploma           | 14 (12)             | 13 (11)             |                |
| High school diploma / GED       | 23 (20)             | 24 (19)             |                |
| Some college                    | 10 (8)              | 5 (3)               |                |
| College degree                  | 3 (1)               | 6 (5)               |                |
| Graduate / Professional         | 1 (1)               | 0 (0)               |                |
| Caregiver employment status     |                     |                     | 8 (5)          |
| Employed                        | 26 (23)             | 35 (24)             |                |
| Unemployed                      | 31 (22)             | 22 (19)             |                |
| Student                         | 1 (1)               | 2 (1)               |                |
| Tests read aloud to caregiver   |                     |                     | None           |
| No                              | 64 (50)             | 58 (43)             |                |
| Yes                             | 1 (0)               | 2 (2)               |                |
| Caregiver relationship status   |                     |                     | 12 (10)        |
| Single                          | 13 (12)             | 11 (8)              |                |
| In a relationship               | 44 (31)             | 45 (34)             |                |
| Caregiver relationship to child |                     |                     | None           |
| Biological mother               | 61 (47)             | —                   |                |
| Other female parent             | 4 (3)               | —                   |                |
| Biological father               | —                   | 51 (38)             |                |
| Other male parent               | —                   | 8 (6)               |                |
| Paternity pending               | —                   | 1 (1)               |                |

*(Table 1 continues)*

*(Table 1 continued)*

| Characteristic               | Female              | Male                | Missing Values |
|------------------------------|---------------------|---------------------|----------------|
| Type of maltreatment         |                     |                     | None           |
| Physical only                | 17 (13)             | 15 (9)              |                |
| Sexual only                  | 2 (2)               | 1 (1)               |                |
| Neglect only                 | 26 (20)             | 21 (17)             |                |
| Emotional only               | 2 (1)               | 1 (1)               |                |
| Truancy / Other legal only   | 1 (1)               | 3 (2)               |                |
| More than one type of abuse  | 17 (13)             | 19 (15)             |                |
| Sex of child                 | 50 (40)             | 75 (55)             | None           |
| <i>M</i> child age in months | 62.12 (71.68)       | 53.89 (51.19)       | 4 (3)          |
| Standard deviation           | 65.59 (68.13)       | 61.43 (58.58)       |                |
| Range                        | 2 to 192 (2 to 192) | 1 to 192 (1 to 192) |                |

Table 2

*Range of Possible Values, Means, Standard Deviations, Skewness, and Kurtosis for CAP and PAI scales for the full sample (N = 125) and the subsample (n = 95).*

| Scale        | Minimum | Maximum       | <i>M</i>        | <i>SD</i>     | Skewness    | Kurtosis    |
|--------------|---------|---------------|-----------------|---------------|-------------|-------------|
| CAP validity |         |               |                 |               |             |             |
| IC           | 0 (0)   | 11 (11)       | 4.02 (4.17)     | 2.13 (2.15)   | .51 (.62)   | .40 (.52)   |
| L            | 0 (1)   | 18 (18)       | 9.19 (8.93)     | 4.00 (3.92)   | -.09 (-.03) | -.62 (-.61) |
| RR           | 0 (0)   | 5 (5)         | 1.94 (1.83)     | 1.24 (1.22)   | .39 (.44)   | -.48 (-.31) |
| CAP Abuse    | 2 (2)   | 364 (364)     | 105.96 (110.55) | 91.81 (92.78) | 1.16 (1.12) | .47 (.48)   |
| PAI validity |         |               |                 |               |             |             |
| ICN          | 34 (34) | 67 (67)       | 49.96 (49.85)   | 8.60 (8.12)   | .06 (.06)   | -.90 (-.91) |
| INF          | 40 (40) | 82 (82)       | 54.52 (54.97)   | 9.37 (9.50)   | .49 (.55)   | -.14 (.04)  |
| NIM          | 44 (44) | 88 (88)       | 50.46 (50.42)   | 8.87 (9.11)   | 1.85 (1.99) | 3.65 (4.24) |
| PIM          | 15 (22) | 77 (75)       | 55.51 (54.66)   | 12.19 (11.45) | -.72 (-.64) | .67 (.36)   |
| PAI AGG      | 32 (34) | 75 (75)       | 47.53 (48.64)   | 10.85 (11.11) | .86 (.89)   | .20 (.11)   |
| PAI DEP      | 35 (36) | 83 (83)       | 52.31 (53.46)   | 11.89 (12.26) | .82 (.79)   | .19 (-.03)  |
| PAI NON      | 37 (37) | 81.50 (81.50) | 48.93 (48.87)   | 11.02 (10.93) | 1.10 (1.24) | .76 (1.21)  |
| PAI STR      | 37 (37) | 84 (84)       | 55.00 (56.16)   | 11.66 (11.95) | .54 (.44)   | -.34 (-.51) |

*Note.* Data for the subsample ( $n = 95$ ) are shown in parentheses.

Table 3

*Percent of agree-disagree responses, corrected item-total correlations, and alpha if item deleted for the 18 Lie scale items for the full sample (N = 125) and the subsample (n = 95).*

| Item  | Correct  | Agree       | Disagree    | $r_{it}$  | $\alpha_d$  |
|---|----------|-------------|-------------|-----------|-------------|
| 1. I sometimes act without thinking               | Agree    | 28.0 (31.6) | 72.0 (68.4) | .42 (.41) | .802 (.789) |
| 2. I am always a good person                      | Disagree | 78.4 (75.8) | 21.6 (24.2) | .36 (.39) | .806 (.790) |
| 3. I never worry about my health                  | Disagree | 34.4 (34.7) | 65.6 (65.3) | .13 (.13) | .820 (.806) |
| 4. I sometimes lose my temper                     | Agree    | 42.4 (44.2) | 57.6 (55.8) | .45 (.46) | .800 (.785) |
| 5. I sometimes think of myself first              | Agree    | 24.8 (25.3) | 75.2 (74.7) | .31 (.25) | .809 (.798) |
| 6. Sometimes I have bad thoughts                  | Agree    | 20.8 (24.2) | 79.2 (75.8) | .53 (.55) | .797 (.780) |
| 7. I always do what is right                      | Disagree | 59.2 (55.8) | 40.8 (44.2) | .62 (.58) | .789 (.777) |
| 8. I sometimes fail to keep all of my promises    | Agree    | 45.6 (45.3) | 54.4 (54.7) | .51 (.54) | .797 (.779) |
| 9. I never get mad at others                      | Disagree | 14.4 (14.7) | 85.6 (85.3) | .47 (.47) | .801 (.786) |
| 10. People sometimes take advantage of me         | Agree    | 48.0 (49.5) | 52.0 (50.5) | .33 (.29) | .808 (.797) |
| 11. I never listen to gossip                      | Disagree | 53.6 (54.7) | 46.4 (45.3) | .21 (.21) | .816 (.802) |
| 12. I sometimes say bad words                     | Agree    | 81.6 (85.3) | 18.4 (14.7) | .47 (.48) | .800 (.786) |
| 13. I never do anything that is bad for my health | Disagree | 39.2 (35.8) | 60.8 (64.2) | .50 (.46) | .797 (.785) |
| 14. I am always happy with what I have            | Disagree | 68.8 (65.3) | 31.2 (34.7) | .42 (.36) | .803 (.791) |
| 15. I sometimes act silly                         | Agree    | 84.0 (82.1) | 16.0 (17.9) | .21 (.18) | .813 (.801) |
| 16. I never raise my voice in anger               | Disagree | 22.4 (20.0) | 77.6 (80.0) | .46 (.43) | .800 (.788) |
| 17. I sometimes think of myself before others     | Agree    | 39.2 (38.9) | 60.8 (61.1) | .22 (.19) | .815 (.803) |
| 18. I always tell the truth                       | Disagree | 63.2 (62.1) | 36.8 (37.9) | .58 (.53) | .792 (.780) |

*Note.* Data for the subsample ( $n = 95$ ) are shown in parentheses. Full scale alpha was .813 ( $N = 125$ ) and .800 ( $n = 95$ );  $r_{it}$  = corrected item-total correlation;  $\alpha_d$  = alpha if item deleted.

Table 4

*Estimated phi coefficients between CAP Lie scale items for the full sample (N = 125) and the subsample (n = 95).*

| Item          | 1    | 2    | 3    | 4    | 5                | 6                | 7    | 8    | 9                 | 10   | 11   | 12                | 13   | 14               | 15               | 16                | 17   | 18   |
|---------------|------|------|------|------|------------------|------------------|------|------|-------------------|------|------|-------------------|------|------------------|------------------|-------------------|------|------|
| 1. Thinking   | —    | .09  | .03  | .44* | .07              | .30*             | .31* | .43* | .22 <sup>a</sup>  | .28* | .07  | .28* <sup>a</sup> | .22  | .08              | .20              | .28*              | .08  | .17  |
| 2. Good       | .15  | —    | .05  | .14  | .12              | .25*             | .54* | .33* | .17 <sup>a</sup>  | .13  | .13  | .17 <sup>a</sup>  | .22  | .16              | .01 <sup>a</sup> | .22 <sup>a</sup>  | .00  | .42* |
| 3. Health     | .00  | .05  | —    | .02  | .02              | .10              | .07  | .18  | .07 <sup>a</sup>  | .07  | .00  | .01 <sup>a</sup>  | .24  | .07              | .05              | .02               | .08  | .34* |
| 4. Temper     | .44* | .14  | .01  | —    | .21              | .34*             | .27* | .26  | .25               | .22  | .09  | .31*              | .22  | .20              | .19              | .29*              | .20  | .14  |
| 5. Self first | .10  | .10  | .01  | .22  | —                | .35*             | .17  | .15  | .04 <sup>a</sup>  | .10  | .04  | .03 <sup>a</sup>  | .02  | .09              | .08 <sup>a</sup> | .01 <sup>a</sup>  | .53* | .20  |
| 6. Thoughts   | .30* | .26  | .12  | .36* | .36*             | —                | .39* | .33* | .24 <sup>a</sup>  | .28* | .13  | .24 <sup>a</sup>  | .22  | .31*             | .07 <sup>a</sup> | .22 <sup>a</sup>  | .25  | .37* |
| 7. Right      | .32* | .47* | .05  | .31* | .28*             | .42*             | —    | .43* | .19               | .18  | .26* | .19               | .36* | .29*             | .19              | .18               | .07  | .49* |
| 8. Promises   | .36* | .22  | .12  | .22  | .22              | .32*             | .42* | —    | .26*              | .24  | .19  | .32*              | .33* | .14              | .09              | .30*              | .01  | .38* |
| 9. Mad        | .21  | .16  | .04  | .26* | .08 <sup>a</sup> | .21 <sup>a</sup> | .20  | .28* | —                 | .11  | .26* | .58* <sup>a</sup> | .31* | .24 <sup>a</sup> | .12 <sup>a</sup> | .61* <sup>a</sup> | .03  | .26* |
| 10. Advantage | .29* | .12  | .01  | .18  | .12              | .26*             | .21  | .25* | .17               | —    | .03  | .17               | .08  | .12              | .13              | .18               | .07  | .18  |
| 11. Gossip    | .01  | .14  | .00  | .05  | .01              | .12              | .27* | .18  | .20               | .07  | —    | .20               | .28* | .09              | .04              | .19               | .12  | .03  |
| 12. Bad words | .30* | .15  | .05  | .32* | .03              | .19 <sup>a</sup> | .23* | .27* | .57* <sup>a</sup> | .21  | .15  | —                 | .37* | .24 <sup>a</sup> | .12 <sup>a</sup> | .61 <sup>a</sup>  | .03  | .20  |
| 13. Health do | .21  | .22  | .21  | .23* | .04              | .25*             | .40* | .34* | .32*              | .12  | .29* | .34*              | —    | .31*             | .17              | .29*              | .03  | .22  |
| 14. Happy     | .16  | .19  | .09  | .23* | .17              | .34*             | .35* | .15  | .23*              | .15  | .10  | .19               | .33* | —                | .01              | .20               | .23  | .25  |
| 15. Silly     | .18  | .02  | .01  | .20  | .01 <sup>a</sup> | .06 <sup>a</sup> | .19  | .14  | .19 <sup>a</sup>  | .16  | .01  | .13 <sup>a</sup>  | .19  | .01              | —                | .04 <sup>a</sup>  | .04  | .08  |
| 16. Voice     | .25* | .24* | .06  | .31* | .04              | .23*             | .25* | .22  | .55* <sup>a</sup> | .21  | .15  | .54*              | .36* | .24*             | .08 <sup>a</sup> | —                 | .09  | .17  |
| 17. Self      | .01  | .02  | .03  | .17  | .49*             | .19              | .13  | .09  | .10               | .11  | .09  | .13               | .01  | .20              | .04              | .04               | —    | .36* |
| 18. Truth     | .19  | .37* | .31* | .18  | .25*             | .39*             | .55* | .40* | .27*              | .23* | .09  | .19               | .31* | .35*             | .11              | .25*              | .31* | —    |

*Note.* Correlations for full sample (N = 125) below diagonal; correlations for subsample (n = 95) above diagonal. Pearson's chi-square for statistical significance.

Coefficients superscripted "a" lacked expected cell frequencies > 5. Each participant contributed data to only one cell of the contingency table.

\* $p < .01$

Table 5

*Corrected item-total correlations and alpha values for the revised 14 Lie scale items.*

| Item  | N = 125  |            |          | n = 95   |            |          |
|---|----------|------------|----------|----------|------------|----------|
|   | $r_{it}$ | $\alpha_d$ |          | $r_{it}$ | $\alpha_d$ |          |
| 1. I sometimes act without thinking               | .450     | .818       |          | .468     | .804       |          |
| 2. I am always a good person                      | .383     | .822       |          | .404     | .809       |          |
| 4. I sometimes lose my temper                     | .460     | .817       |          | .454     | .805       |          |
| 5. I sometimes think of myself first              | .273     | .829       |          | —        | —          |          |
| 6. Sometimes I have bad thoughts                  | .540     | .812       |          | .510     | .801       |          |
| 7. I always do what is right                      | .623     | .805       |          | .578     | .795       |          |
| 8. I sometimes fail to keep all of my promises    | .508     | .814       |          | .544     | .798       |          |
| 9. I never get mad at others                      | .471     | .817       |          | .483     | .804       |          |
| 10. People sometimes take advantage of me         | .336     | .826       |          | .319     | .816       |          |
| 12. I sometimes say bad words                     | .473     | .817       |          | .522     | .802       |          |
| 13. I never do anything that is bad for my health | .472     | .816       |          | .458     | .805       |          |
| 14. I am always happy with what I have            | .417     | .820       |          | .347     | .813       |          |
| 15. I sometimes act silly                         | —        | —          |          | .195     | .821       |          |
| 16. I never raise my voice in anger               | .492     | .815       |          | .483     | .803       |          |
| 18. I always tell the truth                       | .546     | .811       |          | .472     | .804       |          |
|   | Mean     | SD         | $\alpha$ | Mean     | SD         | $\alpha$ |
| Revised Lie scale totals                          | 7.54     | 3.52       | .828     | 6.67     | 3.44       | .817     |

Note.  $r_{it}$  = corrected item-total correlation;  $\alpha_d$  = alpha if item deleted.

Table 6

*Identification rate estimates for selected cutoff scores on the 18-item CAP L scale and 14-item revised CAP L scale for the subsample (n = 95).*

| Scale                       | Cutoff score | Sensitivity  | Specificity  | Base rate = .25 |       | Base rate = .50 |       | Base rate = .75 |       |
|-----------------------------|--------------|--------------|--------------|-----------------|-------|-----------------|-------|-----------------|-------|
|                             |              |              |              | PPP             | NPP   | PPP             | NPP   | PPP             | NPP   |
| 18-item CAP L scale         |              |              |              |                 |       |                 |       |                 |       |
|                             | 7            | 1.000        | 0.568        | 0.436           | 1.000 | 0.698           | 1.000 | 0.874           | 1.000 |
|                             | <b>8</b>     | <b>0.961</b> | <b>0.705</b> | 0.521           | 0.982 | 0.765           | 0.948 | 0.907           | 0.858 |
|                             | 9            | 0.804        | 0.750        | 0.517           | 0.920 | 0.763           | 0.793 | 0.906           | 0.561 |
|                             | 10           | 0.765        | 0.864        | 0.652           | 0.917 | 0.849           | 0.786 | 0.944           | 0.551 |
|                             | 11           | 0.647        | 0.977        | 0.904           | 0.893 | 0.966           | 0.735 | 0.988           | 0.480 |
|                             | 12           | 0.471        | 1.000        | 1.000           | 0.850 | 1.000           | 0.654 | 1.000           | 0.387 |
| 14-item revised CAP L scale |              |              |              |                 |       |                 |       |                 |       |
|                             | 4            | 1.000        | 0.409        | 0.361           | 1.000 | 0.629           | 1.000 | 0.835           | 1.000 |
|                             | 5            | 0.980        | 0.636        | 0.473           | 0.990 | 0.729           | 0.970 | 0.890           | 0.914 |
|                             | <b>6</b>     | <b>0.882</b> | <b>0.750</b> | 0.540           | 0.950 | 0.779           | 0.864 | 0.914           | 0.679 |
|                             | 7            | 0.784        | 0.818        | 0.589           | 0.919 | 0.812           | 0.791 | 0.928           | 0.558 |
|                             | 8            | 0.725        | 0.932        | 0.780           | 0.910 | 0.914           | 0.772 | 0.970           | 0.530 |
|                             | 9            | 0.608        | 1.000        | 1.000           | 0.884 | 1.000           | 0.718 | 1.000           | 0.460 |

*Note.* Fifty-one participants were at or above a PAI PIM cutoff score of 57 *T*, and 44 participants were below a PAI PIM cutoff score of 57 *T*. PPP = positive predictive power; NPP = negative predictive power.

Table 7

*Contingency table for optimal cutoff scores for the 18-item CAP L scale and the revised 14-item CAP L scale (n = 95).*

| Scale           |                        | PIM                     |                      |            |
|-----------------|------------------------|-------------------------|----------------------|------------|
|                 |                        | Fake good ( $\geq 57$ ) | Honest ( $\leq 56$ ) |            |
| 18-item L scale | Fake good ( $\geq 8$ ) | 49                      | 13                   | 13.7% = FP |
|                 | Honest ( $\leq 7$ )    | 2                       | 31                   | 2.1% = FN  |
|                 |                        |                         |                      | 84.2% = TC |
|                 |                        | Fake good ( $\geq 57$ ) | Honest ( $\leq 56$ ) |            |
| 14-item L scale | Fake good ( $\geq 6$ ) | 45                      | 11                   | 11.6% = FP |
|                 | Honest ( $\leq 5$ )    | 6                       | 33                   | 6.3% = FN  |
|                 |                        |                         |                      | 82.1% = TC |

*Note.* PIM  $\geq 57$  T used as criterion for “fake good” responding. FP = false positive rate; FN = false negative rate; TC = total correct classification rate.

Table 8

*Bivariate correlations for variables tested in the mixed effects models (N = 125).*

| Variable       | 1     | 2     | 3     | 4     | 5    | 6    | 7    | 8 |
|----------------|-------|-------|-------|-------|------|------|------|---|
| 1. AGG         | –     |       |       |       |      |      |      |   |
| 2. DEP         | .60*  |       |       |       |      |      |      |   |
| 3. NON         | -.43* | .59*  |       |       |      |      |      |   |
| 4. STR         | .46*  | .59*  | .41*  |       |      |      |      |   |
| 5. Relation    | .14   | -.06  | .15   | -.02  |      |      |      |   |
| 6. Abuse type  | .10   | .12   | .10   | .02   | .11  |      |      |   |
| 7. Lie         | -.55* | -.51* | -.32* | -.52* | .01  | -.11 |      |   |
| 8. Revised Lie | -.58* | -.54* | -.35* | -.56* | -.03 | -.13 | .98* | – |

*Note.* AGG = PAI Aggression; DEP = PAI Depression; NON = PAI Nonsupport;  
 STR = PAI Stress; Relation = caregiver relationship to child; Lie = 18-item CAP  
 Lie scale; Revised Lie = 14-item revised CAP Lie scale.

\* $p < .01$

Table 9

*Results of the random intercept models for the CAP Lie scale and the revised CAP Lie scale (N = 125).*

| Parameter       | Coefficient | SE    | 95% CI |        | <i>t</i> | <i>p</i> |
|-----------------|-------------|-------|--------|--------|----------|----------|
|                 |             |       | Lower  | Upper  |          |          |
| CAP Lie         |             |       |        |        |          |          |
| Intercept       | 22.196      | 1.556 | 19.106 | 25.286 | 14.263   | .0000    |
| AGG             | -0.132      | 0.029 | -0.191 | -0.073 | 4.571    | .0001    |
| STR             | -0.122      | 0.027 | -0.178 | -0.066 | 4.486    | .0001    |
| Revised CAP Lie |             |       |        |        |          |          |
| Intercept       | 19.563      | 1.295 | 16.992 | 22.134 | 15.109   | .0000    |
| AGG             | -0.117      | 0.024 | -0.166 | -0.069 | 4.935    | .0000    |
| STR             | -0.118      | 0.022 | -0.164 | -0.072 | 5.233    | .0000    |

*Note.* Models estimated with restricted maximum likelihood. AGG = PAI Aggression; STR = PAI Stress. CAP Lie  $R_1^2 = .374$ ; Revised CAP Lie  $R_1^2 = .431$ .

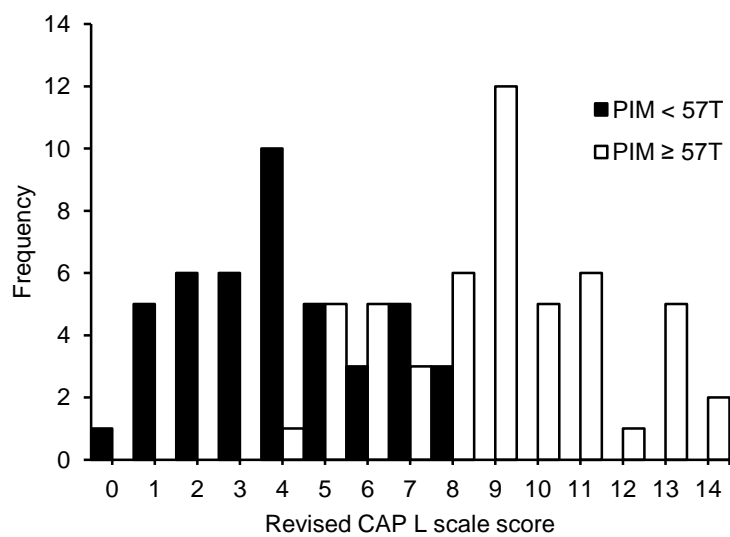
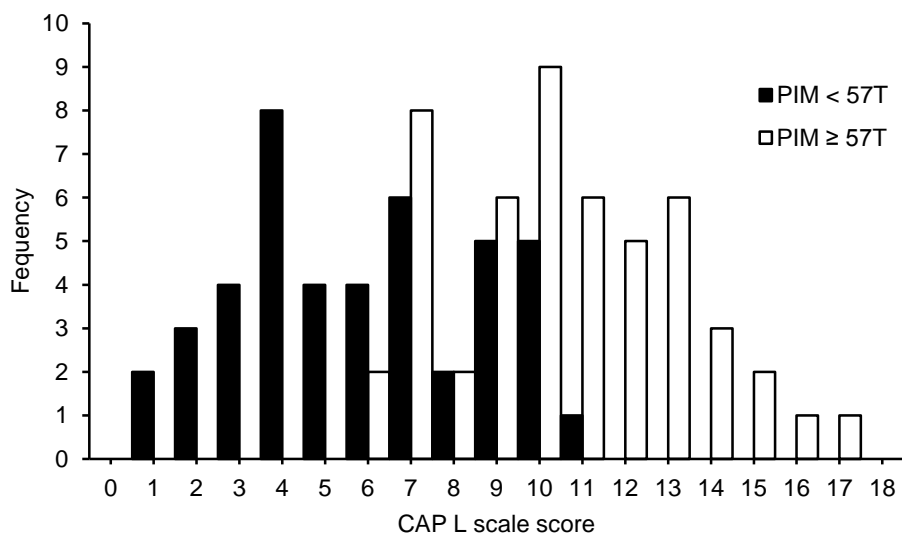


Figure 1. Distribution of CAP L scale and revised CAP L scale scores divided by PIM criterion.

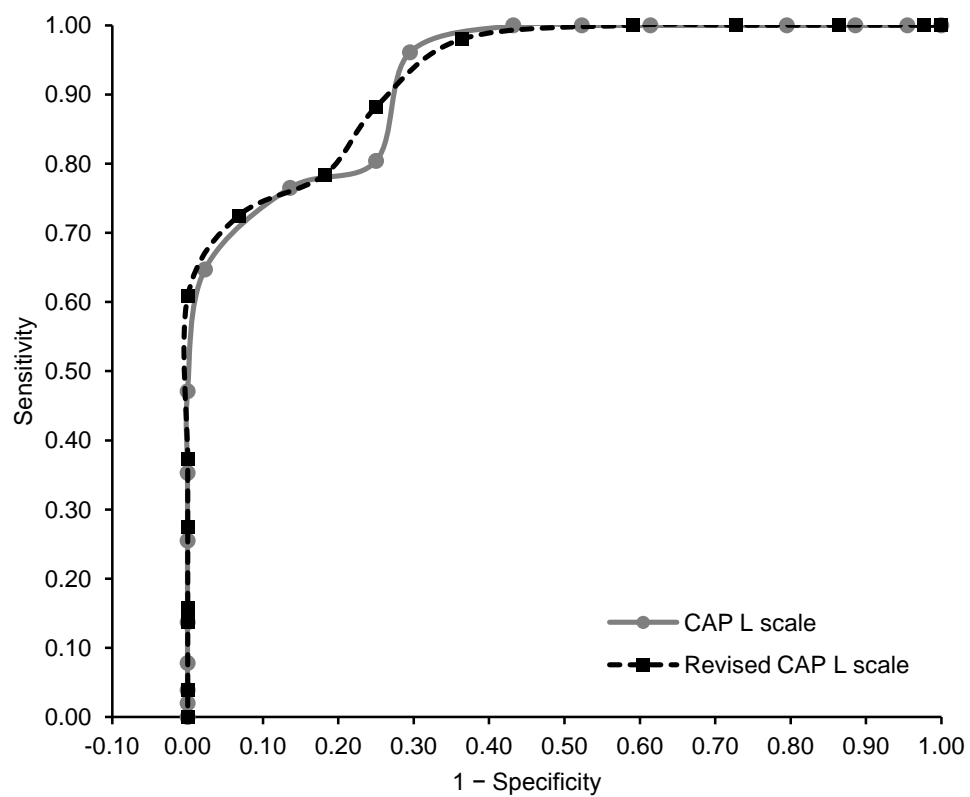


Figure 2. Receiver operating characteristic (ROC) curves for the CAP L scale and revised CAP L scale ( $n = 95$ ).

**Appendix A****DEMOGRAPHIC FORM / CODING SHEET****PARTICIPANT # \_\_\_\_\_****(1) [clinician] Clinician code: \_\_\_\_\_****(2) [datdep] Grouping number: \_\_\_\_\_****How related:**      \_\_\_\_\_ Bio Parent                      \_\_\_\_\_ Other Parent                      \_\_\_\_\_ Kinship**(3) [yrval] Year of evaluation: \_\_\_\_\_****(4) [relation] Caregiver relationship to the child:**

|                     |                              |                            |
|---------------------|------------------------------|----------------------------|
| ____ Bio Mother [1] | ____ Other Male Parent [3]   | ____ Kinship [5]           |
| ____ Bio Father [2] | ____ Other Female Parent [4] | ____ Paternity Pending [6] |

**(5) [phabuse] Physical abuse present: \_\_\_\_ Yes [1], \_\_\_\_ No [2]****(6) [sxabuse] Sexual abuse present: \_\_\_\_ Yes [1], \_\_\_\_ No [2]****(7) [neglect] Physical neglect present: \_\_\_\_ Yes [1], \_\_\_\_ No [2]****(8) [otabuse] Truancy / Other Legal problems present: \_\_\_\_ Yes [1], \_\_\_\_ No [2]****(9) [cage] Caregiver age: \_\_\_\_\_ Years****(10) [cagen] Caregiver gender: \_\_\_\_ Female [1], \_\_\_\_ Male [2]****(11) [chage] Child age: \_\_\_\_\_ Months****(12) [chgen] Child gender: \_\_\_\_ Female [1], \_\_\_\_ Male [2]****(13) [carace] Caregiver race or cultural background**

|                                   |  |
|-----------------------------------|--|
| ____ Black / African American [1] | ____ American Indian / Alaska Native [5]   |
| ____ White / Caucasian [2]        | ____ Asian [6]                             |
| ____ Hispanic / Latina [3]        | ____ Hawaiian / Other Pacific Islander [7] |
| ____ Biracial [4]                 |  |

**(14) [relstat] Caregiver relationship status: \_\_\_\_ Single [1], \_\_\_\_ In a Relationship [2], \_\_\_\_ Unk [3]**

**(15) [caeduc] Highest caregiver education:**

|   |   |  |
|---|---|--|
| <input type="checkbox"/> < HS Diploma [1] | <input type="checkbox"/> Certificate [4]  | <input type="checkbox"/> Bachelors [7]             |
| <input type="checkbox"/> HS Diploma [2]   | <input type="checkbox"/> Some College [5] | <input type="checkbox"/> Graduate/Professional [8] |
| <input type="checkbox"/> GED [3]          | <input type="checkbox"/> Associates [6]   | <input type="checkbox"/> Unk [9]                   |

**(16) [employ] Caregiver employment status:**

|   |                                      |
|---|--------------------------------------|
| <input type="checkbox"/> Employed [1]   | <input type="checkbox"/> Student [3] |
| <input type="checkbox"/> Unemployed [2] | <input type="checkbox"/> Unk [4]     |

**(17) Personality Assessment Inventory:**

|                               |                               |                               |                               |                               |                               |
|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
| <input type="checkbox"/> ICNr | <input type="checkbox"/> ICNt | <input type="checkbox"/> DEPr | <input type="checkbox"/> DEPt | <input type="checkbox"/> DRGr | <input type="checkbox"/> DRGt |
| <input type="checkbox"/> INFr | <input type="checkbox"/> INFt | <input type="checkbox"/> BORr | <input type="checkbox"/> BORt | <input type="checkbox"/> AGGr | <input type="checkbox"/> AGGt |
| <input type="checkbox"/> NIMr | <input type="checkbox"/> NIMt | <input type="checkbox"/> ANTr | <input type="checkbox"/> ANTt | <input type="checkbox"/> VPIr | <input type="checkbox"/> VPIt |
| <input type="checkbox"/> PIMr | <input type="checkbox"/> PIMt | <input type="checkbox"/> ALCr | <input type="checkbox"/> ALCt | <input type="checkbox"/> STRr | <input type="checkbox"/> STRt |

**PIM item scoring:**

[Table deleted: Protected test information.]

**(18) Child Abuse Potential Inventory:**

\_\_\_ Lie

\_\_\_ IC

\_\_\_\_\_ Abuse

\_\_\_ RR

**Lie scale item scoring:**

[Table deleted: Protected test information.]