Functioning of Standardized Self-Report Measures for Caregivers with Active Child Welfare Service Cases

Douglas Klinman

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Gail Joseph, Chair

Virginia Berninger

Monica Oxford

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

Functioning of Standardized Self-Report Measures for Caregivers with Active Child Welfare Service Cases

Douglas Klinman

Chair of Supervisory Committee:
Professor Gail Joseph
Educational Psychology

The purpose of this dissertation was to assist in the development of an assessment system that supports caseworkers in Child Welfare Service (CWS) in making informed decision regarding the families they are serving. Caseworkers CWS have a difficult job of protecting and promoting the well-being of vulnerable children. To accomplish this task, they need assessment tools that both predict future maltreatment and provide guidance on family functioning, service need, and treatment progress. This dissertation presents two studies that examine the results of seven standardized self-report assessments obtained from 318 caregivers involved with CWS (235 families) whose children are in their care, and attempts to use these data to address some of the assessment needs in CWS. Study 1 examined the standardized assessment results and used survival analysis to determine whether the assessments were independently predictive of future child maltreatment and whether they added predictive value to the Structured Decision Making (SDM) tool used by CWS. Due to the coercive nature of caregivers involvement with CWS, the data were analyzed separately for caregivers who were categorized as reporting non-defensively and defensively on the Parent Stress Index. The assessment results for caregivers who were categorized as non-defensive provided insights into the families' struggles and added predictive value both independently and additionally when included in survival analysis with the SDM. For defensive caregivers, the assessments did not appear to provide insights into their struggles,

and scores were not predictive of future child maltreatment. The results indicated no significant difference in survival time between the defensive and non-defensive caregivers. Furthermore, caseworkers appeared to struggle with the assessment of the defensive responders, as indicated by the SDMs lack of predictive value for these families. Study 2 examined how the data from the non-defensive responders could be used to develop a self-report multidimensional assessment of caregivers involved with CWS. Item Response Theory (IRT) was used to examine the functioning of items on two of the administered assessments. The results indicated both adequate item difficulty and item discrimination parameters. IRT was used to reduce the assessment length, resulting in shorter scales with similar predictive validity. Lastly, the study investigated how IRT can assist in the development of assessment tools that address assessment needs of CWS including potentially being resistant to the defensive responding of caregivers.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to acknowledge my chair and committee members, Gail Joseph, Andrew Benjamin, Virginia Berninger, Douglas Cheney, and Monica Oxford. The support, encouragement, and insight of each of these committee members facilitated the completion of this dissertation. I would also like to express my appreciation to Catherine Taylor for her review of my IRT analysis, and for sharing her insights into reliability and validity.

I would like to acknowledge those involved in creating and administering the Comprehensive Assessment Program. Without the creation of this thoughtful program, and the willingness to share the obtained data, this dissertation would not have been possible. I would also like to thank the parents who participated in the Comprehensive Assessment Program, for without their willingness to complete the various assessments this dissertation would not have been possible. Additionally, I would like to thank the staff at Child Welfare Services who assisted in insuring that I received the data needed for this dissertation. I would also like to express my gratitude for having received funding support from the Gatzert Child Development Dissertation Fellowship, the Doi Doctorial Research fund, and for the generous grant from Applied Psychological Measurement.

I would also like to express gratitude to my family. The support that I received from my parents has been invaluable; it has been a long road and they have been there for me every step of the way. Lastly, I would like to thank my children, Alex and James, and my partner Laura. I depended on their support and encouragement throughout this process.

Chapter 1

Introduction

Child Welfare Services (CWS) throughout the United States are expected to provide a safety net for the children most at risk of abuse and neglect. CWS typically becomes involved when an individual contacts the agency with a concern that a child is being maltreated. In 2010, CWS agencies in the United States received an estimated 3.3 million reports involving the alleged maltreatment of approximately 5.9 million children (*Child Maltreatment 2010,* 2011). Once CWS receives a report, a decision must be made as to whether the report meets sufficient criteria to warrant CWS involvement. In 2010, nearly 2 million of these reports were determined to require further CWS involvement, representing 3.6 million children.

When CWS agencies accept these reports, they typically assign the reports to caseworkers for investigation. The caseworkers duties include (a) assessing the immediate level of risk to the child(ren), (b) determining the likelihood of future maltreatment, (c) determining family service needs to promote the safety and well-being of the children in the home, (d) matching family members to appropriate services, (e) and monitoring their progress in these services. Furthermore, the caseworker must determine whether the child is currently safe enough in the family home or whether he/she needs to be placed in alternative care. Making this already complicated task even more difficult is that this work often occurs in a coercive environment, as parents are typically not involved with CWS voluntarily.

The decisions of the caseworkers in CWS also carry significant consequences, as the choices that are made affect the safety and well-being of vulnerable children. Caseworker decisions ideally should be guided by assessments with strong support for their validity. Furthermore, these decisions should be based on assessments with established reliability. To assist caseworkers with these decisions, CWS implemented a number of different assessments. Perhaps due to the significance and complexity of the decisions made by caseworkers, a strong emphasis in the assessment process has been placed on trying

to ensure the current safety of children from being abused or neglected by their caregivers, whereas less emphasis has been placed on assessing family service need and promoting child well-being (Samuels, 2012). Additionally, in a field like child welfare, an emphasis is placed on determining whether the problem is important enough to warrant intervention. In CWS, a problem becomes important enough to warrant intervention when a determination is made that the threat to the child has crossed a predetermined threshold. One way in which this threshold could be established is by determining the criteria for minimum parenting. Intervention would then be warranted when a determination is made that this minimum standard is not being met. However, as Budd (2001) argued, the fundamental issue complicating the task of assessing parenting competency in CWS, "is the *absence of universally accepted standards* of minimal parenting adequacy" (p. 3).

The lack of an accepted minimal standard of parenting is related to the complexity of the task as opposed to the lack of effort. One of the challenges is that severe child maltreatment has a low base rate and as such is difficult to predict (Munro, 2004). In addition, since minimal parenting standards must be understood within the relationship between individual parents and children, the thresholds may be different for different parent/child dyads (Chicchetti, 2010). Furthermore, protective factors may be present in certain situation (e.g., a protective caregiver), mitigating the risks of future harm, but absent in others. Perhaps due to the challenges of ascertaining a clear and universally accepted definition of "minimal standard", the focus in CWS assessment has often been simplified to prioritizing the risk of present and future maltreatment as opposed to ensuring that children receive a minimal standard of care. The recent use of statistical models has enhanced the development of actuarial models to assist caseworkers in determining which children are at an increased risk of being maltreated. However, the actuarial models currently used by CWS provide no guidance regarding treatment planning, progress monitoring, or evaluating treatment outcome. Furthermore, focusing primarily on

risk of maltreatment decreases the attention paid to the well-being of the children with whom CWS comes into contact.

The lack of more comprehensive assessments with support for their reliability and validity becomes problematic for a number of reasons. For example, it is difficult to match services to families without accurate measurement of their current functioning. Furthermore, without accurate measurement, it is not possible to assess progress and adjust interventions accordingly. An example of how this negatively influences current service provision in CWS can be seen when examining the recent push to utilize evidence-based programs. Efforts to increase the use of evidence-based programs in child welfare have recently been documented (Barth, 2009); however, evidence-based intervention have been designed and researched to intervene with well-defined problems, which typically can only be determined with adequate assessment. Without adequate assessments, caseworkers must base the decisions about which programs they should refer families to on unassisted clinical judgment, which has been shown to be poor at diagnosing clients' needs (Grove & Meehl, 1996). This is particularly concerning with the child welfare clientele, for whom the mental health concerns of children (Burns et al. 2004) and parents (Child Welfare Education and Research Programs, 2011) are high. Another area of concern is that child well-being, which is a CWS goal according to Adoption and Safe Families Act (ASFA, Public Law 105-89) and an area needing greater emphasis (Samuels, 2012), requires a more holistic examination and a more refined assessment system than the one that CWS caseworkers currently use.

This dissertation first provides an overview of reliability and validity, with the focus on their relation to assessment in CWS (Chapter 2). Subsequently, this dissertation provides a review of the current tools used by CWS in one state, the state of Washington, to get an understanding of the strengths and shortcomings of the current assessment system used by CWS. This is followed by an overview of the struggles facing families involved with CWS, and then an examination of the

Comprehensive Assessment Program (CAP). The CAP is an alternative assessment model that administers a number of standardized assessment tools, and it has been piloted with caregivers believed to be at a high-risk of future child maltreatment (Chapter 3). Two studies, which were based on client level data obtained from both the CAP program and from CWS, will then be reported. Study 1 (Chapters 4, 5, and 6) examines how the various assessments used by both the CAP and CWS function in determining family needs and predicting future child maltreatment. By examining the same data used in Study 1, Study 2 (Chapters 7, 8, 9, and 10) explores the ways in which procedures based on Item Response Theory can be used to create a more refined and comprehensive assessment system for CWS.

Chapter 2

Theoretical Background: Issues Related to Reliability and Validity of Assessment in Child Welfare

Various assessment tools have been implemented to assist Child Welfare Service (CWS) caseworkers. Before analyzing these assessment tools, it is valuable to first examine some concepts related to reliability and validity. In particular, concepts that are pertinent to both the assessments of clients served by CWS and to the context in which these assessments are used will be highlighted. Concepts of reliability and validity are of central importance in CWS. The potential consequences of determining which children are safe in their homes, which children need to be removed from their homes, or what services family members must engage in are significant and must be guided by reliable and valid assessment procedures.

2.1 Reliability

In order for CWS to support consistent and appropriate treatment of clients, there must be clear guidelines and standards in place to ensure the reliable assessment of the clients. The guidance should both outline the domains that should be assessed as well as provide assessment tools to assist caseworkers in accurately estimating families' "true score" (i.e., actual level of functioning) in various domains. Although defining the domains that should be included belongs to the realm of validity, reliability is concerned with the measurement error around the obtained scores and the likelihood that if the assessment is repeated the results would be the same. Reliability creates a foundation on which the credibility of any assessment must be built. Two general principles should be taken into account: providing evidence in support of the consistency of the obtained scores over repeated measurements, and establishing the parameters for which the generalizations of the scores are warranted (Miller, Linn & Gronlund, 2009).

Decisions in CWS are typically made by caseworks who gather information by investigating and/or interacting with their clients. Under these conditions, error (variance from the "true score") is related to three general components: persons (clients), raters (caseworkers), and the assessment items (Feldt & Brennan, 1998). Caseworkers primarily assess family functioning during face-to-face contact with the client. This interaction typically occurs in a highly emotive setting, as the CWS caseworker is questioning the ability of the caregivers to ensure the safety and well-being of their children. Different caregivers are going to respond to this situation in different ways. Some parents may answer questions in a manner that approximates how things are typically (their true score), other parents may try to present an overly positive picture, and still others may become hostile and present an overly poor picture of their family's functioning. The difference in the presentation of information can be considered the variance in the score attributable to client presentation. In addition to client variability in reporting, caseworker inconsistencies can affect family assessment in numerous areas.

Caseworker variance in scoring can occur either between different workers or within individual workers on different occasions. Variance between individual caseworkers' assessments depends on variables such as the workers experience and knowledge, which may influence the type of information sought and the manner in which this information is documented. Factors, such as a caseworker's current caseload or workers' feelings on a given day influence the assessment performed by individual workers. One method that often increases the accuracy and reliability of the assessment of family functioning is to encourage caseworkers to obtain information from multiple sources as well as to look into the family's history of functioning. Even though these steps may mitigate some of the errors related to gathering information solely from direct contact with the clients, it also introduces further challenges to creating consistency in the assessment of families. This might occur if the same inter and intra-caseworker differences that lead to the initial bias become further confounded during the collection and interpretation of information obtained from the collateral contacts.

The variance in client presentation and caseworker assessment becomes exacerbated when you include the client by caseworker interaction. For example, caseworkers who have limited de-escalation skills will likely obtain different information from an angry confrontational client than will a caseworker with strong de-escalation skills. Furthermore, client and caseworker bias, which comes into play based on their own previous experiences and stereotypes, may impact both the provision and interpretation of relevant information.

Another area that can contribute to variances in the obtained information is related to the assessment items. As noted previously, current assessments are based largely on caseworker interviews, observations, and the collection of collateral information. Although inter-rater consistency in the assessment of the gathered information overlaps with both the rater and client facets of reliability previously outlined, additional variability can depend on the assessment format; in particular, the level of specificity provided in the assessment items can influence consistency in the gathering and reporting of information. To control for inter-rater variability, assessments should be based on clearly explicated criteria, which are laid out in well organized scoring rubrics (Miller, Linn, & Gronlund, 2009). These guidelines are particularly applicable for assessing something as complex as family functioning, child safety, and child well-being. The extent to which these guidelines are present and utilized in the assessment process is yet another factor, which will independently influence inter and intra-rater reliability. Furthermore, the level of specificity and clarity required for the completion of the assessment will interact with both the caseworkers' approach to gathering the information as well as the client's presentation of information, further compounding the reliability of the information gathered.

As can be seen, numerous factors can influence the reliability of the assessment of clients involved in CWS. Due to both the independence and interdependence of the variance surrounding the

potential errors in the assessments, it would be inappropriate to use classical test theory to examine reliability estimates. Classical test theory is based on the model:

$$X = T + E$$

where the observables (X or assessed score) are listed on the left side of the equals sign and the unobservable True score (T) and Error score (E) are listed on the right side of the equation. This model has only one error term; consequently, all sources of error are confounded into this one term. Furthermore, the model assumes that all error is random and that none is systematic. However, as outlined previously, there are many likely sources of measurement error in assessing CWS involved clients, and much of this error may be systematic as opposed to random. Due to these concerns, the more appropriate statistical model would be that of Generalizability Theory, as it allows for a closer examination of the obtained error by separating the multiple sources of error that are of interest to the investigator (Brennan 2011). Understanding the independent errors from the various sources is an important consideration in assessment by CWS, since separating the proportion of errors due to client presentation, caseworker skill, and assessment tool clarity, as well as the interaction between these factors has significant practice and policy implications. For, example if a significant source of error is determined to come from client presentation, this might indicate that a stronger emphasis should be placed on obtaining information using multiple methods and relying on multiple sources. If caseworker skill appears to be a major concern, efforts should focus on increasing caseworker knowledge and skills in assessing the domains of interest. Likewise, if the tools seem to be creating a significant amount of error, the efforts should then be made to strengthen the quality of the assessment items.

The complexities of determining the sources of error impacting reliability measurement becomes clearer when all possible error, and combination of errors, are presented. For example, if *c* =

client, *w* = caseworker, and *i* = item (assessment tool), the following formula represents the variance (error) of the observed scores, assuming that the effects are not correlated (Brennan, 2011):

$$\sigma^2(X_{cwi}) = \sigma^2(c) + \sigma^2(w) + \sigma^2(i) + \sigma^2(cw) + \sigma^2(ci) + \sigma^2(wi) + \sigma^2(cwi)$$

The model's complexity is daunting; however, it provides error variance for each factor and the interactions among factors and consequently, it provides guidance for where the system should be strengthened. Furthermore, as indicated by Brennan (2011), if classical test theory were utilized in this scenario, the estimated error variance would likely be an under-estimation of error, as there is more than one random factor present.

2.2 Validity

A comprehensive examination of validity is beyond the scope of this dissertation. Rather, the focus of this section is to provide an overview of the concepts of validity that are especially pertinent to the assessments by CWS. First, it is important to note the considerable relationship between reliability, as outlined in the previous section, and validity. Brennan (2011) emphasized this point, noting that Generalizability Theory blurs the distinction between reliability and validity. Messick (1989) also indicated this overlap in his seminal work on validity when he wrote that:

> "Tests do not have reliabilities and validities, only tests responses do. This is an important point because test responses are a function not only of the items, tasks, or stimulus conditions but of *persons* responding and the *context* of measurement. This latter context includes factors in the environmental background as well as the assessment setting." (p. 14)

In addition to indicating the overlap between reliability and validity, Messick also highlighted that validity and reliability are characteristics belonging to test responses and not to the test, which has

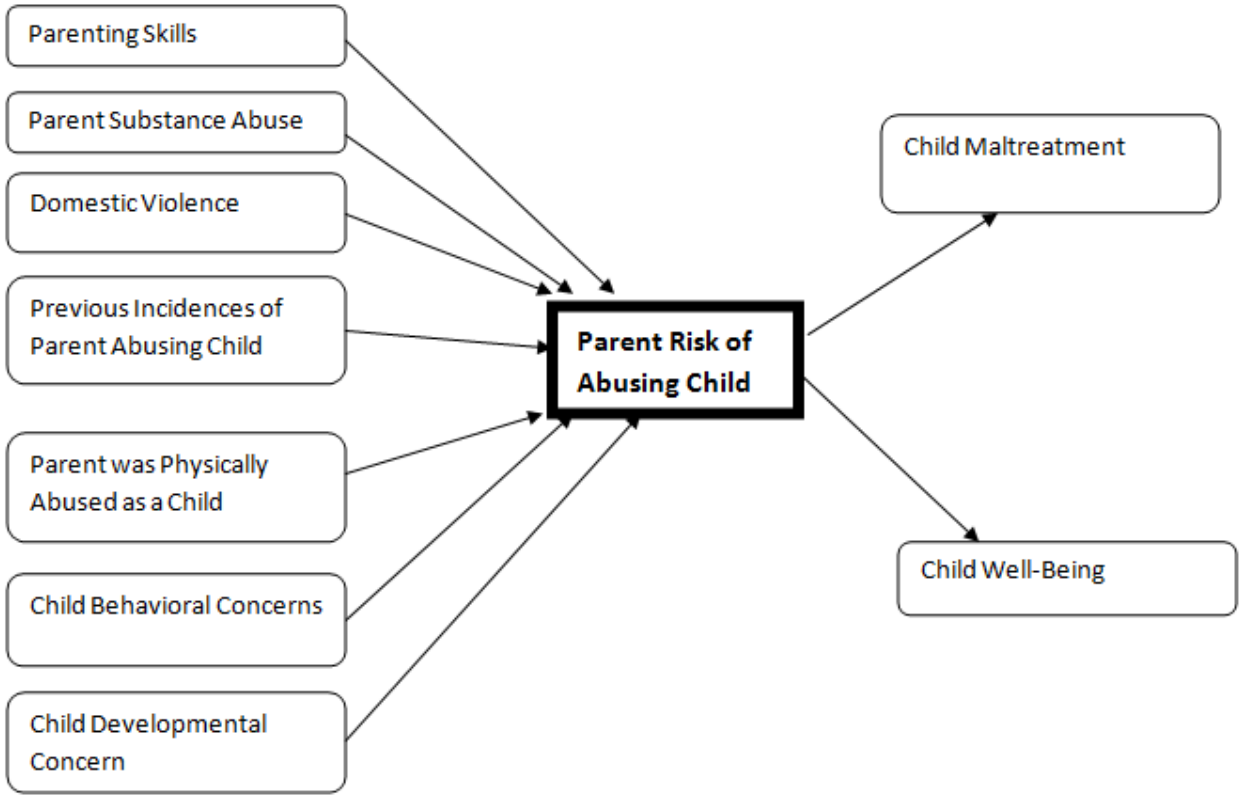
significant implications for the assessment of clients involved with CWS. The context of assessment in CWS is unusual in that it occurs in a coercive atmosphere; consequently, evidence supporting the validity of assessment scores in other contexts would not necessarily transfer over to clients involved with CWS. This point is emphasized by Shepard (1993) who noted that "validity must be established for each particular use of a test" (p. 406).

The distinction between reliability and validity lies in the evidence on which they each focus. Whereas reliability is concerned with the consistency of the obtained scores, and it assesses how closely these represent a "true score", validity is concerned with the meaning of the score in relation to the domains or constructs of interest. This includes the inferences, interpretations, and conclusions, which are drawn based on the attributed scores meaning (Taylor, 2013). With the focus of validity being placed on the overall meaning of the score, Messick (1998) and others (Kane, 2013; Shepard, 1993) have presented a unified theory of validity in which all strategies for obtaining evidence for the validity of test scores (e.g., content-related evidence for validity and criterion-related evidence for validity) are considered parts of the more general notion of construct validity. According to the unified theory of validity, the primary concern of validity is the provision of "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (Messick, 1998, p. 13). The unified theory of validity has various implications for the assessments by CWS. In particular, it not only supports score meaning using content and criterion validity, but also extends the traditional definition of validity to include utility, value implications, and social consequences (Shepard, 1993).

To assist in showing the interplay between various validation methods and score meaning, it is helpful to present a hypothetical nomological network. The concept of the nomological network was

presented by Cronback and Meehl (1955) who asserted that "we shall refer to the interlocking system of laws which constitute a theory as a *nomological network"* (p. 290). Although the 'system of laws' contention has been challenged as not accurately fitting experimentation in the social sciences (Shadish, Cook, & Campbell, 2002), the nomological network does provide a useful representation for examining many of the potential sources of evidence for the validity of inferences from test scores. A hypothetical nomological network representing variables associated with parent risk of physically abusing their child is presented in Figure 2.1 (this is a simplified network that should include both more variables and more interconnections but serves the purpose of clarifying the concept). The variables on the left side of the network are represented as being causally related to the construct of interest, Parent Risk of Physically Abusing their Child. This construct is then represented as being causally related to both child abuse and child wellbeing. The construct in this model, Parent Risk of Abusing their Child, is also a measurable variable. However, the nomological network emphasizes that the Parent Risk of Abusing their Child is the variable of focus in the model. Furthermore, it is important to note that the outcomes in this model are measurable and observable. For example, child abuse can be determined by interviewing the child or by witnessing bruising on the child, while child's wellbeing can be measured by assessing success in school, mental health status, and the like. Using this nomological network, we can now examine how content and criterion related evidence for validity, as well as the utility, value implications, and consequences of the obtained assessment scores support construct validity.

Figure 2.1: Nomological Network of Parent Risk of Abusing Child



Content-related evidence for validity is determined by how well the variables on the left side of the nomological network are theoretically related to and represent the construct of interest as well as the judgment of the adequacy of the assessment tool in representing these variables. Experts in the field under study usually determine the content-related evidence for validity. As Messick (1989) wrote, "content validity provides judgmental evidence in support of the domain relevance and representativeness of the content of the test instrument, rather than evidence in support of the inferences to be made from the test score" (p.17). The greater the theoretical and correlational links between the variables informing the construct and the actual construct the greater the support is for the content validity. Consequently, the higher the levels of knowledge about what leads to parents physically abusing their children and the more this knowledge is represented in the assessment tool, the greater will be the resulting content related evidence for validity. Another point regarding content-

related evidence for validity is worth noting. Some of the variables on the left side of the network best fit in what can be termed as 'attribute variables' (Kerlinger, 1986). Attribute variables are impossible or at least difficult to manipulate. An example of an attribute variable in the sample nomological network would be "previous incidences of parent abusing child"; also, depending on the context and the developmental issue, "child developmental concern" could also be attribute variables. Kerlinger (1986) also refers to 'active-attributes', which are generally flexible and modifiable (e.g., parenting skills and domestic violence). In the field of child welfare, this distinction is important because if the intervention is to be successful in reducing child maltreatment in the family home, the active-attributes must understood, measured, and targeted.

Messick (1989) noted that "criterion related validity is based on the degree of empirical relationship, usually in terms of correlations or regressions, between the test scores and criterion scores" (p. 17). In the nomological network example, the criterion scores would be the items listed on the right side of Figure 2.1, that is, incidences of child abuse and status of child well-being. Criterion related evidence in support of an interpretation of a test score can be obtained either concurrently or predicatively. Concurrent validation would examine whether the Parent Risk of Abusing their Child is related to current physical abuse and the current well-being status of the child while predictive validation is concerned with the assessment of the score's relationship with future child maltreatment and child well-being. In order to have meaningful concurrent and predictive evidence for validity, the criterion measures must be well defined (e.g., what constitutes physical abuse and what areas of child well-being are of concern) as well as accurately measured. This can be a significant problem in fields such as child welfare in which the outcome criterion are often not clearly defined nor accurately measured (this point will be outlined more fully throughout this dissertation). Miller, Linn, and Grunlund (2009) noted concerns with overreliance on criterion measures and contended that this increases the importance of both content-related and construct-related evidence for validity. An

additional concern with overreliance on criterion measures was highlighted by Messick (1995) who wrote that, “as a salient social value, validity assumes both a scientific and a political role that can by no means be fulfilled by a simple correlation coefficient between test scores and purported criterion” (p. 742). With these thoughts in mind, we now take a closer look at the unifying theory of construct validity.

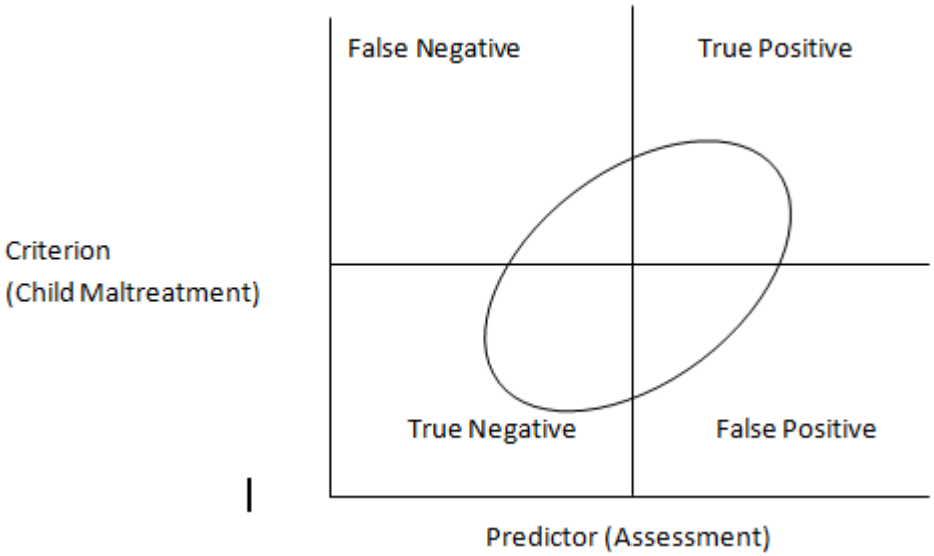
Construct validity is typically given the highest priority by measurement specialists (Miller, Linn, & Gronlund, 2009). However, for a number of reasons, construct validity is more than the combination of content and criterion-related validity. One of the concerns related to overreliance on content and criterion-related evidence for validity involves potential errors related to construct underrepresentation and construct irrelevant variance (Kane, 2013; Messick, 1989; Miller, Linn, & Grunland, 2009). Construct underrepresentation would indicate that the necessary variables are not being included either at the content level or at the criterion level. For example, perhaps certain aspects of child well-being that should be included in the construct or in a criterion measure are not included. Construct-irrelevance would occur if irrelevant factors were included in the content informing the construct or in the method of assessment or if irrelevant factors were influencing the criterion scores. An example of irrelevant variance might be seen if a CWS-involved parent were given an assessment and were motivated to provide misleading information to reduce potential negative consequences. Due to the concerns of construct-underrepresentation and construct-irrelevant variance, Messick (1989) argued for a unified theory of construct validity that includes not only evidence from content and criterion sources, but also issues of the instrument’s utility, value implications, and social consequences. Messick argued that evidence from each of these areas contributes to a holistic interpretation of, “What does the score mean?’’

An assessment's utility reflects the cost/benefits analysis of testing (Messick, 1995). In the field of child welfare where caseworkers are typically overburdened and often do not have expertise in testing, the utility of the assessment tool may largely be based on the instruments administrative and interpretive burden (Miller, Linn, & Grunland, 2009). Furthermore, the utility of the tool is also judged by the relevance of the obtained scores to the objectives of CWS (e.g., does the score provide information indicating the level of risk that a child is currently facing in her/his family home?).

Two additional items included in the unified theory of construct validity are the value implications and social consequences of the scores. These are part of what Messick (1989, 1995) referred to as the consequential basis of score interpretation. Value implications play a significant role in the field of child welfare. For example, in the nomological network presented in Figure 2.1., child abuse and child well-being are maintained as separate outcomes. However, one might contend that child abuse could, and possibly should, be subsumed under child well-being. However, the determination to keep these criteria separate is value laden (e.g., perhaps defining the CWS role as preventing future child maltreatment is seen as more realistic than the promotion of child well-being). Another example of how values impact the interpretation of scores in CWS is that assessments are often used to help determine which children should remain with their parents and which children should be placed in alternative care. This type of dichotomous decision (remove or do not remove) utilizing assessments with continuous measures is represented in Figure 2.2. In this figure, true positives represent children who are placed in out-of-home care, and if they had not been placed, they would have continued to suffer significant maltreatment. True negatives represent those children who were correctly determined not to be at a risk of being exposed to significant future maltreatment. Children in the false negative quadrant were predicted not to be repeatedly abused but in fact, they were consequently significantly maltreated. Lastly, those in the false positive quadrant were incorrectly assessed as needing to be placed due to concern of future maltreatment. The shape of the oval

represents the accuracy of the assessment, with a narrower oval representing a more accurate tool. The determination of where to set the criterion (e.g., significance of maltreatment) and predictor [likelihood of maltreatment] cut scores are based on social values. Namely, the ratio of false positives to false negatives is a policy determination based on social values.

Figure 2.2: Representation of Assessment Accuracy



Consequential validity is the last aspect included in the unified theory of construct validity. Consequential validity is concerned with the resulting effect of the test scores, which might be determined, for example, by whether or not the child(ren) was (were) better off having been removed from their caregiver. Messick (1989) drew a distinction between negative consequences due to test invalidity versus test misuse. He wrote that in general, "If the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced – or if the validation process can discount sources of test invalidity as the likely determinants, or at least render them less plausible – then the validity of the test use is not overturned. Adverse social consequences associated with valid test interpretation and use may implicate attributes validly assessed" (Messick, 1989, p. 88). Messick (1989) further noted that

negative consequences of test invalidity are often due to construct underrepresentation and/or construct-irrelevant representation. Messick did not negate the important consequences of test misuse but rather drew a distinction between social consequences due to test misuse versus test invalidity. However, in the past 20 years or so, validity theorists have argued that social consequences due to test misuse should be more fully incorporated into the unified theory of construct validity (Kane, 2013; Shepard, 1993; Taylor, 2013).

Kane (2013) proposed an "interpretation/use argument" (IUA) to assess the validity of the obtained score, where the IUA includes all claims based on test scores. Kane (2013) expanded Messick's (1989) view by including the consequences of the use of the instrument as evidence that can be used to support or challenge an instrument's validity in a given situation. As Kane (2013) proposed, "the evaluation of test score uses requires an evaluation of the consequences of the proposed uses, and the negative consequences can render a score use unacceptable" (p.46). Kane (2013) did however agree with Messick (1989) in that he too contended that, "the rejection of a score use does not necessarily invalidate a prior, underlying score interpretation" (p. 46). Taken together, this implies that a test score may be a valid measure of the construct, although additional evidence is needed to support the validity of the measure in each new situation. Kane (2013) contended that three main outcomes/consequences must be addressed to provide supporting arguments of an instruments use, "(1) the extent to which the intended outcomes are achieved, (2) differential impact on groups (particularly adverse impact on legally protected groups), and (3) positive and negative systemic effects" (p. 48). The importance of evaluating the consequences of an assessment's use is particularly relevant when applying cut-scores, or decision rules, to the assessments results. As Kane (2013) argued, "Decision rules are evaluated in terms of their overall consequences (or utilities, or values) for some population" (p. 47). As assessments in child welfare are often associated with decision rules, it is clear that it is important to consider consequences when determining the validity of an assessment tool in the context of child welfare.

Child welfare uses various assessment tools to assist with decision rules about which client cases should be closed after the investigation, which children should be removed from their caregiver, what services a family should be offered, when children should be returned to their caregivers, and so on. Assuming that an assessment tool could accurately guide inferences made in these various situations, adverse social consequences could occur in spite of these appropriate inferences. As Kane (2013) contended, an essential component of assessment is to improve social conditions; therefore, one must examine the consequences of the actions taken as a result of the scores when determining the instrument's validity in a given situation. For example, based on accurate indicators, children may be removed from their parents' care and placed in foster care due to the concerns about future child maltreatment. However, it may be that the removal to a foster home leads to a different form of trauma, questioning the validity of the tool for this particular purpose. Likewise, Kane (2013) would contend that if CWS determines that a family needs services but the resulting services do not improve or perhaps even worsen the conditions in the family home, the validity of using the assessment for the purpose of service provision is called into question. The challenge of consequential validity ultimately becomes a pragmatic one, measuring and balancing the intended and incidental positive effects of using assessment scores with the negative effects of using these scores.

In summary, the unified theory of validity is concerned with the meaning of the scores. Each of the validation strategies covered add to the construct validity or overall meaning of the scores obtained during the assessment. The aspects of reliability and validity outlined in this chapter provide a framework for how these terms apply to assessment in CWS. Much of the rest of this dissertation focuses on the current and potential use of assessment in the field of CWS, and as such, it utilizes the concepts outlined in this chapter.

Chapter 3

Review of the Literature

Overview

Caseworkers in CWS have difficult jobs, ensuring that the most vulnerable children in society are safe and cared for appropriately. To aid caseworkers in their jobs, CWS has developed a series of assessment tools to assist with major decision points (e.g., should children be placed in alternative care, should the case be opened for services, or should the case be closed post investigation). These types of decisions are primarily based on the perceived level of risk to the child(ren) in the family home, and the available assessment tools help the caseworker in establishing the level of risk.

This literature review will initially examine and analyze the assessment tools used by CWS in the state of Washington. The subsequent sections outline the caseworkers' roles as service brokers and the duties they need to fulfill to accomplish this role adequately. The review will then turn to an examination of the clients who are involved in CWS, with an emphasis on the major struggles they face. Lastly, a pilot program that was implemented in the state of Washington to assist in assessing family needs is reviewed.

3.1 Current Assessment of Child Welfare Service Involved Families

3.1.1 Safety assessment

The primary responsibility of CWS is ensuring child safety. The Safety Assessment is the first tool that caseworkers use when investigating allegations of maltreatment. The Safety Assessment used in the state of Washington was adapted from assessment tools developed at the National Resource Center for Child Protective Services in 2011 (Children's Administration Practices and procedures Guide, 2013). This Safety Assessment tool is used to determine the present and imminent safety threats to the

child(ren) in the home, and it is also supposed to prioritize safety and guide safety focused practice while CWS is involved with the family. The safety assessment consists of four steps, gathering information, assessing information, analyzing the information, and designing and implementing a safety plan if deemed necessary.

In the gathering phase, caseworkers obtain information to help them answer the following six questions:

1) What is the nature and extent of the maltreatment?
2) What surrounding circumstances accompany the maltreatment?
3) How does the child (or children) function on a daily basis?
4) How does the parent(s)/caregiver discipline the child?
5) What are the overall parenting/child care practices used by the caregiver?
6) How does the parent(s)/caregiver manage his/her own life on a daily basis? (This focuses on how the parent functions in an adult role outside of his/her parenting role).

Caseworkers obtain the information needed to answer these questions by interviewing those directly related to the alleged maltreatment (e.g., caregivers, children in the family) and by reviewing documentation regarding the maltreatment. The obtained information should be comprehensive, and it should be corroborated by other sources. The information gathered during this phase informs the next three phases of the safety assessment process.

The second step involves assessing the gathered information. Two tools are available to assist with the assessment of the information (Children's Administration Practices and Procedures Guide, 2013). The first comprises a list of 17 safety threats. These threats include items like, "Caregiver(s) perceive child in extremely negative terms", and "A child has serious physical injuries or serious physical condition resulting from maltreatment". (See Appendix A for a complete list of the safety threats). The

caseworkers must determine, which safety threats, if any, are present in the family home. The 17 safety threats are comprehensive and include many reasons for which CWS would be involved in a family's life. The next step in the process involves the determination of whether the safety threat(s) have crossed a predefined threshold, indicating that the safety threat must be addressed. This threshold is quite high. For example, for the two safety threats mentioned earlier, the safety threshold, as set by the Caseworker Guide and Definition Manual, would be:

- Caregiver(s) perceive child in extremely negative terms. "Extremely" is meant to suggest a perception, which is so negative that when present, it creates child safety concerns. In order for this threat to be checked, these types of perceptions must be present and the perceptions must be inaccurate. Examples of these include the following: Child is perceived to be the devil, demon-possessed, or evil. Child is considered to be punishing or torturing the parent/caregiver.
- A child has serious physical injuries or serious physical condition resulting from maltreatment. The key word is "serious" and suggests that the child's condition has immediate implications for intervention (e.g., need for medical attention, extreme physical vulnerability). The presumption related to this safety threat is there is some connection, either alleged or confirmed, that the physical injuries or physical symptoms are related to maltreatment.

In addition to determining that a safety threat is present, the following four criteria must also be met to cross the safety threshold: (1) The maltreatment must be immediate or likely to occur in the near future (next 30-60 days); (2) Child is vulnerable in relation to the safety threat; (3) The threat must be 'out-of-control' in that no responsible parent or adult in the home can prevent the threat; and (4) Behaviors and conditions are specific, observable, and clearly understood. When all four of these conditions are met, a safety threat to the child is considered present. This assessment is done in an effort to focus the caseworker's efforts on families in which significant harm is likely to occur in the

immediate or near future. After determining that a safety threat exists (i.e., crossing the threshold), a safety plan must be put in place. The next step in the child safety framework is to determine/analyze the appropriate type of plan (Children's Administration Practices and Procedures Guide, 2013).

When creating a safety plan, the focus is on controlling the safety threat. The first issue to consider is, "Can this threat be controlled in the family home, or does the child need to be placed elsewhere in order for the threat to be controlled?" Since one of the factors resulting in the need to implement a safety plan is that the caregiver(s) must not be able or willing to control the concerning threat, other resources must be mobilized to ensure the safety of the child. The following four questions must all be answered positively to indicate that the threat can be managed in the family home; otherwise out of home placement is required. (1) There is a parent/caregiver or adult in the home, (2) the home is calm (stable) enough to allow safety providers to function in the home, (3) the adults in the home agree to cooperate with, and allow for, an in-home safety plan, and (4) sufficient, appropriate, reliable resources are available to provide safety services and tasks. Service plans are then developed to enhance caregiver's capacity to reduce/manage the safety threat(s) and increase protective capacities. The safety plan remains in place until the safety threat no longer crosses the safety threshold.

Although the safety framework has been implemented in the state of Washington as well as in a number of other states, to date there has been no work to provide evidence in support of the reliability of the Safety Assessment tool. This is a significant concern. Although efforts have been made to define safety threats and to determine whether they are out of control, there is still considerable room for subjectivity in determining whether these factors are present. For example, the following phrases from the earlier examples are quite ambiguous: "Child is considered to be punishing or torturing the parent/caregiver", "extreme physical vulnerability", and "the home is calm (stable) enough to allow

safety providers to function in the home". Without careful examination of the consistency with which families are scored on these types of items, it would be an error to attribute any reliability to the obtained scores. Without reliability, it is impossible to objectively establish the meaning of an obtained score, consequently questioning the validity of the Safety Assessment.

Nevertheless, the safety assessment has some content validity, as most would agree that if present, the safety threats would pose a risk to the child(ren) in the home. However, content validity without other converging forms of validity provides only weak support regarding the validity of the assessment tool. As Messick (1989) wrote, "content judgments alone do not provide a sufficient evidential basis for the validity of inferences and actions based on test scores" (p. 42). Indeed, validity involves a direct examination of content validity on the one hand and criterion validity on the other, and without such examination, one simply has an untested theory. Based on these concerns as well as others related to the utility of the tool and the unexamined social consequences of the tool, the Safety Assessment is perhaps best seen at this point as a practice guide that may be used to help educate caseworkers, rather than as an assessment tool, which has evidence supporting its reliability or validity.

As the other assessment tools used by CWS in the state of Washington are next outlined, it is important to keep in mind that the safety assessment model was adopted to guide caseworkers in determining which families cases should remain open and receive services from CWS. When safety threats are present, caseworkers must take actions to control these threats. When safety threats are not present, the cases should be closed. However, another assessment tool used by CWS caseworkers somewhat contradicts the focus of the Safety Assessment. After the initial Safety Assessment, the caseworker then uses an actuarial tool, the Structured Decision Making (SDM) tool, to determine which families are at risk of maltreating their children in the future.

3.1.2 Structured decision-making

After the initial investigation of the family, if it is determined that the child is safe enough to remain in the family home, the next crucial decision is the determination of whether the family's case should remain open and receive supervision and services or whether the case should be closed. This is a key decision, as inaccurate classification could lead to cases being closed in which children are subsequently maltreated or could result in unnecessary intervention and potential unwarranted removal of children (Shlonsky & Wagner, 2005).

Historically, in child welfare, the determination of which cases should remain open had been based on clinical judgment. Due to concerns of the inaccuracy of unassisted clinical judgment (Grove & Meehl, 1996), tools were developed to support caseworkers in making these decisions. Initially, consensus-based tools were developed that were guided by available research and expert opinion (content validity) of factors that were most likely associated with future reports of child maltreatment (Cash, 2001; English & Pecora, 1994). Caseworkers would use these judgment-based tools to determine what risk factors were present (e.g., domestic violence, parent drug/alcohol use, mental illness…). If present, the worker would determine the significance of this threat in the family home (e.g., low, moderate, high) (Shlonsky, Saini, & Wu, 2007). The worker would then add up the scores in the various domains, and the family would then receive a total score indicating the overall perceived risk/likelihood of future maltreatment in the household. However, a number of concerns were reported with these consensus-based assessment tools, including poor conceptualization and inconsistency in terms of the type and number of variables included (Shlonsky et al., 2007). In addition, there were concerns regarding inconsistent scoring of these tools (Baird, Wagner, Healy, & Johnson 1999; DePanfills & Zuravin, 2001). It is noteworthy that many of these concerns could also apply to the Safety Assessment tool described previously. For instance, the Safety Assessment relies primarily on a content validity, ignoring other sources of evidence that are needed to support the instrument's reliability and validity. Another concern regarding the consensus-based models is that they failed to make a needed distinction

between variables associated with the occurrence versus the reoccurrence of maltreatment (Lyons, Doueck, & Wodarski 1996). Due to these concerns, there was a significant push in CWS towards actuarial models.

The actuarial models used in CWS are based on correlations between certain variables and a specific outcome of interest (Rycus & Hughes, 2003). The goal in the creation of these tools was that the resulting scores correlated with the following two criterion measures: 1) the likelihood of future abuse/neglect, and 2) the severity of the future abuse/neglect. Actuarial tools are developed using standard statistical procedures used to analyze the data collected in domains typically examined during a CWS investigation or otherwise available to the assigned caseworker (e.g., age of children, number of previous referrals, mental health status of the parents…). When developing the actuarial tools, the information collected in various domains that correlates with the criterion measures of interest are then placed into multivariate models to control for potential overlap between risk factors, and this information is then weighted to determine its importance based on strength of association with the outcome relative to other included measures (Shlonsky & Wagner, 2005).

Compared to consensus-based models, actuarial models have demonstrated considerably more validity in predicting the criterion of concern (e.g., future reports of child maltreatment and severity of future maltreatment). One study by Baird and Wagner (2000) compared three risk assessment instruments, two consensus-based tools (one that had been used in the state of Washington and one in California), and one actuarial tool (used in Michigan). Based on their study, the researchers concluded that:

> "Actuarial-based systems are more accurate than consensus-based or expert systems and, therefore, have the potential to improve CPS decision making. The value of increased accuracy is perhaps best demonstrated by comparing cases classified to different risk levels by each system. Cases classified high risk by the Michigan system,

but low or moderate risk by the Washington and California systems, had high rates of subsequent maltreatment. Conversely, families classified high risk by the Washington and California system but moderate or low risk by the Michigan scales had subsequent maltreatment rates at or below the average for the entire sample" (p. 868).

These findings are consistent with findings from numerous other fields in which complex judgment is needed, suggesting that actuarial tools produce outcomes that are equal or superior to more informal clinical judgment (Grove & Meehl, 1996). However, the actuarial models developed in the field of child welfare are far from perfect.

Although superior to consensus-based models or unstructured clinical judgment in predicting future incidences of child maltreatment, actuarial models still demonstrate a fair amount of error. The actuarial model can make appropriate distinction between higher risk cases and lower risk cases in terms of future maltreatment. For example, a study of the actuarial tool currently used in California showed that about 8% of families rated as low-risk, 15% rated as moderate-risk, 30% rated as high-risk, and 45% rated as very high-risk had a future substantiated report of child maltreatment. Therefore, the model does a reasonably good job at distinguishing between the groups. However, it still has a fair amount of error, as 70% of the high-risk families and 55% of the very high-risk families did not have a future substantiated report of child abuse/neglect (Shlonsky & Wagner, 2005). Similar concerns regarding the sensitivity of an actuarial tool were also found in a study by Loman and Siegel (2004). Findings such as these has led authors to note that although better than non-statistically based models, "actuarial models are rarely able to predict re-abuse at acceptable levels of sensitivity (correctly classifying those children who will be re-abused)" (Gambrill & Shlonsky, 2000. p. 825). In other words, high sensitivity of the current actuarial models seems to lead to both correct classification of high-risk parents who re-abuse (true positives) and incorrect classification of parents, as a significant percentage of parents scored as high-risk do not re-abuse their children (false positives).

In addition to assisting in the prediction of which children are at a risk of future maltreatment, actuarial tools also have established cut scores that are used to determine the cases that should remain open and receive services and ongoing monitoring from CWS. The cut scores, which determines the tools sensitivity, is based on values that are often influenced by political and popular/media pressure (Messick, 1989). The use of cut scores with such high error rate has led some to write that due to external pressures, CWS has moved from probabilistic models to a possibilistic model (Pithouse et al. 2011). It is possible that in response to this high error rate, the Safety Assessment tool described in the previous section was introduced to help focus the attention on the true true-positives. However, this creates a somewhat circular argument, as the actuarial tool was implemented in response to a lack of accuracy and inconsistency in scores based on models that relied primarily on clinical judgment and content validity, like the Safety Assessment. Another inconsistency between the Safety Assessment and the actuarial assessment is that the Safety Assessment is designed to focus on maltreatment that is likely to occur immediately or in the near future (30-60 days), whereas the statistical method used to develop the actuarial tools was built on re-referral data obtained further in the future. For example, Johnson (2004) monitored cases for 2 years from the index referral to determine the predictive validity of an actuarial tool used by CWS. The potential contradiction between the two tools may create confusion among workers as to which cases to prioritize.

The actuarial risk assessment model used in the state of Washington as well as many other parts of the country is called the Structured Decision-Making Risk Assessment tool (SDM) (National Council on Crime and Delinquency). The SDM is broken up into two categories, abuse and neglect, with each section providing an independent score (See Appendix B for a sample of the SDM used in Washington State). In order for a case to be considered moderate-high or high risk, and consequently referred for

ongoing services/monitoring, the total score for either the abuse or neglect section must be at least 5[1]. The instrument accounts for a history of involvement with CWS. For instance, in the neglect sub-score, three or more previous referrals for neglect would receive 3 points, the household having received services before would receive 1 point, and the current referral for neglect would add an additional point; thus, a family would already reach 5 points. Families can also receive points for a child in the home being developmentally delayed, current caregiver mental illness or substance abuse, domestic violence, and the like. Based on this weighted scoring system, once a family receives a moderate-high or high score on the SDM, it likely to stay in that range if encountered in the future by CWS, assuming that the caseworkers scored the instrument reliably. On the other hand, a family that is being referred for the first time would have to have a number of the other current risk factors in order to reach the threshold.

The SDM tool was developed primarily using criterion related validity measures. As noted earlier, various factors that were found to correlate with the criterion measures of interest were placed into multivariate models to control for potential overlap and were then weighted in terms of their importance according to the strength of association with the outcome of interest. A number of limitations are apparent when examining a model that relies so heavily on criterion-related measures and does not give equal footing to other forms of validity. First, the SDM has only a limited amount of content validity, which compromises the interpretation and the utility of the obtained score. For example, the SDM tool does not provide significant insight into what areas a family is struggling with and does not provide guidance as to what services may be helpful to family; moreover, it cannot be used to monitor a family's progress, thus greatly reducing the utility of the tool (Rycus & Hughes, 2003; Shlonsky & Wagner, 2005). Furthermore, it has not yet been established whether interventions that focus on families with a high SDM score are helpful in reducing further maltreatment or in improving child well-

[1] If a family scores below 5 on the SDM, caseworkers are able to override the scoring and open the case for services.

being. This highlights significant concerns when examining the consequential validity of the SDM. To help with one of these concerns, namely the lack of guidance the SDM provides in determining what services to offer family members, if it is determined that a family is at risk based on the SDM score, the caseworker must do a family assessment to help determine the services that are most appropriate for family members.

3.1.3 Family assessment

In the state of Washington, a comprehensive family assessment is used to guide and monitor service provision and family progress. All families who have their children placed in alternative care or who are referred to ongoing in-home services must have a family assessment within 45 days and a re-assessment at least every 6 months. A practice model adopted by CWS, which is referred to as Solution-Based Casework (SBC) (Christensen, Todahl, & Barrett, 1999), informs the family assessment. SBC is based on an ecological model, which assumes that human behavior can only be understood in the context in which it occurs. In addition to considering environmental factors, case planning also focuses on client competencies, family development theory, and relapse prevention strategies.

The developers of SBC contend that historically, in part due to influences of physical and mental health treatment models, there has been an over emphasis on individual and family dysfunction. In contrast, they note that recent models of service provision stress the need to collaborate with clients and promote working partnerships. SBC builds on this approach by emphasizing client competencies in "detailing (the family members) attempted solutions, identifying moments of success, and encouraging the use of underutilized resources" (Christensen et al. 1999). By assuring clients that they are more than their symptoms and by helping them build their strengths, it is believed that clients will be more likely to engage in services.

Family development theory, which informs SBC, proposes that all families face similar challenges dependent on the stage of development that the family is encountering. For example, there are

commonalities between the experiences of all parents when caring for an infant, a toddler, or an adolescent. By focusing on these commonalities, the challenges that families face can be normalized and partnerships can be based on shared understandings of the most relevant struggles. SBC places particular emphasis on everyday life struggles and tasks; for example, potty training is a common struggle for most families with a toddler. As Christensen et al. (1999) noted, "anchoring case plans in the developmental tasks of everyday life, where the danger occurs, is instrumental in developing clear, pertinent and tailored case plans" (p. 10).

Another element of SBC is its use of Relapse prevention. Relapse prevention theory focuses on "(1) recognition of patterns; (2) learning the details of high-risk patterns; (3) practicing small steps toward change; and (4) creating a relapse prevention plan" (Christensen et al. 1999, p. 16). Relapse prevention is included in the model, as it is often the case that the concerning maltreatment is recurrent in nature; hence, the focus should be on interrupting the sequence of events that typically leads to the inappropriate behavior. In order for the relapse prevention plan to be effective, there must be a clear understanding of the problem so that the plan can match the likely scenario as closely as possible.

SBC combines the elements of client competency, family development theory, and relapse prevention to create a model that is problem centered and solution focused. Christensen et al. (1999) noted that the caseworker in CWS should remain focused on "known and highly probable issues of risk" and further noted that the "temptation to include 'everything but the kitchen sink'" (p. 112) should be avoided. The primacy placed on risk in the SBC model is reflected in the family assessment that caseworkers must fill out (see Appendix C for a sample family assessment). The central focus of the assessment is on the adults' high-risk behavior, which presumably led to their involvement with CWS. The first part of the family assessment incorporates the earlier described Safety Assessment. The Safety Assessment highlights one or more of the 17 safety threats that have been indicated, and the plan that follows is designed to directly address these threats. The family assessment draws attention to the

parents' daily functioning and skills, and the caseworker is asked to describe the caregivers' behavioral, cognitive, and emotional capacity to protect their child(ren). The assessment also asks the caseworker to report on the functioning of the child(ren) in the home. This includes describing the child(rens) general behavior, emotions, temperament, physical capacity as well as their daily functioning. The focus is on gathering information to assist in planning intervention around the factors that are most closely related to the high-risk behavior while building family and individual competencies that would consider the families existing strengths. In summary, the SBC process builds on the Safety Assessment, essentially providing a set of strategies to partner with the family to address the identified safety threats.

The SBC model includes the development of both family and individual level objectives. The family level objective is a broad statement that describes the desired outcome for the case (Christensen et al., 1999). For example, a family level objective might be that the "family will prevent abuse and remain safe together in the home" while the "*individual-level objective* represents the new behavior that the individual(s) will be exhibiting in order to participate successfully in the *family-level objective"* (Christensen et al., 1999, p. 111). These individual level objectives need to be clearly tied to concerning family events that need to improve. An example of an individual level objective might be, "I will use my 'keep cool plan' to ensure my family's safety from physical or emotional hurt." The 'keep cool' plan, that is, a relapse prevention plan, would then detail the skills that the individual will use/acquire to avoid losing their temper. The risk of harm reduction is then measured by documenting the acquisition and use of these skills during specific high-risk times.

SBC does not advocate the use of standardized assessment tools and, as implemented in the state of Washington, does not rely on any standardized assessment tools. Although, in outlining SBC, Christensen et al. (1999) did not explicitly state that standardized assessments do not have a place in the model, they strongly emphasized that clients are more than their symptoms. They noted that

caseworkers using SBC are more interested in the clients' ascribed meaning to the problems and their ideas about solutions rather than their diagnostic categories, which they noted are often used to promote rote intervention schemes. The authors further argued that "the individual-focused, deficit orientation assumes that conflict predominantly resides within the individuals, that individuals are largely unaware of their needs, and that they therefore require the assistance of enlightened professionals" (p. 27). Rather, they suggested that caseworkers should "join with the family around their frustration with resolving certain family issues and should re-instill hope that the problem will get solved" (p. 53). As noted earlier, learned cognitive and behavioral skills that are assumed to prevent the recurrence of child maltreatment should determine the effectiveness of the intervention. The caseworker's assessment of learned skills is based largely on interviews with the client as well as on the reports of others involved in the case.

A couple of recent studies have examined the effectiveness of SBC in CWS. One study examined the effect of SBC on recidivism of child maltreatment (Antle, Barbee, Christensen, & Sullivan, 2009). The study design was quasi-experimental, comparing recidivism rates of clients served by offices that were implementing SBC with high fidelity and clients served by offices that were implementing SBC with low fidelity. The researchers found that recidivism rates (i.e., incidence of future maltreatment of children) were significantly lower in the high fidelity offices compared to the low fidelity offices. However, the method used to select these two groups in this study had limitations. For example, by definition, the two groups differ in their ability to implement the model, which is likely due to other variables not explored in this study (e.g., staffing differences, administrative differences, client differences…). The second study examined the relation between SBC and individual case files meeting the federal standards for casework in child welfare (Antle, Christensen, van Zyl, & Barbee, 2012). In this study, cases that were in high compliance with SBC were compared with cases that were in low compliance. The results indicated that the high compliance cases met the federally defined standards at a significantly higher

rate. However, like the first study, this study has significant limitations related to selection bias. In addition to office and caseworker confounds, this study adds the element of error associated with some clients being perhaps more willing to engage in the SBC process, which in turn may have been the variable influencing the outcome. Based on the design flaws in these studies, SBC is seen as a promising practice rather than as an evidence-based practice (California Evidence-Based Clearinghouse for Child Welfare).

As noted, the lack of standardized measures implemented in the SBC is due to the model's focus on wanting to avoid a deficit orientation. However, a vast amount of research has indicated that more clinical and open-ended methods of assessment lack accuracy and that standardized tools greatly enhance the diagnostic/predictive validity of assessments (Grove & Meehle, 1996). This brings up concern related to the evidence in support of the reliability and validity of plans created by caseworkers based solely on clinical judgment. This is particularly concerning in a field such as child welfare, where the risks associated with classification errors are significant and the complexity of the issues surrounding the assessment of family functioning are great. To understand better the need for reliable and valid assessments in CWS, it is helpful to examine the role of the caseworker and client motivation more closely and to have a broad perspective on how to conceptualize child well-being. It is also important to have an understanding of the complexity and significance of the issues facing parents and children involved in CWS.

3.2 Caseworkers as Service Brokers

In addition to determining child safety, CWS caseworkers are also responsible for assessing service need as well as for referring clients to the appropriate service. How closely caseworkers are able to match the client's needs with the available resources will fundamentally influence the effectiveness of the interventions. The role of the caseworker in this scenario has been described as that of a service broker. In the service broker model, "the case manager refers clients to formal and informal care-giving

systems and monitors client's progress and delivery of services but does not directly provide services" (Rohland, Rohrer, & Tzou, 1998). Some studies on the influences of service brokers (Stiffman et al., 2001; Stiffman, Pescosolido, & Cabassa, 2004) have found that services brokers, or gateway providers as they are referred to in these articles, play a significant role in youth accessing mental health services. These researchers found that the greatest predictor of service provision was the service broker's perception of the youth's mental health needs, highlighting the importance of accurate assessment.

Another study, which examined CWS caseworkers' role as service brokers, implemented a training and consultation model to improve the capacity of the caseworkers to act as service brokers. One of the primary goals of the study was to ensure that CWS caseworkers who act as service brokers for youth in foster care have the necessary skills and training to identify mental health needs and make appropriate referrals (Kerns et al., 2010). The researchers provided training to CWS caseworkers to increase their understanding of common mental health needs of youth in foster care and to identify and classify the youths' mental health needs using the existing data. The trainings also focused on assigning youth based on their identified mental health needs to the appropriate and available evidence-based programs (EBP's). The trainings were followed up with four months of biweekly case consultations. The results indicated that the caseworkers reported and demonstrated increased knowledge of the available EBP's and increased ability to identify appropriate EBP referrals for particular mental health problems. However, when compared to the control group, caseworkers who went through the training did not refer significantly more clients to the EBP's than did the control group caseworkers (Dorsey et al., 2012). This finding seems to indicate that knowledge of family needs and of available resources may be necessary but not sufficient to ensure that families obtain appropriate services. It may be necessary to develop decision-making aids to help service brokers identify family needs and the resources to best match these needs (Stiffman et al., 2004). The need to accurately assess individual and family needs becomes even more pressing as the states increasingly push for the implementation of services that

have been found effective by previous research. These EBPs have been designed and shown effective for clients with particular problems, and caseworker's ability to accurately match clients to the appropriate services is fundamental to increase the rate of success.

The National Institutes of Health has provided a comprehensive road map to promote improvements in national public health system (Landsverk et al., 2011). The first two steps of this process include promoting basic research that informs the development of clinical interventions and testing these interventions in carefully controlled clinical trials, resulting in the development of EBPs. The third step outlined in the road map, which has historically received less attention, is the process associated with bringing the treatments to and testing their effectiveness in usual care settings. Focusing on this third step and a review of the literature, Fixsen et al. (2009) proposed a model outlining the core implementation components that work together to sustain the effective use of human service innovation, such as the implementation of EBP's, in usual care settings.

The authors identified seven interactive core implementation components: staff performance evaluation, decision support data systems, facilitative administrative supports, system interventions, recruitment and selection, pre-service training, and consultation and coaching. The system's intervention component of this model refers to the external systems supporting the implemented interventions. Fixsen et al. (2009) drew attention to the need to coordinate with "external systems to ensure the availability of the financial, organizational, and human resources required to support the work of the practitioners" (p. 535). Although not highlighted in their model, it also seems that one of the necessities of the external systems is to refer the appropriate clientele to the interventions. In their role as service brokers, caseworkers are the link between the client and service, and if the caseworker does not have accurate information to link the client to appropriate service, the result is likely to be frustrating for both the provider and the client. There is nothing provided in the Safety Assessment,

SDM tool, or the Family Assessment that would markedly enhance caseworker's ability to optimally match clients and services. This diminishes the utility of the assessments, as they are not providing caseworkers with the necessary information to do their job adequately. Furthermore, if the goal is to provide appropriate services to clients, then the lack of adequate problem identification could be seen as construct underrepresentation and affect negatively the evidence supporting the consequential validity of the assessment process. In other word, inadequate assessment may lead to clients not being referred to appropriate services, undermining their treatment progress.

The understanding of the fallibility of unaided judgment led to the implementation of actuarial tools to assist with decisions regarding which cases should be opened to receive services or closed. It is unclear why it is thought that clinical judgment would be adequate for determining immediate safety threats (in the case of the Safety Assessment) and for determining service need (in the case of the Family Assessment) and inadequate for predicting future maltreatment (the SDM tool). The same cognitive biases are present in each of these situations, and research appears to indicate that inaccuracies would be similar during each of these stages of assessment (Grove & Meehle, 1996; Nisbet & Ross, 1980; Oskamp, 1965; Tversky & Kahneman, 1974). It would seem that decisions made with the family might mitigate the bias; however, there is no indication that this would be the case and there are likely significant reliability issues related to both worker perception and family reporting. For example, just at the client level, it is likely that the power balance between the caseworker and the parents is such that the family members might try to please the caseworker and acquiesce to the caseworker's assessment or perhaps under-represent their challenges in order to get CWS out of their lives. These issues become particularly problematic when the caseworkers focus on and create plans to address their greatest perceived threat to child safety, which may not be what the family members would prioritize as their greatest need (Courtney, 2010).

3.3 Client Motivation and Standardized Assessments

According to SBC, client motivation can be enhanced when "the social worker is more interested in the client's ascribed meaning to the problem and his/her ideas about solutions than diagnostic categories and rote intervention schemes" (Christensen et al., 1999, p.13). The importance of client motivation to engage in services is well supported by research (Chaffin et al. 2009; Nock & Kazdin, 2005). However, how the use of standardized assessment tools reduces this motivation is less clear. The use of standardized assessment tools may result in increased motivation by increasing the precision and perceived objectivity of the assessment process. A recent example of a model that incorporates standardized tools and motivational interviewing techniques into a family assessment can be found in the Family Check-Up (FCU) intervention (Dishion et al., 2008; Gill, Hyde, Shaw, Dishon, & Wilson, 2008; Lunkenheimer et al. 2008).

FCU is a "brief, motivational intervention that supports parents' existing strengths as well as their engagement in additional parent training services when needed" (Lunkenhiemer et al., 2008, p. 1739). The intervention protocol includes three sessions. Parent-child interactions are recorded and coded during an initial meeting, during which caregivers also fill out several standardized questionnaires about their child as well as their own functioning. The second session involves an interview with the parent, focusing on the parents' concerns and on family issues that are most critical to the child's well-being. This information is also gathered from other important people in the child's life (e.g., teachers, relatives, and other care providers). The third session involves providing feedback to the parent about the results of the assessment. "An objective of the feedback session is to explore the parents' willingness to change problematic parenting practices, to support existing parenting strengths, and to identify services appropriate to the family need" (Dishion et al., 2008, p. 1402). Families are provided with an easy to read profile that indicates both family strengths and areas that need attention. The profile includes a report on child (temperament, behavior, language development), family well-being (daily hassles, emotional well-being), observed parent/child interaction (positive play, follows direction,

proactive parenting), and parent report of family support (social support, neighborhood resources). The process of changing the family dynamics follows a step-wise fashion, focusing first on issues of safety and security followed by behavioral management, parenting skills, and relationship building (Gill et al., 2008).

The model was tested with families who were receiving WIC services as well as facing other struggles (e.g., child behavior problems, maternal depression, substance-use problems). Families who were randomly assigned to FCU showed significant improvements in positive behavior support for children ages 2-3, which in turn promoted children's inhibitory control and language development measured when the children were 3-4 years old (Lunkenheimer et al., 2008). Although the target population and intervention goals were not identical to those of CWS, FCU presents a model where standardized tools can be used in concert with clinical judgment and motivational techniques to assess family functioning, promote family engagement, and provide insight into appropriate intervention. The use of standardized tools adds support to the reliability and validity of the assessment, offers insight into service need, and also provides a foundation on which future assessments can be compared and family change/progress monitored. These benefits could be seen either while the family is engaging in an intervention or if the family is encountered by the service providers in the future. Although an assessment of this sort may be beyond the skill level of caseworkers and may require the assistance of others, it still shows a promising direction for how assessments in CWS might be able to incorporate clinical impressions with standardized measures while also promoting family engagement.

An important distinction between the population served by FCU and that served by CWS is related to the nature of involvement. In FCU, as with most interventions, the client's involvement is voluntary and the client is looking for support. This increases the likelihood that the clients will engage in the assessment process in an honest manner; filling out the assessments and presenting in a manner that approximates their perception of their current circumstances. For CWS involved clients, this is

typically not the case. CWS involved clients are typically not volunteering for intervention and are often coerced into services with the threat of having the child(ren) removed from their care. As such, it seems reasonable to assume that they may be motivated to present in a manner that downplays their current struggles. A client reporting in this manner undermines the reliability of standardized assessments, for example, a client may be motivated to present a certain picture to the caseworker as opposed to honestly sharing about oneself and ones struggles (Converse et al., 2010). Models like SBC highlight the need to interact with clients in a manner that is presumed to decrease this type of defensiveness; however, it has not been demonstrated how this type of interaction influences disclosure of the current struggles of CWS involved clients. Further, it is not clear that clients reporting more openly in an unstructured format in and of itself leads to accurate assessment; if this were the case, one would need to question the value of standardized personality assessments in any situation.

3.4 Child Well-Being

Another concern regarding the current assessment system used by CWS in the state of Washington is the lack of focus on child well-being. If, after investigation, it is determined that the family needs ongoing services, this indicates that the family is presenting significant challenges for the child(ren) living in the home. This assessment is based on family history and current circumstances, which place the child(ren) at risk of future maltreatment in the immediate or near future. The current system focuses on the need to remove the safety threat(s) that is placing the child(ren) at risk and to close the case once the safety threat is removed. Additionally, but not a focus of this review, in CWS there is also a focus on permanency, with a goal of ensuring that children who were removed from their home are returned in a timely manner or placed relatively quickly into an adoptive home (Adoption and Safe Families Act, 1997). Safety and permanency are two of the elements of the federal law that directs priorities in CWS. These two priorities have become the focus of CWS, resulting in an outcome measure that focuses on tracking how long it takes children to gain permanency and on monitoring the re-

occurrence rate of child abuse/neglect. The third priority outlined in ASFA is that emphasis should also be placed on supporting the well-being of the children that CWS encounters. In recent years, efforts to ensure child well-being have focused primarily on assuring that children in out-of-home care attend school, receive medical and mental health care, and have their basic needs met in other ways. However, little attention has been paid to overall well-being of children who remain in their family home. This lack of attention to the overall well-being of children in their family home can be seen in the current assessment system that focuses on controlling safety threats but does little to determine what is needed to support child well-being.

Recently, the priorities of CWS have been challenged. Some of this criticism contends that CWS has prioritized agency-level risk management over the well-being of its clients. As Pithouse et al. (2011) argued, "Once a case has been accepted and investigation made, this too is to be completed within the time constraints imposed by a system that is as much about providing data to satisfy external surveillance as to processing a more reflective and considered response to individual need" (p. 174). The focus on permanence, measured by the length of time in care and number of moves while in care, and safety, as measured by reoccurrence of maltreatment, are easily measurable, and CWS employees at all levels can be held accountable for these numbers. However, they provide a very limited definition of what it means to support the most at-risk children in society, and in fact, they are proxy measures as opposed to focusing accountability directly on how the children are doing. This point is strongly stated by Mansell, Ota, Erasmus, and Marks (2011):

> Current preoccupation with risk and risk assessment seem not entirely motivated by concerns about the safety and wellbeing of children. They are just as likely to be driven by the need to provide mechanisms to ration resources and allocate responsibility and subsequent blame (usually to social workers) when true positives are missed and cases

> ‘go wrong’… Child protection systems are putting large numbers of families through an investigation process and, in doing so, are losing sight of children’s needs (p. 2007).

A recent Information Memorandum from Bryan Samuals (2012)(Commissioner of U.S. Department of Health and Human Services, Administration on Children, Youth and Families) challenges CWS in the United States to better meet the well-being of the children they serve. He noted that, “there is a growing body of evidence indicating that while ensuring the safety and achieving permanency are necessary to well-being, they are not sufficient” (p. 2). He suggested that a well-being framework should be adopted by CWS, which would focus on (a) cognitive functioning, (b) physical health and development, (c) behavioral/emotional functioning, and (d) social functioning. Although not explicitly stated in this memorandum, the push towards a more comprehensive focus on child well-being highlights a problem with the federal mandate that has separated safety, permanency, and well-being. A clearer and more accurate mandate might be that government agencies that focus on supporting/protecting vulnerable children should all be concerned predominantly with child well-being, with safety and permanency representing two areas related to child well-being.

This more comprehensive view of child well-being adds complexity and challenges CWS, and the complexity becomes even greater when there is an understanding that concern for child well-being must be both present and future oriented. As Ben-Arieh and Frones (2012) noted, “children’s rights [to well-being] refer both to their rights here and now, and to their rights to develop and ‘become’” (p. 463). Assessments focused on the here and now must accurately describe the current well-being of children while focusing on ‘becoming’ requires reliance on the predictive validity of the assessment. It seems that the current assessments used by CWS focus on one aspect of child well-being in the here and now: Can we call this child “safe” right now? However, the research as it relates to many children encountered by CWS would indicate that a shortsighted view of child safety does not promote adequate

care to encourage anything close to satisfactory development (this is further explored in the *child* section 3.7).

Due to the enormity of the challenges that CWS involved families face, assessments that focus only on present or impending safety threats cannot adequately promote long-term well-being of children who live in ongoing stressful and high-risk environments. If CWS hopes to make a difference in the developmental trajectories of children, assessments used to guide CWS intervention must be reliable and valid, include indicators of how the child is doing in their current environment (physically socially, emotionally, cognitively), provide guidance for the most appropriate interventions, and be able to measure if the current service plan is promoting change that is likely to lead to a more promising future.

Before looking at an alternative assessment system, it is helpful to first take a closer look at the parents and children involved in CWS.

3.5 Frequently Encountered Families

The majority of referrals to CWS are for child neglect. For example, in the state of Washington, 70.4% of the referrals are for neglect. Child neglect in the state of Washington includes an act or failure to act or the cumulative effects of a pattern of conduct, behavior, or inaction that evidences a serious disregard of consequences of such magnitude as to constitute a clear and present danger to the child's health, welfare, or safety (RCW 26.44.020). In the state of Washington, 20.2% of the referrals are for physical abuse, defined as non-accidental infliction of physical injury or physical mistreatment of a child, and 6.5% of the referrals are for sexual abuse, defined as committing a sexual offense or allowing a sexual offense to be committed against a child, as defined in the criminal code (Children's Administration Profile-Selected Demographic and Performance Measures Spanning 2004-2009, 2010).

These numbers are similar to the national numbers, which indicate that 78.3% of the referrals allege neglect, 17.6% physical abuse, and 9.2% sexual abuse (Child Abuse and Neglect Data System, 2010).

In general, those families investigated by CWS represent a population with numerous risk factors, and many of these factors are associated with living in poverty. For example, a recent survey done by Partners for Our Children (Courtney, 2010; Marcenko, Newby, Lee, Courtney, & Brennan, 2009) provides a wealth of information about the CWS clients in the state of Washington. The survey included 809 parents who had recent CWS involvement. The survey found that 47% of the parents reported a total household income of less than $10,000 a year, and 69% reported incomes of less than $20,000. Two-thirds of the respondents reported being unemployed. Furthermore, there were other indicators of significant poverty in the surveyed families. For example, in the last 12 months, 52% reported going to a food pantry or community meal program, 29% were homeless, 35% moved in with a family friend, and 26% had their utilities shut off. The economic struggles of parents involved in CWS have also been documented by other researchers, indicating that children who enter CWS are much poorer compared to children in the general population (Barth, Wildfire, & Green, 2006). Barth et al. (2006) found that about half of the children in their study lived in households with income below 50% of the poverty level.

As these numbers indicate, the population served by CWS is struggling with a great financial hardship, which appears to be closely associated with child neglect. Of particular concern are those families that are struggling with the stress of ongoing financial hardship and where the children are chronically maltreated by their caregivers. Studies have indicated that the extent and continuity of maltreatment contributes to the prediction of behavior and emotional trauma symptoms in children (Anda et al., 2006; English, Graham, Litrownik, Everson, & Bangdiwala, 2005; Perry et al., 1995). One detailed study done by Loman (2006) examined chronically maltreated children by investigating about 33,000 families referred to CWS in the state of Missouri (Loman, 2006). The researcher looked closely at families with four or more new referrals over the course of a five-year period; referred to as frequently

encountered families. Loman (2006) noted that although these families were referred primarily for neglect, there was also a significant amount of reported physical and sexual abuse. About 21% of the families in the study fit the criteria of being frequently encountered by CWS. Loman (2006) found that the frequently encountered families were almost twice as likely (20.8%) to be in severe financial difficulty (unable to pay for one or more basic necessities, such as rent, heat, light, food or clothing) as non-frequently encountered families (11.6%). It is also worth noting that during the 5-year follow up, about 37% of the frequently encountered families had at least one child placed in out-of-home care (Loman, 2006) in comparison to studies of the general population involved with CWS in which the placement rate was about 9.4% (Horwitz, Hurlburt, Cohen, Zhang, & Landsverk, 2011)

Loman (2006) also examined the cost of frequently encountered families to the CWS system. Although the frequently encountered families accounted only for about one fifth of the CWS involved families, when looking at costs associated with family services, foster care, and group care, they accounted for half of all expenditures to CWS involved families. The average cost of each frequently encountered family was about $13,000 over the course of 5 years. This figure does not include administrative, case management, or court-related costs associated with serving these families. Due to the risk of severe developmental consequences for children living in households in which chronic maltreatment is occurring and due to the high cost associated with serving these families, an assessment process with high support for its validity must be in place to provide guidance for the service provision to this population.

3.6 Caregiver struggles

In addition to struggling with poverty, the Partners for Our Children study also found that parents investigated by CWS face challenges stemming from their own childhood experiences. For example, 55% of those surveyed suffered sexual abuse as children (Marcenko, Newby, Lee, Courtney, & Brennan, 2009). This finding is consistent with a survey done in the state of Pennsylvania in which 54%

of the parents reported that they were involved in CWS when they were children (Child Welfare Education and Research Programs, 2011). Not surprisingly, in addition to poverty and negative childhood experiences, research has shown that caregivers involved in CWS struggle with a number of other challenges at a greater rate compared to the general population.

Mental Health

Caregivers involved in CWS experience increased rates of mental health struggles, particularly struggles related to depression. For example, 32% of families with recent investigations by CWS report taking medication for anxiety and depression (Child Welfare Education and Research Programs, 2011), with 45% suffering major depression within their lifetime (Marcenko, Newby, Lee, Courtney, & Brennan, 2009). An examination of National Survey of Child and Adolescent Well-being (NSCAW) data done by Kohl, Kagotho, and Dixon (2011) found that 21% of mothers questioned shortly after being investigated for child maltreatment reported meeting the diagnostic criteria for a major depressive episode, which compares to about 6.8% in the general population (Child Maltreatment, 2013). Kohl et al. (2011) also found that maternal depression was significantly related to child emotional maltreatment and child neglect. Looking at the same NSCAW data, Mustillo, Dorsey, Conover and Burns (2011) found that depression was directly associated with behavioral and emotional problems of all youth while for younger children, parental depression also indirectly influenced child outcomes through neglectful parenting. These findings are consistent with those by Chaffin, Bard, Hecht, and Silovsky (2011) who examined caregivers who had been referred to, and received services from CWS. They assessed a number of different domains and based on the results, they created different categories of caregivers. Caregivers in stable high category, that is, those who remained at a high-risk even after receiving services, were significantly depressed at a rate of 73%, while only 2% of those in the stable low group had significant scores on the depression measure. They also found a strong correlation between

number of previous reports of child maltreatment and caregiver depression, that is, those who were higher on depression had significantly more previous reports.

Domestic Violence

Many studies have shown a link between domestic violence (DV) and Child Welfare involvement (Hamby, Finkelhor, Turner, & Ormrod, 2010; Kohl, Barth, Hazen, & Landverk, 2005; Marcenko, Lyons, & Courtney 2011). One study examining over 3000 female caregivers who were a part of the NSCAW found that child welfare workers indicated that DV was present in 12% of the families investigated for maltreatment. In contrast, when items from the violence portion of the Conflict Tactics Scale (CTS) were administered to caregivers who had recent CWS involvement, 31% of the female caregivers reported domestic violence in the past year while 45% reported lifetime DV on this measure (Kohl et al., 2005). Furthermore, 19% of female caregivers in this study reported severe physical DV in the past year while 33% reported a lifetime prevalence of severe physical DV. Similar numbers were found in a study done by Marcenko, Lyons, and Courtney (2011). Based on the interviews with 747 parents with open CWS cases, they found that 35.8% of the female caregivers reported verbal threats, physical aggression, or physical injury in their most recent relationship. The rate was 31.8% for mothers whose children were in their care while the rate for mothers whose children had been placed was 38.8%. Furthermore, it has been found that high scores of physical assault on the Conflict Tactics Scale is associated with a significantly higher odds ratio of children being placed in out-of-home care (Horwitzet al., 2011).

Other researchers have focused on the relationship between DV and child neglect (Antle et al., 2007; Nicklas & Mackenzie. 2013). Antle et al. (2007) found that about 29% of the cases alleging child neglect also included DV. Nicklas and Mackenzie (2013) also investigated the connection between DV and child neglect and found that out of 151 mothers reporting at least some neglect, 38% reported earlier interpersonal violence. These authors pointed out that in addition to putting children at risk due

to the direct exposure to DV, the caregiver's emotional resources and ability to care for the child was also negatively impacted by the presence of DV.

Others who have found negative effects of DV on children involved with CWS support these findings. For example, Kernic et al. (2003) found that children who had been exposed to DV without concomitant child maltreatment were borderline or at clinical levels of internalizing behaviors and externalizing behaviors at rates of 21% and 29%, respectively. In comparison, those who were exposed to both maternal interpersonal violence and child maltreatment showed internalizing behavior problems at a rate of about 46% and a rate of externalizing behavior problems of 54%.

Substance Abuse

Studies examining the rates of substance abuse with CWS involved parents have reported varied findings. Young, Boles, and Otero (2007) conducted a literature review and found that reported rates of substance abuse ranged from 11% to 80% across studies. Based on their analysis of the various studies, they estimated that 11% of children who were victims of child maltreatment and received in-home services had parents who would have met criteria for substance use disorder. Further, they conclude that this rate was between 43% and 70% for the parents of children who were placed in out of home care. A recent study examining data from the NSCAW found that 12.5 % of parents indicated harmful use or dependence on alcohol and/or other drugs (Chuang, Wells, Belletiere, & Cross, 2013). Another study based on confidential interviews with caregivers who have been recently involved with CWS and assessed as high risk by the CWS caseworker post investigation found that 33.3% of the parents reported alcohol abuse (Proctor et al., 2012). This is consistent with findings from Marcenko et al. (2011) who found that about 30% of caregivers met criteria for alcohol or drug abuse/dependency in the past 12 months. Furthermore, when comparing the re-report rates for those families with no new reports with those categorized as continuously re-reported for child maltreatment, alcohol abuse was

associated with an increased odds ratio of 4.86 for being a member of the continuous re-reporting group (Proctor et al., 2012).

3.7 Child struggles

Many studies outline the adverse effect of child abuse and neglect on children's well-being and development (Bada et al., 2008; Bank & Burraston, 2001; Bellamy, 2008; Felitti et al. 1998; Halle, 2009; Perry, Pollard, Blakley, Baker, & Vigilante, 1995). Perry et al. (1995) described a process of how children form maladaptive behaviors and neural structures in the presence of stressful stimuli. The authors highlighted that because the developing brain organizes and internalizes new information in a use-dependent fashion, the more children are in a state of hyperarousal, the more likely they are to have neuropsychiatric symptoms following trauma, which over time can become maladaptive traits. Children in abusive and particularly in neglectful situations are often repeatedly exposed to stressful stimuli. For example, in a neglectful environment, an infant may be dependent on an under-responsive and depressed caregiver multiple times every day and have their basic needs unmet or met only marginally. Anda et al. (2006) reiterated this point, noting that the developing child is, "vulnerable to extreme, repetitive, or abnormal patterns of stress during critical or circumscribed periods of childhood brain development that can impair, often permanently, the activity of major neuroregulatory systems, with profound and lasting neurobehavioral consequences" (p. 174).

Many studies support this model, showing the negative effects of early, more long-term and repeated exposure on the child's functioning. In one study, researchers examined the effects of early abuse and/or neglect (ages birth to 2) versus later abuse and/or neglect (ages 4, 6, and 8) and later childhood aggression (Kotch et al., 2008). The researchers examined data of 1318 predominantly at-risk children. They found that particularly early neglect scores, prior to the age of two, were significantly predictive of higher aggression scores at ages 4, 6, and 8. In another study, Hussey et al. (2006)

examined the prevalence of child maltreatment in the United States and its effects on adolescent health. They examined surveys administered to 10,828 individuals longitudinally in 1995, 1996, and 2001-2002. The researchers found a significant relationship of family income with supervision neglect, physical neglect, and contact sexual abuse. Each type of maltreatment was associated with at least 8 of the 10 adolescent health outcomes examined. These health risk factors included outcomes such as fair/poor health, overweight status, depression, violence, cigarette, drug, and alcohol use. In another study, Bank et al. (2001) closely followed 182 boys on various developmental outcomes. Like the previously reviewed studies, they found an enduring and powerful effect of an abusive/neglectful home environment on later emotional and behavioral problems. These findings are also supported by those of Vance, Bowen, Fernandez, and Thompson (2002) who found that behavioral functioning in adolescents was related more to risk and protective factors in the child's environment than it was to children having prior psychiatric symptoms.

The previously mentioned parent survey done by Partners for Our Children in the state of Washington highlights a number of challenges facing the children involved in CWS. The survey investigated 2,382 children with an average age just under 9 years. Parents identified 737 of these children as having special needs (59% mental health conditions, 41% learning disabilities, 38% speech, hearing or vision problems, and 16% physical disabilities) (Marcenko et al., 2009). The results of a recent initiative in Pennsylvania that has promoted screening all 0-5 year old children encountered by CWS for developmental and emotional concerns are even more striking. Using the Ages and Stages Questionnaire (ASQ) and the Ages and Stages Questionnaire: Social-Emotional (ASQ-SE), the researchers found that 45% of the children aged 0-5 years old screened positive for a developmental or social-emotional concern (Child Welfare Education and Research Program, 2011). Furthermore, the researchers found no difference in prevalence between children that were investigated for being maltreated but not removed from their caregivers and children who were placed in foster care;

additionally, the rate of concern did not differ based on whether or not the allegations of maltreatment were substantiated. Although not examined in these studies, but based on the earlier mentioned research, it is likely that a disproportionate number of the children with developmental or emotional concerns come from the more chronic and higher risk families.

These findings are consistent with other findings based on large national samples. Examining data on 3,803 children ages 2-14 collected by the NSCAW, showed that the 47.9% of the youth who had been investigated by CWS had clinically significant emotional or behavioral problems (Burns et al., 2004). Furthermore, only one fourth of these youth received any special mental health care during the previous year. The study also found that youth in out-of-home care were at least twice as likely to receive mental health services, although it is not clear that their mental health needs were greater. Additionally, mental health service provision differed significantly by type of abuse rather than by the child showing clinically significant emotional or behavioral problems. For example, of those identified to be in the clinical range, 81% of the children who had been referred for sexual abuse received mental health services while only 19% of the physical abuse victims and 13% of the neglect victims received mental health services. However, although type of abuse did predict service provision, it did not correspondingly predict frequency of mental health problems; 52% of those reported for physical abuse scored in the significant range while 45% of those sexually abused and neglected scored in the significant range (Burns et al., 2004),

Other studies have also indicated caseworker struggles in identifying mental health problems in children. In another study looking at the NSCAW data, McCrae and Barth (2008) found that, “using worker indications of child mental health problems alone correctly classified just 48% of symptomatic children, no more than would be expected by chance” (p. 155). They noted that this translates to 470,000 children a year who may not be recognized by their caseworker as having clinically significant

levels of mental health concerns following an investigation. In an effort to increase identification, the authors tested a model where they simply looked at risk factors present in a family, which are typically available in the CWS database. These risk factors include such things as prior reports of maltreatment, poor parenting skills, active or history of domestic violence, and caregiver history of abuse or neglect as a child. It is important to note that these risk factors are similar to those included in the SDM for predicting future maltreatment. The authors noted that compared to caseworker judgment identifying 48% of the children, the cumulative risk model identified 73% of the children with mental health concerns. They also found that the model had somewhat low specificity (52%), but they argued that initial over identification is preferable to under identification. The authors contended that implementing universal screeners for all children with whom CWS comes into contact would be costly and suggested that the cumulative risk model may be a reasonable alternative for identifying children in need of further assessment. This is consistent with the English et al.'s (2005) contention that, "if resources are limited, those children with the most chronic maltreatment histories should receive services first" (p. 591).

Although chronicity of maltreatment seems to be a strong predictor of childhood developmental problems, it is important to keep in mind the heterogeneity of caregivers and children when determining service need. Some children seem more resilient and some home situations seem to provide supports that buffer some of the adverse consequences to children (Cicchetti, 2010). For example, a study done by Martinez-Torteya, Bogart, von Eye, and Levendosky (2009) examined protective factors and resilience in children exposed to domestic violence. The researchers examined 190 mother-child dyads at 2, 3, and 4 years of age. The researchers found that chronic exposure to domestic violence predicts the development of internalizing and externalizing behavior problems. The findings suggested that the experiences of children who are continuously under stress is qualitatively different from those of children exposed to intermittent domestic violence. However, the researchers

also found that certain factors, namely non-depressed mothers and child's easy temperament, emerged as significant buffers that predict the childhood resilience. Given the strong relationship between poverty and depression, one can assume that one of the better buffers, a non-depressed mother, is missing from many families involved with CWS (Halle, 2009; Knitzer, 2008; Wilson & Homer, 1995). However, these studies show a need to closely examine both the risk and protective factors in the family home.

As the above studies indicated, caregivers and children involved with CWS are facing a number of challenges, and if appropriate intervention is not implemented, the developmental trajectory of the children will be greatly compromised (Felitti et al. 1998; Hussey et al., 2006; Kotch et al., 2008). Further, a fundamental component of appropriate service provision is adequate assessment, which, as outlined earlier, is missing from CWS.

3.8 Assessment Needs in CWS

Halverson (1995) noted that assessments may include (a) screening and general disposition; (b) definitions, which may include diagnosis, labeling, or quantification of problem severity; (c) planning or matching treatment; (d) monitoring treatment progress; and (e) evaluating treatment outcome (p.23). CWS currently focuses on the first part of this model, particularly as it relates to one risk factor, namely the screening and identifying caregivers most at risk of maltreating their children. However, less attention has been paid to the rest of the model, and as explored in section 3.1, the current assessment tools that focus on identifying and monitoring the needs and progress of individuals provide little evidence supporting their reliability or validity. However, if an evidence-supported assessment system is not in place for each of the steps in the assessment model, there will be a lack of precision in assessing family needs, challenging the process of appropriately matching individuals with available treatments (Klein & Harden, 2011) and compromising the ability of the caseworker to monitor individuals' progress.

This lack of precision in assessment becomes particularly troublesome with the current emphasis on the implementation of evidence-based programs (EBPs). Evidence-based interventions have been designed and tested with individuals who have specific needs. For example, most of the evidence-based parenting programs have shown to be helpful for parents seeking treatment for their child with externalizing behavior problems (Eyberg, Nelson, & Boggs, 2008); additionally, some research has supported their effectiveness with parents who abuse their child(ren) physically (Chaffin et al., 2004). However, there is no evidence that these programs are effective for children who are neglected by their parents (Chaffin et al., 2004). These studies would indicate a need to closely match caregiver and child need for appropriate interventions; however, without reliable and valid assessments, this is not possible. Furthermore, if you do not have valid measures to determine the client's condition prior to starting the treatment, it will be difficult to know the effect of the intervention on the identified concerns, to determine whether the treatment plan needs adjustment, and to establish the treatment outcome. Samuels (2012) emphasized this point when he wrote that reliable and valid instruments for screening and assessing various aspects of social-emotional well-being of CWS involved clients is needed, and further that although child welfare staff may not be responsible for delivering interventions, "they should be able to appropriately assess and refer children and families to these evidence-based treatment providers and determine whether or not the interventions being delivered are having positive effects on child and family functioning" (p. 15).

3.9 The Comprehensive Assessment Program (CAP)

In the state of Washington, there has been one relatively small-scale attempt to introduce an assessment system that includes the use of standardized tools in determining family functioning. In 2009, a program was introduced to assist with the assessment of children in high-risk situations while they are still in their family home. The program, called the Comprehensive Assessment Program (CAP),

is administered to individuals referred by the assigned caseworker. The family receives a onetime assessment, and a report is written and provided to the CWS caseworker. The following is a description of the CAP.

(The following is an excerpt from the description on the program's website. It has been slightly modified for clarification purposes) [http://depts.washington.edu/hcsats/CAP/About_us.html].

> In 2007, the Governor's Office in the state of Washington asked the Children's Administration (CA) for recommendations to increase child safety. CA proposed a program that would improve the safety of children in their home by guiding decision-making and safety planning in complex, high-risk cases. CA proposed to make this program available at the front end of a case, when children were still in the care of their parents, with the overall goal to reduce recurrence of child maltreatment.
>
> The proposal was included in the Governor's budget and was funded by the legislature in a budget proviso, "solely to contract with medical professionals for comprehensive safety assessments of high risk families receiving in-home child protective services or family voluntary services. The safety assessments will use validated assessment tools to guide intervention decisions through the identification of additional safety and risk factors."
>
> On May 15, 2008, a planning meeting was held to develop recommendations on the structure and implementation of the Child Assessment Program (CAP). The planning team was comprised of CA regional staff, CA Policy and Practice Improvement, CA Indian Child Welfare, CA Practice Model, Harborview Medical Center, Office of the Ombudsman, and the University

of Washington Evidence Based Practice Institute. The planning team recommendations are included in the proposal below.

On July 10, 2008, CA Leadership approved that the CAP be administered through a contract with Harborview Medical Center. Community partners trained by Harborview would conduct assessments regionally.

Program Model Structure –Providers were approved to administer a systematic and standardized initial assessment that identifies problems and needs of parents and children. The program would provide baseline information on current family functioning.

Target Cases – Higher risk cases involving children aged birth to twelve who have complex issues, such as serious physical abuse, severe/chronic neglect, and sexual abuse. These are cases where the social worker is extremely concerned and is not sure how to proceed with service planning. Generally, CAP targets the needs of Family Voluntary Services overseeing high-risk, in-home cases.

Provider – The contract is held by Harborview Medical Center to employ and/or subcontract with trained Master's level staff to complete the assessments for referred clients. In addition, Harborview Medical Center provides supervision, support, and continuous quality assurance.

Tools – The assessment will be based on empirically identified problem areas and validated assessment tools such as:

- Brief Child Abuse Potential Inventory
- Parenting Stress Index
- Patient Health Questionnaire

- Pediatric Symptom Checklist
- Child Distress/Trauma Assessment
- Alcohol, Smoking and Substance Abuse Involvement Screening Test
- Conflict Tactics Scale
- Parents Evaluation of Developmental Status

Referral Process – CA staff will forward the following to the provider:

- DSHS Consent Form (DSHS 14-012)
- Current CPS referral including case notes
- Referral history for the family
- Safety assessment and plan
- Global Appraisal of Individual Needs - Short Screener (GAIN-SS) assessment
- Structured Decision Making (SDM) and Risk Assessment (if available)
- Family assessment (if available)
- Medical reports (if available)
- Child Protection Medical Consultation Report (if applicable)

Timelines – Families will be seen within five business days of the referral being made and the final report/findings will be submitted within 15 business days of the referral being made. A follow-up staffing occurs with CA staff and the family to review the results and identify any needed intervention and services.

Report – Short in length, with recommendations on safety/risks, can be used to inform service planning with the family.

The CAP report that is provided to CWS is typically 8-10 pages long and contains the following sections:

Sources of information: Lists various sources of information obtained and used in the assessment process.

CWS Intake and History: This section summarizes the current allegation and the family's history with CWS. This section also summarizes the results of the assessments that CWS caseworkers have done on the family as well as other records provided to the evaluator by CPS (medical records, school records…).

Family Situation: In a narrative form, based on an interview with the care providers, this section provides information on the family. This information might include family's current living situation, length of time parents have been together, parent's level of education, financial situation, parents own childhood history, mental health history, drug/alcohol abuse history, history with supportive services, and the like.

Family Perspective on the Intake: This section allows the parents to give their perspective regarding the most recent report made to CWS.

Observation of the Parent-Child Interaction: This section reports on observation of the child's behavior when in the proximity of a caregiver and of the responsiveness/interaction pattern between the caregiver and the child(ren).

Results of Standardized Measures: This section reports on the results of each of the standardized measures administered to each of the caregivers. The standardized measures administered to parents include the: Brief Child abuse Potential inventory, Pediatric Symptom Checklist-17, Parent Evaluation of

Developmental Status, Conflict Tactics Scale 2, Alcohol Smoking and Substance Involvement screening tool, Parent Health Questionairre-9, and the Parent Stress Index-Short Form.

Family Response to Measure Results and Service Options: Each of the standardized measures is reviewed with the caregivers. The measures are explained to the family and in this section, the parents' response to and elaborations on the results are reported. Service options are also discussed with the parents. The result of the Brief Child Abuse Potential Inventory is not initially provided to parents; instead, the results are discussed with parents during the follow-up session when the caseworker is present.

Barriers to Service: Potential barriers to service are outlined.

Impressions: This section outlines the evaluator's impressions and recommendations. This includes the evaluator's assessment of the caregiver's motivation to change and engage in services. This section also outlines any service recommendations and assists the evaluator in making placement recommendation if the evaluator felt the child(ren) needed to be removed from the home.

Follow-up: A follow-up staffing then occurs with CA staff and the family to review the results and identify any needed intervention and services.

3.10 Dissertation Focus

As can be seen, the CAP relies heavily on standardized measures but also includes family input, review of case records, and clinician input. In many ways, the CAP is similar to the programs like Family Check-Up (FCU) outlined in section 3.2 in that it uses standardized assessments to assess and engage families in action planning. However, the assessment also differs in some significant ways. In the case of the FCU assessment, the client is the family being assessed while the client of the CAP assessment is

primarily the state caseworker who is ultimately responsible for determining the services in which the family must engage. This has a number of significant implications. First, it is not known how the standardized assessment tools used during the CAP assessment will function with a population that is not engaging in the assessment process voluntarily. The coercive nature of the involvement with CWS may influence the caregiver's response to the assessment items, as it is likely that at least a portion of the caregivers will be motivated to respond in a manner to reduce the potential harm of CWS involvement in their lives. This motivation will likely influence the reliability and validity of the assessment tools. This is contrary to the process used to gather information from family members involved with CWS for the various studies reviewed in this chapter, as the assessment process in these various studies guaranteed the clients confidentiality (e.g., not sharing the results of the assessment with the caseworker) and often conducted the evaluations after the client's case had been closed with CWS. The assurance of confidentiality during the assessment process likely increases the reliability of the client's answers, resulting in assessment scores closer to the client's true score. However, a central question of this dissertation is, "How can standardized assessments be used to help in the assessment process of CWS involved clients?" As such, it is vital to closely examine how the caregivers respond to the standardized assessment tools during the CAP assessment.

Two studies are carried out in this dissertation in order to examine the use of the standardized tools used during the CAP assessment process and to assess whether the obtained information can be used to assist in the creation of a comprehensive assessment tool, with evidence supporting its reliability and validity. The first study examines the utility of various assessment measures with the CWS involved caregivers. The focus is on the assessment scale scores with an examination of whether these scale scores are in the range and function in a manner that would be predicted in a population as high risk as that referred to the CAP. This study also examines the evidence in support of the predictive validity of the various assessment tools. Throughout the study, efforts will be taken to closely examine

how honestly caregivers are responding to the measures. This is achieved through examining separately those who appear to be responding in a non-defensive and in a defensive manner, with a goal of seeing how this information may be used to further the understanding of the assessment of CWS involved clients. The specific research questions examined in Study 1 are presented in the methods section (Chapter 4).

The second study examines the data from a different perspective, with the goal of exploring how various assessment results might be used to create a self-report assessment for clients involved with CWS to better meet some of the goals of assessment outlined in this chapter. To accomplish this, Item Response Theory (IRT) will be used to examine the functioning of various items on selected assessment tools used by the CAP. Subsequently, the study explores how the results may be used to create a self-report assessment tool that not only assesses client struggles more adequately, but also is resistant to defensive (faking) behavior.

Chapter 4

Methods

4.1 Participants

Between 8/2009 and 7/2013, 348 families were referred to the CAP program. However, the CAP data base indicates that 98 of these families did not receive an assessment for various reasons (e.g., caseworker withdraw referral, family refused, child was placed in out-of-home care…). Additionally, even though the CAP referral criteria indicates that the child had to be in the family home at the time of the CAP assessment, 20 of the target children were placed in out-of-home care at the time of the assessment. Five of these children were returned home within 60 days of the assessment and were included in the analysis. The other 15 were not returned home within 60 days and were removed from the analysis. Consequently, the analyses in this dissertation are based on 235 families that were referred to and received a CAP assessment between 8/2009 and 7/2013. All families that were referred for a CAP assessment had an open case with CWS.

4.2 Data and Data Collection Process

After obtaining IRB approval from the Department of Health and Human Services, Harborview Medical Center who facilitates the CAP was contacted to provide the information they had acquired on each family. Harborview Medical Center maintains a client information system, which contains descriptive information on the participants as well as outcome information from the various assessments for each of the clients (see Table 4.1 for a list of the variables and section 4.3 for a description of the variables). Information obtained during the CAP assessment but not included in the client information system was not examined. For example, the final report, which included the results of interviews with caregivers and caseworkers, observation of parent-child interactions, and summary recommendations, was not examined. The researcher accessed only de-identified data while a spreadsheet that included identifying information was provided to three CWS employees. The CWS

employees then gathered the requested data from the CWS computer system by examining each client's electronic file and entering the data into a spread sheet (see Table 4.2 for the list of variables obtained from the CWS and section 4.4 for a full description of the variables). All information gathered by the CWS employees was then provided to this researcher to be matched with the de-identified CAP data.

Table 4.1
Information obtained from the CAP data base

Household information
Caregivers gender, date of birth*, and CWS case number
Target Childs gender, date of birth, and ethnicity*
Number of children in the family home
Type of maltreatment alleged in most recent report to CWS*
*This information was obtained from the CWS database if it was not included in the CAP database.
Results from the following assessment tools administered during the CAP assessment process
Parent Stress Index-Short Form (PSI) Including the subscales of: • Defensive Responding (score indicating if inventory was responded to in a defensive manner) • Parent Distress (PSI-PD) • Parent-Child Dysfunctional Interaction (PSI-PCDI) • Difficult Child (PSI-DC) • Total Stress (PSI-TS)
World Health Organization–Alcohol, Smoking and Substance Involvement Screening Test (ASSIST 3.0) *Substance Abuse Screener*
Patient Health Questionnaire (PHQ-9) *Depression Screener*
Conflict Tactics Scale (CTS-2) *Domestic Violence Screener*
Brief Child Abuse Potential Inventory (BCAP)* • Total score • Score indicating if the instrument was responded to in a valid manner. *Only parents where physical abuse was indicated were administered the BCAP
Parents Evaluation of Developmental Status (PEDS) -Used in cases were target child was Birth thru 3
Pediatric Symptom Check List – 17 (PSC17) Including the subscales of: • Internalizing • Attention • Externalizing • Total Score -Used in cases were target child was 4 and older

Table 4.2
Information obtained from the CWS data base

Structured Decision Making scores (SDM)-the SDM done in closest proximity to the CAP assessment was obtained and sub-scores in the following areas were recorded: Total scores for- • Neglect • Abuse Item scores for the following dichotomous questions (present/not present) on the SDM were also obtained: • Domestic violence in previous year (Question A6 in Appendix B) • Domestic violence prior to the previous year (Included on the SDM as a question, but not included in the scoring matrix) • Drug use in previous year (Question N8 in Appendix B) • Alcohol use in previous year (Question N8 in Appendix B)
Number of reports of child maltreatment on the family prior to the CAP assessment
Placement of the Target Child prior to the CAP assessment
Dates of reports of child maltreatment on the family after the CAP assessment
Dates of placement of the Target Child after the CAP assessment

4.3 Domains Assessed and Measures Used by the CAP

Child Development and Behavior

When a family was referred for a CAP assessment, a Target Child was identified. The Target Child was the child in the family home identified as being most at risk of maltreatment and/or being most challenging for the caregivers. The Target Child was the child of focus during the CAP assessment, and the child development/behavioral measures completed by the caregivers were done regarding this child. The child assessment tools included the Parents' Evaluation of Developmental Status (children birth thru 3) and the Pediatric Symptom Checklist-17 (children 4 and older).

Parents' Evaluation of Developmental Status (PEDS) (Used for children age 0-3)

The Parents' Evaluation of Developmental Status (PEDS) measures early childhood developmental and behavioral problems as reported by parents. Although the tool has been shown to be valid for children birth to 8 years of age, it has been used in the CAP to assess children birth thru 3 years while the PSC-17 has been used for older children. The PEDS consists of 10 questions written at

the average reading level of a 10-year-old. The first question is an open-ended question asking the parent to "list any concerns about his/her child's learning, development, and behavior". The last question is also an open-ended question asking the parent to list "any other concerns". The remaining 8 questions focus on different areas of development, asking the parent to indicate "no", "yes", or "a little" concern in each domain. The PEDS has shown to have high sensitivity in detecting children with disabilities, with detection rates ranging from 74-79% across age levels (Glascoe, 2000). However, concerns regarding the specificity of the PEDS have also been noted. Limbos and Joyce (2011) found the specificity of the PEDS to be only at 64%. The scoring of the items indicated by the parent on the PEDS depends on the child's age (e.g., question, Are you concerned with your child's behavior?, is scored differently for a 3 month old and a 3 year old). The scoring of the PEDS results in children being placed in one of 5 categories. Category A means two or more predictive concerns were noted (was scored a 3), category B means one predictive concern was noted (was scored a 2), category C means a non-predictive concern was noted (was scored a 1), category D means that a parent has difficulties communicating (e.g., a language barrier, this will not be scored, as this was not indicated on any of the assessments), and category E means no concerns were reported (was scored as 0). The scores were also dichotomized to indicate the number of children presenting with at least one predictive concern (category A and B). The dichotomized scores reporting the number of children for whom caregivers were indicating significant concern was used as covariates in the Cox proportional hazard model (see section 4.3 for details on the methods of analysis). The continuous scores were used to examine mean differences between groups and were used to examine correlations between measures.

Pediatric Symptom Checklist-17 (PSC-17) (Used for children age 4-17)

The Pediatric Symptom Checklist (PSC-17) is a parental report assessment used to assist with the screening of psychosocial problems in children (Gardner, Murphy, & Childs, 1999). The original PSC assessment includes 35 items that parents answer regarding their child's symptoms and behaviors.

Using factor analysis, Gardner et al. (1999) were able to shorten the form to 17 items scored on a Likert scale. This revised PSC includes subscales measuring internalizing, externalizing, and attention problems. Gardner et al. (1999) found the subscale for internalizing disorders to correlate with the Screen for Anxiety-Related Emotional Disorders (SCARED), the externalizing scale to correlate with the Iowa Conners aggression subscale, and the attention score to correlate with the Iowa Conners inattentive-over activity scale. The developers noted that the PSC-17 does not provide a diagnosis; hence, it should not be used to label a child. Instead, it should be used to determine whether further examination of the child and family is necessary.

In one study, the scores obtained on the PSC-17 were compared with scores obtained on a number of other measures, including the Child Behavior Check List (CBCL), The Screen for Child Anxiety Related Emotional Disorder (SCARED), The Children's Depression Inventory (CDI), and The Schedule for Affective Disorders and Schizophrenia for School-Aged Children-Present and Lifetime version (K-SADS) (Gardener, Lucas, Koloko, & Campo. 2007). Children in this study were recruited at a primary care clinic. The results indicated that 30% of the children had a positive PSC-17 total score, 15% obtained a positive score on the Attention subscale, 20% on the Externalizing subscale, and 36% on the Internalizing subscale. Based on the correlation with the other tools used in the study, the results supported the validity of the scores obtained on the PSC-17 as a screener. Gardner et al. (1999) reported that the PSC-17 has alpha coefficients of .79 for internalizing, .83 for externalizing, .83 for attention, and .89 for total score. The authors also reported that the PSC-17 has reasonable sensitivity (between 79% and 87% depending on the scale) and reasonable specificity (between 68% and 81%). Each scale also has a recommended cut score, which signifies a problem. A score of 5 or greater for internalizing problems, 7 or more for attention or externalizing problems, and a total score of 15 or greater indicate that the child should be referred for further assessment. The cut scores were used to report on the number of children for whom parents were reporting significant behavioral problems, and they were used as

covariates in the Cox proportional hazard model. The continuous scores on each of the scales were used to examine mean differences between groups and correlations between measures.

Domestic Violence

The Conflict Tactics Scale-2 (CTS-2) is used to identify domestic violence/intimate partner violence. The CTS-2 is the most widely used instrument for assessing intimate partner violence, with strong evidence supporting both the reliability and validity of the measure (Straus, Hamby, Boney-McCoy, & Sugarman, 1996). The CTS-2 is based on conflict theory, which assumes that conflict is an inevitable part of all human association; however, violence as a tactic to deal with conflict is not (Straus et al., 1996). The CTS-2 includes 5 subscales, physical assault, psychological aggression, negotiation, injury, and sexual coercion. The CAP assessment utilizes the physical assault and injury scales of the assessment. The developers of the CTS-2 indicated that it is acceptable to use selected scales when the length of assessment is an issue or when just particular types of conflict are of interest (Straus et al., 1996; Straus & Douglas, 2004). The abuse subscale consists of 12 physical assault questions. Five of these questions assess minor assault, such as "I pushed or shoved my partner", and 7 questions address severe assault, such as, "I beat my partner up". The injury scale includes 2 minor injury questions, such as, "I had a sprain, bruise or small cut because of a fight with my partner", and 4 severe injury items such as, "Had a broken bone from a fight with my partner". Each of the questions is also administered to the respondent as the perpetrator rather than the victim (e.g., "My partner has done this to me"). This resulted in 36 items scored on an 8-point scale indicating the frequency of behavior. Zero indicates never, 1 thru 6 indicate varying frequencies over the last year, and a 7 indicates that it has occurred but not in the previous year. The CTS-2 can be scored in various ways (Straus et al., 1996; Straus & Douglas, 2004; Straus, 2004). This study examined the prevalence of both any DV in the past year and any DV ever.

Drug and Alcohol Abuse

The World Health Organization - Alcohol, Smoking and Substance Involvement Screening Test (ASSIST 3.0) was used as a Drug/Alcohol screening tool. The ASSIST 3.0 has been shown to be significantly correlated to a number of other screeners, and it has shown to discriminate between substance use, abuse, and dependency (Humeniuk et al., 2008). The ASSIST 3.0 covers 10 substances: tobacco, alcohol, cannabis, cocaine, amphetamine-type stimulants, inhalants, sedatives, hallucinogens, opioids, and 'other drugs'. Clients receive a low, moderate, or high score in relation to each of the 10 substances as well as an overall score. The ASSIST 3.0 investigates the frequency of use and problems associated with each substance. The Tobacco section of the instrument was not used for the CAP assessment. For this study, those scoring as either "moderate" or "high" on any substance abuse were considered as indicated for substance abusing while those that scored "low" were considered as not indicated.

Depression

The CAP used the Patient Health Questionairre-9 (PHQ-9) as the depression measure. The PHQ-9 consists of 9 items measured on a Likert scale. The 9 questions of the PHQ-9 are closely aligned with the 9 diagnostic criteria for depressive disorder on the DSM-IV. The internal reliability of the PHQ-9 was reported as being 0.89, and test-retest reliability of the PHQ-9 was also high (Kroenke, Spitzer, Janet, & Williams, 2001). A previous study has supported the validity of PHQ-9 in making diagnoses and assessing severity of depressive disorders (Kroenke, Spitzer, Janet, & Williams, 2001). Other research has shown that the PHQ-9 may be useful to measure and monitor the outcomes of depression therapy (Titov et al., 2011). The 9 items of the PHQ-9 are scored on a four-point Likert scale, with a total of 27 possible points. Caregivers reporting minimal (score of 0-4) and mild depression (score of 5-9) were considered as not indicated while those scoring in the moderate (score of 10-14), moderate severe (a

score of 15-19), and severe range (20-27) were considered as indicated. The indicated versus not indicated scores were used in the analysis using the Cox proportional hazard model. The continuous measures were used to examine mean differences between groups and correlations between the various measures.

The next two measures, the Parent Stress Index and the Brief Child Abuse Potential Inventory, are not domain specific but rather take into consideration risk factors from various domains.

Parenting Stress Index-Short Form

Caregivers referred to the CAP were administered the Parenting Stress Index-Short Form (PSI-SF). Based on factor analyses of the Parent Stress Index, a shorter version, the Parenting Stress Index-Short Form (PSI-SF), was developed. The PSI-SF is a brief, 36-item self-report measure of parenting stress. Respondents use a 5-point Likert scale to indicate their level of agreement with each statement. The index provides a total score and three subscale scores, Parental Distress, Parent-Child Dysfunctional Interaction, and Difficult Child. Each subscale contains 12 items. Research on the PSI-SF has indicated that the Parent-Child Dysfunctional Interaction scale in combination with Difficult Child scale was a unique predictor of child abuse, whereas the Parent Distress scale was not a unique predictor (Haskett, Ahern, Ward, & Allaire, 2006). The Haskett et al.'s (2006) study, which examined both parents who had a documented history of physically abusing their child and parents who had no known history of abuse, found that parents with histories of abusing their child(ren) had a mean total score of 89.2 while the comparison group had a mean total score of 79.0. The researchers also found that 7.5% of all parents had scores indicating defensive responding, and these subjects were removed from further analysis in their study. There is also indication that Parent-Child Dysfunctional Interaction and Difficult Child scales correlate strongly with measures of child externalizing and internalizing behavior problems while Parent Distress correlates more strongly with measures of depression (Costa, Weems, Pellerin, & Dalton 2006;

Huth-Bocks & Hughes 2008). The alpha reliability coefficients of the PSI-SF have been shown to be .87 for Parental Distress, .80 for Parent/Child Dysfunction, .85 for Difficult Child, and .91 for total stress score (Kelley, 1998).

The PSI-SF includes a Defensive Responding Scale that indicates the degree to which the parent might be attempting to deny or minimize problems. This scale comprises 7 of the 12 items included in the Parent Distress scale. A total score on this measure of 10 or less represents a parent reporting at or below the 10th percentile and is indicative of defensive responding. The defensive scale was used to identify two groups of respondents, those whose score on this scale indicated defensiveness responding (Defensive group) and those for whom the score on this subscale did not indicate defensive responding (Non-Defensive group). These two groups were compared on all acquired data. It was anticipated that the results of this comparison could help provide insight into the performance of various assessments with CWS involved caregivers as well as inform the development of a new assessment for CWS involved caregivers that could potentially be resistant to defensive reporting.

Scores at or above the 85th percentile on the various scales of the PSI-SF are considered significant. This represents a score of 33 or higher on the Parent Distress scale, 26 or higher on the Parent-Child Dysfunction scale, 33 or higher on the Difficult Child scale, and a Total Stress score of 86 or higher. Summary data indicating the percentage of scores meeting the 85% on each of the scales are reported, and the indicated compared to not indicated conditions were used in the Cox proportional hazard model. The continuous score for each of the scales was used to both compare the mean scores of the caregivers and examine the correlations between various measures.

Brief Child Abuse Potential Inventory (BCAP)

The Child Abuse Potential Inventory is the most widely used and thoroughly researched measure of parental abuse risk (Ondersma, Chaffin, Mullins, & LeBreton, 2005). The Child Abuse

Potential Inventory is a 160-item self-report measure comprised of agree-disagree statements. Published studies on the Child Abuse Potential Inventory have indicated that it can be used to distinguish abusive from non-abusive parents and assess a range of difficulties associated with increased risk for physical abuse (Walker & Davies, 2010). However, the Child Abuse Potential Inventory takes about 15-20 minutes to complete. Due to the concern about the length of the survey, the Brief Child Abuse Inventory (BCAP) was created (Ondersma et al., 2005). The BCAP comprises 33 agree-disagree statements. Six of these items are validity items, three are random response items, and 24 items measure the construct of abuse potential. The BCAP is written at the fourth grade level and can be completed in less than 5 min. The BCAP contains selected items from the Child Abuse Potential inventory and maintains similar psychometric properties. The internal consistency reliability estimate of the BCAP is .89. Using factor analysis, it appears that the 24 items of the BCAP (33 items, excluding the 6 validity and 3 random response items) represent seven different factors (Ondersma et al., 2005). The 7 factors include distress, family conflict, parent rigidity, happiness, feelings of persecution, loneliness, and financial insecurity. The BCAP seems to do well at predicting future reports of both child abuse and neglect with general populations, but has not shown to be a good predictor of future behavior for those who had previously been reported for child abuse or neglect. The BCAP has been shown to correlate with measures of depression and parenting practices (Ondersma et al., 2005).

As mentioned, the validity measure on the BCAP consists of 6 items, when the respondent responds positively to more than three of these items the assessment is considered invalid. When administered to parents who had enrolled in some form of abuse treatment or prevention programs, 31.9% of the BCAP scores were determined to be invalid (Ondersma et al., 2005). The authors reported that the average abuse risk score was 5.2 in protocols deemed invalid and 9.2 in protocols deemed valid. It seems plausible that this apparent underreporting evidenced on the invalid BCAP scores is also present when caregivers complete other standardized measures. This concern is referenced in a study

by Chaffin et al. (2004), who noted, “current self-reported use of drugs or alcohol was vastly lower than lifetime prevalence reports, and current report levels were inversely correlated with scores on the CAP (Child Abuse Potential Inventory) Lie scale, suggesting that although lifetime prevalence was substantially endorsed, current use was not accurately reported” (p. 505).

Like the Defensive Reporting Measure of the PSI-SF, the BCAP was used in this study to examine reporting trends of those who have “valid” versus “invalid” scores. The CAP only administered the BCAP to families who had a history of allegations of physical abuse; consequently, only 121 caregivers received the BCAP. Due to the potential selection bias (situations were physical abuse is suspected) and the reduced number of participants receiving this assessment, the BCAP was not relied upon as heavily; rather the PSI Defensive scale was used to compare Defensive and Non-Defensive respondents. However, there will be some examination of the score differences between the valid and invalid BCAP results. It is also noteworthy that the BCAP used in the CAP program comprised 34 not 33 items, with the additional item being in the abuse potential score (i.e., not a validity item). All other items in the two versions of the BCAP are identical. The continuous score on the BCAP was used to compare the mean scores of the caregivers and examine the correlations between the various measures. For other analysis, as suggested by Ondersma et al. (2005), a cut score of 12 was used, with those scoring 12 or higher being considered as indicated. Using the cut score of 12, Walker and Davies (2012) found that 7.7% of the respondents from a community sample in England were considered as indicated on the BCAP.

4.3 Measures from Child Welfare Services (CWS)

Reports of child maltreatment

The CWS case attached to the primary caregiver identified in the CAP database was examined. The reports of child maltreatment linked to the caregivers’ case were included in this study if the report

alleged some form of child maltreatment by a caregiver that was accepted for some form of intervention. This includes reports that were accepted for an investigation of child maltreatment as well as those accepted for an alternative response, such as an intake worker calling the family to discuss the concerning information without initiating a formal investigation. All calls of concern received by CWS, regardless of whether they received a response, are documented in the computer system by the intake workers and attached to the family's case file; however, the cases that do not cross a pre-established threshold of maltreatment are entered into the system for information purposes only. Consequently, there are no criteria for the level of concern in the information only referrals; therefore, information only reports were not included in this study.

As the reports of child maltreatment included in this study were based on the primary caregiver's case file, this would include all reports on the family, regardless of whether or not the Target Child was involved in the allegations. Due to resource limitations, it was not possible to indicate which caregiver in the family home was alleged for what type of maltreatment (e.g., child abuse or child neglect) on each report. However, the type of maltreatment in the target report (i.e., the report received prior to and in closest proximity to the CAP assessment) was recorded. The types of maltreatment recorded included risk only, neglect, physical abuse, sexual abuse, or a combination of two or more abuse types.

When examining reports of child maltreatment, some researchers have focused on substantiated reports (Fluke, Yuan, & Edwards 1999; Solomon & Asberg, 2012) while others have used reports of maltreatment regardless of substantiation as the measure of interest. Kohl, Johnson-Reid, and Drake (2009) demonstrated that the risk of recidivism is similar for both substantiated and unsubstantiated cases. Furthermore, English, Marshall, Brummel, and Orme (1999) argued that many factors unrelated to whether or not a child has been abused or neglected influence the substantiation of

a report of child maltreatment. For example, work load, resources, office and individual worker practice, and standards of proof all influence the substantiation of allegations. Therefore, this study examined reports of maltreatment data regardless of substantiation. The numbers of reports of child maltreatment made prior to the CAP assessment were tallied, and the dates of all reports made after the CAP assessments were also recorded. The first report received after the CAP assessment was used in the Cox proportional hazard model.

Structured Decision Making (SDM)

Data from the SDM tool, discussed in detail earlier (Section 3.1.2 and also see Appendix 2 for a sample SDM), which was done by the caseworker in closest proximity to the CAP assessment, was examined. The overall risk score for neglect and the overall risk score for abuse were both recorded. These overall risk scores were separated into low, moderate, moderate high, and high categories. Per CWS policy and as indicated in the SDM instruction manual, scores in the moderate high and high range are considered at risk of future child maltreatment and should be offered ongoing services/case management. Although keeping the case open is not a requirement for those who score in the low and moderate range, the policy allows the caseworker to override the established criteria and monitor and provide services to these families as well. Based on these guidelines those cases that had a moderate high or high score on either the neglect or abuse scale were coded as indicated while those in the low and moderately range were coded as not indicated.

Two of the individual items that make up the SDM were also examined. The individual items selected from the SDM were those that lined up closest with the various CAP measures. Indication that the caregiver had a current drug and/or alcohol problem (item N8 on the SDM) was collected and compared with the results of the ASSIST 3.0. Moreover, indication of domestic violence was gathered and compared with the caregivers report on the CTS-2. Although the wording on the SDM items

differed from the comparable CAP tool, for example, the SDM item regarding domestic violence stated two or more incidences in the previous year, the items on the assessments in these areas seemed close enough to warrant comparison between caregiver report and caseworker report. Child-level data could not be compared between the various assessment tools as the SDM questions focused on child behavior/development of any child in the family home, consequently it was not possible to connect the child referenced on the SDM with the child assessed during the CAP assessment.

Placement of Target Child

As noted earlier, the CAP client information system lists one of the children in the home as the Target Child. This child would be the focus of the PEDS or the PSC-17. CWS case records were examined to determine the dates of placement in out-of-home care of this child as well as return dates of this child to the family home, both prior to and after the CAP assessment. Placements were included if the removal occurred through a court order, which typically occurs when the CWS caseworker petitions the court requesting authority to place the child in out-of-home care. Additionally, placements were recorded if the placement occurred under a voluntary placement agreement, which are typically obtained when it is determined that a child is not safe in the family home and the parent and caseworker agree that the child should be temporarily placed in out-of-home care. Placements were also included if Law Enforcement placed the child into the protective custody of CWS, resulting in the child being removed from the family home. Since all three of these methods of removal are based on risk to the child in the family home, a decision was made to include all three in this study. As indicated earlier, even though part of the criteria for the CAP program was that the child be in the family home at the time of the assessment, there were a number of instances in which the Target child had been placed in out-of-home care. This resulted in 15 families being removed from the study. Five families in which the Target Child was returned within 60 days of the CAP assessment were included in the study.

4.5 Study 1 Research Questions and Analytic Strategy

The first three questions focus on examining the CAP data and the data collected from CWS services separately. An additional 5 questions were examined that are based on a combination of the two sources of data.

1) What are the characteristics of the family members according to the various CAP assessment tools?

Descriptive data were examined for all family members involved in the CAP assessment process. Means were examined on each of the assessment tools as well as the percentage of family members meeting various cut score criteria on each of the assessments. The defensive responding sub-score on the PSI was used to compare the results of those indicated for defensiveness and those not indicated for defensiveness on each of the administered assessments. It was anticipated that the scores on the standardized assessment tools used in the CAP would be slightly lower than would be expected in the high risk population referred to the CAP program due to caregiver underreporting. However, when the Defensive and Non-Defensive groups are analyzed separately, it was anticipated that the scores of the Non-Defensive responders would be closer to what would be expected for this population while the scores of the Defensive responders would be significantly less on the various measures compared to what would be expected.

2) What percentage of families received new reports of child maltreatment or placement of the Target Child in out-of-home care and when did these incidents occur?

In general, studies on new reports of child maltreatment have indicated that new reports are most likely to occur soon after the reference report. For example, Fluke, Shusterman, Hollinshead, and Yuan (2008) found the 16% of children were re-reported for child maltreatment within one year of the reference report, and 22% were re-reported within two years. This rate is slightly lower compared to the rate of 28% found by English and Marshall (1999) over a 1.5-year period. English and Marshall

(1999) also found a much higher re-report rate for families with more prior reports of child maltreatment. For example, families with more than 4 prior reports had a re-report rate of about 40% over the 1.5-year period. Research on placement of children in out-of-home care has also shown a strong relationship between out-of-home placement and the families' frequency of historic contact with CWS. A study of a national database found that 9.4% of children investigated for maltreatment reports were placed in out-of-home care (Horwitz et al., 2011). The authors further noted that the risk of out-of-home placement varied greatly depending on the risk factors associated with the family. For example, the placement rate for children deemed to be at a lower risk level was only 3% over the three year study period while the placement rate for those children at a higher risk was 25%. Another study looking at families frequently encountered by the CWS found that 37% of the frequently encountered families had at least one child removed and placed in out-of-home care during the five year period examined (Loman, 2006). It was anticipated that the data from the CWS system would indicate that the families referred to the CAP are a high-risk group with a significant number of prior referrals and a high rate of prior placements of the Target Child. It was predicted that this population would have new reports of child maltreatment and/or placement of the Target Child at a rate consistent with previous studies that have examined families at high-risk for involvement with CWS.

3) Does the SDM score predict the rate and timing of new reports of child maltreatment or placement of the Target Child after the CAP assessment?

Survival analysis was done using the Cox proportional hazard regression program in SPSS-19. Time to either first new report of maltreatment or placement of the Target Child after the CAP assessment was entered in the time function. The time to event represented the outcome variable in the model (Kleinbaum & Klein, 2012). The event was also entered as either occurring (coded as 1) or not occurring (coded as 0). Time in the study was determined for each participant, with the study time ending either when an event occurs (i.e., new report or placement) or when the observation period

ends (in this case, when the data was pulled from the CWS system in October 2013). Thus, a strength of survival analysis is that it maximizes the use of available information. To examine the question related to the predictive validity of the SDM, the Cox proportional hazard model was utilized to compare those who scored as indicated or not indicated on the SDM. It was predicted that those classified as indicated on the SDM would have significantly higher rates of new reports of child maltreatment compared to those classified as not indicated on the SDM.

4) Is there a significant difference between the Defensive and Non-Defensive groups regarding a new event occurring (either re-report or placement of the Target Child)?

To examine whether those who scored in the Defensive group and Non-Defensive group differ in time to new incident, the two groups were entered as covariates into the Cox proportional hazard model. It was anticipated that the two groups would have similar time to new events, indicating similar risk in both groups.

5) Does the SDM predict future maltreatment differently for the Defensive and Non-Defensive groups?

To compare the Defensive and Non-Defensive groups in the functioning of the SDM, separate Cox proportional hazard models was run for the Defensive group and the Non-Defensive group while entering indicated/not indicated on the SDM as a covariate in the model. As the SDM is strongly influenced by history (e.g., number of previous referrals) and current non-modifiable factors (e.g., number of children in the home), it was anticipated that the SDM would function equally well for both Defensive and Non-Defensive responders.

6) Does the predictive validity of the modifiable variables (e.g., depression, child behavior problems…) measured during the CAP assessment add to the predictive validity of the SDM?

To examine the predictive value of the various CAP measures, each measure was entered into the Cox proportional hazard model as a covariate separately for the Defensive and the Non-Defensive responders. All these variables were then entered together with the SDM score using the Backward LR

function in SPSS to examine variables that significantly improve the predictive validity of the SDM. It was anticipated that those responding in a non-defensive manner could add insight into the current struggles they are facing; consequently, for the Non-Defensive group, at least some of the scores on the self-report measures could predict future child maltreatment as well as improve the predictive validity of the SDM. It was anticipated that the Defensive group might significantly underreport their problems; consequently, it was predicted that their scores on the various measures would not increase the predictive validity of the SDM.

7) Do any of the non-modifiable variables (e.g., caregiver age, child gender…) improve the predictive validity of the SDM?

The SDM includes a number of non-modifiable variables, including number and age of children in the home and number of previous referrals. To avoid duplicating any of these measures, only variables not included in the SDM were entered in the Cox proportional hazard model. These variables included ethnicity of the child, gender of the child, and previous placement of the Target Child in out-of-home care. It was anticipated that previous placement of the Target Child would indicate a high-risk family situation and that this variable would add significantly to the model for both the Defensive and Non-Defensive groups. It was not anticipated that the other non-modifiable variables would add significantly to the model.

8) What is the correlation between the caseworkers' assessment and the clients' report of the presence of domestic violence and substance abuse?

Previous research has indicated a discrepancy between client report and caseworker assessment of risk in these areas. For example, Kohl, Barth, Hazel, and Landsverk (2005) found that about 33% of female caregivers with recent CWS cases reported any DV; however, workers reported any DV only 12% of the time. Young, Boles, and Otero (2007) reported that caseworkers failed to identify substance abuse problems in 61% of caregivers who actually met the criteria. It was anticipated that

these types of discrepancy would also be apparent in this study, providing some indication as to why the caregivers' self-report on the standardized assessments adds to the predictive value of the SDM.

Chapter 5

Results of Study 1

5.1 Family Characteristics

Two hundred and thirty five families who completed the CAP assessment between 8-2009 and 7-2013 were included in this study. All families referred to the CAP had an open CWS case. When the CWS caseworker referred a family to the CAP program, a "Target Child" in the family was identified. The Target Child was typically considered the child in the family that was most at risk and/or most challenging for the caregiver(s). Table 5.1 presents the descriptive information for the 235 Target Children. The population included slightly more boys compared to girls (53.2% to 46.8%). The mean age of the children was 6.3, with a standard deviation of 4.8 years. The majority of the children fit within the CAP referral criteria of being between the ages of birth to 12; however, 22 children were 13 or older when referred to the program. The ethnicity of the children referred to the program was diverse, with the primary categories being Caucasian (42.6%), African American (17.4%), Native American (10.2%), and Multiracial (10.2%).

Table 5.1
Target Children referred to the Comprehensive Assessment Program (CAP)

Number of Children	**Gender**	**Age at CAP Assessment**	**Ethnicity**
235	53.2% Male 46.8% Female	Mean: 6.3 Std. Dev: 4.8 Median: 5.5 Range: 0-18 years* Quartiles: 0-2.1 years 2.1-5.5 years 5.5-10.5 Year 10.5 and over *Although one child had just turned 18 at the time of the CAP, the family was kept in the dataset, as there was number of other children in the home.	Caucasian: 42.6% African American: 17.4% Native American: 10.2% Multi Racial: 10.2% Hispanic/Latino/Mexican: 6.8% Asian: 3.0% Native Hawaiian/Pacific Islander: .9% Other/Unable to determine: 8.9%

Concerning the caregivers, when referred to the CAP program, a primary caregiver was identified and assessed, but in 83 cases, a secondary caregiver was also assessed. More information was available on the primary caregiver than on the secondary caregiver. For instance, on a number of occasions, the secondary caregiver was not named in the CAP database. Hence, it was difficult to match the secondary caregiver with the SDM assessment done on the family. Consequently, data from all caregivers (primary and secondary) were used to examine correlations between the various CAP measures and percentages meeting cut scores on these measures; however, only CAP data on the primary caregiver was matched to the CWS assessment (the SDM). Moreover, only the primary caregiver's case was examined to determine both incidents of reports to CWS and placement of the Target Child by CWS. Table 5.2 lists the descriptive data on the caregivers. The primary caregivers were predominantly female (91.9%), with the mean age of 31.9. For secondary caregivers whose gender could be determined, 62.7% were male. Information was not available to determine secondary caregivers' age. Additionally, as indicated in Table 5.2, each family had on average 2.37 children.

Table 5.2
Caregiver Information

Number of Primary Caregivers	**Gender**	**Caregiver Age at CAP Assessment**	**Number of Children in the Family Home**
235	91.9% Female 8.1% Male	Mean: 31.9 Std. Dev: 9.3 Minimum: 16 Maximum: 70	Mean: 2.37 Std. Dev: 1.3 Minimum: 1 Maximum: 8
Number of Secondary Caregiver Assessed by CAP			
83 (*Total of 318 caregiver*)	4.8% Female 62.7% Male 32.5% Missing		

5.2 CAP Data

Table 5.3 lists various measures administered during the CAP assessment and the number of caregivers completing each measure. All caregivers completed the PHQ-9, and nearly all caregivers completed the PSI-SF, ASSIST 3.0, and CTS-2. The BCAP was only administered in cases with a history of physical abuse or when physical abuse was suspected. Consequently, only 121 BCAPs were administered. The PEDS evaluation was used with children birth thru 3, while the PSC-17 was used with children 4 years of age and older. Hence, the numbers of caregivers completing these two measures decreased. Because nearly all caregivers completed four of the measures and the other measures contained data that were missing systematically, listwise deletion was used to address missing data in all analysis. The demographics and descriptive statistics for those that completed the PSI-SF, ASSIST 3.0 and CTS-2 were nearly identical to the total population (Table 5.1 and 5.2). Whereas more boys (53.2%) than girls (46.8%) were the target children for the total study population, the children whose caregivers completed the BCAP (n = 87 families), there were slightly more girl's (52.9%) than boy's (47.1%). The PEDS was completed on 87 children, with the percentage being more boys (60.9%) compared to girls (39.1%). Also, the primary caregivers who completed the PEDS were slightly younger than the total population (mean of 27.3 compared to 31.9 years), and there were fewer children in the home (mean of 1.86 compared to 2.37). The PSC-17 was completed on 143 children (51% girls and 49% boys). The primary caregivers who completed the PSC-17 were slightly older than the total population (mean = 34.72 years) and had more children in the family home (mean = 2.68).

Table .5.3
Measures from the CAP

Assessment Tool Name and Sub-Scores of the Assessments	**Number of Caregivers Completing the Measure**
Parent Stress Index-Short Form (PSI-SF) • Defensive Responding (PSI-Def) (score indicating if inventory was responded to in a defensive manner) • Parent Distress (PSI-PD) • Parent-Child Dysfunctional Interaction (PSI-PCD) • Difficult Child (PSI-DC) • Total Stress (PSI-TS)	315
World Health Organization –ASSIST 3.0 *Substance Abuse Screener*	312
Patient Health Questionnaire (PHQ-9) *Depression Screener*	318
Conflict Tactics Scale (CTS-2) *Domestic Violence Screener*	316
Brief Child Abuse Potential Inventory (BCAP) • Total score • Score indicating if obtained scores are valid -Only parents where physical abuse was indicated were administered the BCAP	121
Parents Evaluation of Developmental Status (PEDS) Used in cases were target child was Birth thru 3	122
Pediatric Symptom Check List – 17 (PSC-17) • Internalizing • Attention • Externalizing • Total Score Used in cases were target child was 4 and older	187

As one of the goals of assessing the data was to examine the effect of defensive responding, the results from the various assessments are reported for the entire population as well as for those who scored non-defensively and defensively on this PSI-SF subscale. Since only 121 caregivers completed the BCAP, limited analysis was done contrasting those that completed the BCAP in a valid compared to an invalid manner. However, it is noteworthy that while about a quarter of those who completed the PSI-SF was classified as being defensive (24.1% defensive and 75.9% non-defensive), over half of those who completed the BCAP had invalid scores (53.7% invalid and 46.3% valid). This indicates that the validity score on the BCAP may be more sensitive to misleading responses than the PSI-SF defensiveness scale.

The caregiver characteristics for the defensive and non-defensive responders were similar. The caregivers in the Defensive group had an average age of 31.29 and had an average of 2.26 children in the home. The average age for the Non-Defensive caregivers was 32.01 and they had an average of 2.39 children in the home. The data for the Target Child was also similar for the two groups. The target children of the Defensive caregivers were 46.1% female and had an average age of 5.25. While the target children of the Non-Defensive caregivers were 48.5% female and had an average age of 6.35.

Table 5.4 shows the scores for the entire group on those CAP measures that could be measured as a continuous variable. Table 5.5 shows a comparison between the Non-Defensive and Defensive groups on these measures. As can be seen in the Table 5.5, the means of the scores on all measures were significantly different between the Defensive and Non-Defensive groups, with the exception of PSC17-Attention score. It is important to note that the sample sizes differed between the two groups, with 239 in the Non-Defensive group compared to 76 in the Defensive group. The Levene statistic, which does not assume homogeneity of the variance, is also included in Table 5.5. A significant Levene statistic would indicate that the standard deviation between the two groups is significant; thus, the second value in the significance column should be applied.

Table 5.4
Scores for the Total Sample on the various CAP Measures

	N	Mean	Std. Deviation
PHQ (Depression)	318	5.50	6.368
BCAP Score	121	5.17	5.574
PSC17 Attention	187	4.72	2.946
PSC17 Internalizing	187	2.93	2.591
PSC17 Externalizing	187	5.82	4.244
PSC17 Total	187	13.46	8.254
PEDS Score	123	.83	1.046
PSI Parent Distress	315	23.98	8.220
PSI Parent Child Dysfunction	315	21.82	8.821
PSI Difficult Child	315	27.60	10.850
PSI Total Stress	315	73.40	23.792

Table 5.5

Comparison between the Non-Defensive and Defensive Caregivers on Various CAP Measures

		N	Mean	Std. Deviation	Std. Error Mean	Levene Statistic[1]		t	df	Sig. (2-tailed)
						F	Sig.			
PHQ -Depression	Non-Defensive	239	6.57	6.746	.436	41.076	.000	5.551	313	.000
	Defensive	76	2.11	3.349	.384			7.678	258.030	.000*
BCAP (using all BCAP scores)	Non-Defensive	90	6.18	5.988	.631	24.498	.000	3.567	119	.001
	Defensive	31	2.23	2.432	.437			5.149	115.842	.000*
BCAP[2] (using just valid BCAP scores)	Non-Defensive	45	9.49	6.159	.918	9.431	.003	3.004	54	.004
	Defensive	11	3.73	2.936	.885			4.518	34.121	.000*
PSI Parent Distress	Non-Defensive	239	26.82	7.306	.473	46.217	.000	13.736	313	.000
	Defensive	76	15.05	2.668	.306			20.893	307.738	.000*
PSI Parent/Child Dysfunction	Non-Defensive	239	23.35	8.679	.561	9.384	.002	5.722	313	.000
	Defensive	76	17.02	7.475	.857			6.180	144.695	.000*
PSI Difficult Child	Non-Defensive	239	29.64	10.633	.688	7.095	.008	6.272	313	.000
	Defensive	76	21.18	8.890	1.020			6.878	149.044	.000*
PSI Total Stress	Non-Defensive	239	79.81	22.172	1.434	13.204	.000	9.637	313	.000
	Defensive	76	53.25	16.359	1.876			11.245	169.947	.000*
PSC17-Attention	Non-Defensive	146	4.92	2.839	.235	2.064	.152	1.613	184	.109
	Defensive	40	4.08	3.238	.512			1.496	56.495	.140
PSC17-Internalizing	Non-Defensive	146	3.18	2.631	.218	3.073	.081	2.827	184	.005*
	Defensive	40	1.90	2.134	.337			3.183	74.747	.002
PSC17-Externalizing	Non-Defensive	146	6.23	4.110	.340	.021	885	2.376	184	.019*
	Defensive	40	4.45	4.446	.706			2.266	58.354	.027
PSC17-Total	Non-Defensive	146	14.32	8.065	.667	.033	.856	2.685	184	.008*
	Defensive	40	10.43	8.385	1.326			2.625	60.238	.011
PEDS Score	Non-Defensive	85	1.02	1.091	.118	9.114	.003	3.164	118	.002
	Defensive	35	.37	.770	.130			3.650	88.820	.000*

Independent Sample T-test between those who had non-defensive and defensive PSI scores.

[1]Test of homogeneity of the variance, if significant the second row in the significance column is more appropriate as indicated by the *, as this measure does not assume equal variance.

[2]BCAP mean for all valid scores was 8.36 (n = 56) while the mean for invalid scores was 2.42 (n = 65).

Table 5.6 shows the correlations between various CAP measures listed in Table 5.5. The table is split between those responding in a non-defensive manner, shown in the lower left portion of the table, and those responding in a defensive manner, shown in the upper right portion of the table. In addition to showing the strength of the correlations between various measures, the bottom score on the lower left side of the table also includes Fisher ʐ' scores. The Fisher ʐ' score is being used to compare the strength of the correlation between the Non-Defensive and Defensive groups. As indicated in the table, the correlations are more frequent and significantly stronger for the Non-Defensive group.

Table 5.7 presents the results using cut scores, as opposed to continuous measures and consequently, the measures of domestic violence (CTS-2) and the drug/alcohol screener (ASSIST 3.0) were included. The CTS-2 was classified as indicated if the caregiver reported that domestic violence was present in the last year or present ever. Drug/alcohol was classified as indicated if the risk on the ASSIST 3.0 was moderate or high. The PEDS was classified as indicated if one or more predictive concern was reported. The instruments established cut scores were used to classify all other measures (outlined in section 4.3). Table 5.7 presents the number and percentage of caregivers of the total study sample classified as indicated and not indicated as well as reports these numbers separately for those categorized as non-defensive or defensive on the PSI. A Pearson Chi Square test was used to determine whether the difference between the Non-Defensive and Defensive groups were statistically significant. The results signified that on all measures, the Non-Defensive group reported a higher percentage of concerns in the indicated category, and many of these differences were statistically significant.

Table 5.6

Correlations between the Non-Defensive and Defensive Groups on Various CAP Measures.

		PHQ Depression	BCAP Score	PSI Parent Distress	PSI Parent Child Dysfunction	PSI Diff. Child	PSI Total Stress	PSC17 Attention	PSC17 Internalizing	PSC17 Externalizing	PSC17 Total	PEDS Score
PHQ Depression	Correlation		*.238*	*.087*	*-.079*	*-.031*	*-.038*	*-.024*	*-.087*	*.006*	*-.028*	*.210*
	N		*31*	*76*	*76*	*76*	*76*	*76*	*40*	*40*	*40*	*35*
BCAP Scores	Correlation	.663**		*.189*	*-.291*	*-.291*	*-.266*	*.105*	*.138*	*-.165*	*-.029*	*-.454*
	N	90		*31*	*31*	*31*	*31*	*21*	*21*	*21*	*21*	*9*
	Fisher z'	2.56**										
PSI-Parent Distress	Correlation	.610**	.670**		*.122*	*.220*	*.339**	*.096*	*-.090*	*.211*	*.127*	*.157*
	N	239	90		*76*	*76*	*76*	*40*	*40*	*40*	*40*	*35*
	Fisher z'	4.64**	2.85**.									
PSI-Parent/Child Dysfunction	Correlation	.309**	.284**	.374**		*.830**	*.928**	*.379**	*.308*	*.761**	*.630**	*.204*
	N	239	90	239		*76*	*76*	*40*	*40*	*40*	*40*	*35*
	Fisher z'	2.98**	2.72**	2.02*								
PSI-Difficult Child	Correlation	.325**	.312**	.399**	.761**		*.958**	*.469**	*.331*	*.833**	*.709**	*.379*
	N	239	90	239	239		*76*	*40*	*40*	*40*	*40*	*35*
	Fisher z'	2.75**	2.86**	1.48	-1.42							
PSI-Total Stress	Correlation	.478**	.500**	.667**	.880**	.909**		*.443**	*.313*	*.833**	*.694**	*.314*
	N	239	90	239	239	239		*40*	*40*	*40*	*40*	*35*
	Fisher z'	4.17**	3.78**	3.38**	-2.0*	-2.98**						
PSC17-Attention	Correlation	.329**	.149	.286**	.523**	.633**	.606**		*.628**	*.567**	*.848**	*.a*
	N	146	69	146	146	146	146		*40*	*40*	*40*	*0*
	Fisher z'	1.98*	0.17	1.07	0.98	1.29	1.23					
PSC17-Internalizing	Correlation	.489**	.383**	.423**	.503**	.565**	.609**	.462**		*.538**	*.783**	*.a*
	N	146	69	146	146	146	146	146		*40*	*40*	*0*
	Fisher z'	3.37**	1.0	2.94**	1.27*	1.61	2.08*	-1.29				
PSC17-Externalizing	Correlation	.206*	.101	.259**	.642**	.713**	.685**	.652**	.512**		*.888**	*.a*
	N	146	69	146	146	146	146	146	146		*40*	*0*
	Fisher z'	1.10	1.01	0.28	-1.29	-1.65	-1.95*	0.74	-0.19			
PSC17-Total	Correlation	.380**	.231	.371**	.675**	.770**	.761**	.835**	.749**	.906**		*.a*
	N	146	69	146	146	146	146	146	146	146		*0*
	Fisher z'	2.32*	0.99	1.42	0.43	0.73	0.78	-0.24	-0.45	0.5		
PEDS	Correlation	.180	.120	.167	.293**	.483**	.390**	.a	.a	.a	.a	
	N	86	21	86	86	86	86	0	0	0	0	
	Fisher z'	-0.15	1.29	0.05	0.46	0.73	0.42	N/A	N/A	N/A	N/A	

The Defensive group is represented in the upper right portion of the table, and the Non-Defensive group in the lower left portion of the table. *Indicates correlations significant at the .05 level, and **indicates significance at the .01 level (2-tailed). Fisher z' score is presented in the lower left portion of the table and indicates if the correlations in the given cells are significantly different between the Non-Defensive and Defensive groups.

Table 5.7
Comparison using Cut Scores

		Total Population	Not Defensive on PSI	Defensive on PSI
PSI-Parent Distress	Not Indicated Indicated	*n* = 270 85.7% *n* = 45 14.3%	*n* = 194 81.2% *n* = 45 18.8%**	*n* = 76 100% *n* = 0 0%
PSI-Parent/Child Dysfunction	Not Indicated Indicated	*n* = 230 73% *n* = 85 27%	*n* = 158 66.1% *n* = 81 33.9%**	*n* = 72 94.7% *n* = 4 5.3%
PSI-Difficult Child	Not Indicated Indicated	*n* = 223 70.1% *n* = 92 28.9%	*n* = 154 64.4% *n*= 85 35.6%**	*n* = 69 90.8% *n* = 7 9.2%
PSI-Total Stress	Not Indicated Indicated	*n* = 233 74% *n* = 82 26%	*n* = 160 66.9% *n* = 79 33.1%**	*n* = 73 96.1% *n* = 3 3.9%
BCAP (Using all BCAP scores)	Not Indicated Indicated	*n* = 101 83.5% *n* = 20 16.5%	*n* = 70 77.8% *n* = 20 22.2%**	*n* = 31 94% *n* = 2 6%
BCAP (Using Valid BCAP scores)	Not Indicated Indicated	*n* = 37 66.1% *n* = 19 33.9%	*n* = 26 57.8% *n* = 19 42.2%**	*n* = 11 100% *n* = 0 0%
PHQ (Depression)	Not Indicated Indicated	*n* = 249 78.3% *n* = 69 21.7%	*n* =174 72.8% *n* = 65 27.2%**	*n* = 73 96.1% *n* = 3 3.9%
CTS-2 Domestic Violence-Previous Year	Not Indicated Indicated	*n* = 261 82.1% *n* = 55 17.3%	*n* = 192 81% *n* = 45 19%	*n* = 66 86.8% *n* = 10 13.2%
CTS-2 Domestic Violence-Ever	Not Indicated Indicated	*n* = 155 49.1% *n* = 161 50.9%	*n* = 107 45.1% *n* = 130 54.9%*	*n* = 46 60.5% *n* = 30 39.5%
ASSIST 3.0 (Drug/Alcohol)	Not Indicated Indicated	*n* = 225 72.1% *n* = 87 27.9%	*n* = 165 70.5% *n* = 69 29.5%	*n* = 58 77.3% *n* = 17 22.7%
PSC-Attention	Not Indicated Indicated	*n* = 123 65.8% *n* = 64 34.2%	*n* = 94 64.4% *n* = 52 35.6%	*n* = 28 70% *n* = 12 30%
PSC-Internalizing	Not Indicated Indicated	*n* = 140 74.9% *n* = 47 25.1%	*n* = 105 71.9% *n* = 41 28.1%	*n* = 35 87.56% *n* = 5 12.5%
PSC-Externalizing	Not Indicated Indicated	*n* = 113 60.4% *n* = 74 39.6%	*n* = 80 54.8% *n* = 66 45.2%**	*n* = 32 80% *n* = 8 20%
PSC-Total	Not Indicated Indicated	*n* = 112 59.9% *n* = 75 40.1%	*n* = 82 56.2% *n* = 64 43.8%	*n* = 29 72.5% *n* = 11 27.5%
PEDS	Not Indicated Indicated	*n* = 87 70.7% *n* = 36 29.3%	*n* = 57 66.3% *n* = 29 33.7%	*n* = 29 82.9% *n* = 6 17.1%

*Indicates significant Pearson Chi Square between the Non-Defensive and Defensive PSI reporting groups at the 0.05 level using two sided significance test. ** Indicates significant Pearson Chi Square test between the Non-Defensive and Defensive PSI reporting groups at the 0.01 level using two sided significance test.

5.3 Child Welfare Service (CWS) Data

Table 4.2 outlines the data obtained from the CWS system. The measures gathered included scores for the identified primary caregiver obtained on the Structured Decision Making assessment (SDM). Results from the SDM completed in closest proximity to the CAP assessment were examined.

Seventy five percent of the SDMs examined were done prior to the CAP assessment program and 25% were done after. The median number of days between the time of the CAP assessment and the time of SDM completion was 27; however, the SDM was completed on average 67 days, with a standard deviation of 85 days, from the time of the CAP assessment, indicating considerable variability in its proximity to the CAP. The information gathered from the SDM included the total neglect and total abuse score. Consistent with CWS policy as well as the SDM scoring manual, individuals with low and moderate scores were considered not indicated for future maltreatment, and individuals with moderate high and high scores were considered as indicated for being at risk of future reports of child maltreatment. Each family was then given a total SDM score as either indicated or not indicated based on both the neglect and abuse score. If they were indicated on either/both the abuse and/or neglect score, then they were considered indicated. As shown in Table 5.8, 62 cases (26.4%) were not indicated on either the abuse or neglect scale while 167 cases (71.1%), were indicated on at least one of the scales (2.6% of the cases did not have an SDM score).

Table 5.8
Summary of CWS Data

Structured Decision Making (SDM) Score			
	Neglect Score	Abuse Score	Indicated on either Abuse/Neglect Score
Low	Frequency: 12 5.1%	Frequency: 47 20%	Low or Moderate (Not Indicated): Frequency: 62 26.4%
Moderate	Frequency: 63 26.8%	Frequency: 112 47.7%	
Moderate High	Frequency: 130 55.3%	Frequency: 51 21.7%	Moderate High or High (Indicated): Frequency: 167 71.1%
High	Frequency: 24 10.2%	Frequency: 19 8.1%	
Missing	2.6%	2.6%	2.6%
Reporting History of Child Maltreatment Related to Family Cases Prior to CAP Assessment:		Mean: 5.66 Sd. Dev: 4.00 Median: 4 Range: 1-28	
Placement of Target Child at Some Point Prior to the CAP Assessment.		165 not placed prior to the CAP assessment (70.2%) 70 placed some time prior to the CAP assessment (29.8%)	

In addition to the total SDM scores, the data was also collected on the SDM items indicating alcohol and/or drug use in the previous year as well as domestic violence in the previous year or domestic violence ever. These items were selected because they matched the corresponding CAP measures of similar domains (ASSIST 3.0, CTS-2). Table 5.9 presents the rate with which caseworkers alleged primary caregivers struggled in these two domains. About 33% of caseworkers alleged that the primary caregiver was struggling with current drug/alcohol use while 11.3% of caseworkers alleged that caregivers were struggling with current domestic violence.

Table 5.9
Caseworker Assessment of the Presence of Client Struggle in Various Domains as Indicated on the SDM

	Not Indicated	**Indicate**
Drug/Alcohol Use	150 or 67%	74 or 33%
D/V Previous Year	197 or 88.7%	25 or 11.3%
D/V Prior to Previous Year	176 or 79.3%	46 or 20.7%

Other information collected from the CWS records showed that families referred to the CAP program had an average of 5.66 accepted reports of child maltreatment prior to the CAP assessment (Table 5.8). Additionally, 29.8% of the Target Children were placed out of the family home sometime prior to the CAP assessment. All but 5 of these children were in the family home at the time of the CAP assessment, and these 5 children were returned to the family home within 60 days of the CAP assessment (Table 5.8). When looking at the reason for the reports of child maltreatment closely preceding the CAP referral, about 60% of the reports involved child neglect, 17% physical abuse, 12% risk only, 11% a combination of concerns, and 1 report involved a family that requested services from CWS (Table 5.10).

Table 5.10
Type of Maltreatment Alleged on the Report Prior to CAP Assessment

Allegation type listed on report prior to the CAP assessment	Frequency	Percent
Risk Only	28	11.9%
Neglect	140	59.6%
Physical Abuse	41	17.4%
Combination	25	10.6%
FVS Intake (Request for Services)	1	.4%

Table 5.11 lists the number and percent of families who had a new report of child maltreatment after the CAP assessment as well as the number and percent of Target Children placed in out-of-home care after the CAP assessment. As indicated in the table, about 65% had a new report of child maltreatment during the study period, and about 20% of the Target Children were placed into out-of-home care after the CAP assessment. The last two rows of Table 5.11 combine new report or placement. The date of which of these two events occurred first was used to create the time variable for the Cox proportional hazard model. As indicated in Table 5.11, about 66% of families had either a re-report of child maltreatment or placement of the Target Child into out-of-home care after the CAP assessment. Through the use of the Life Tables function in SPSS, it is was determined that for those who had event indicating child maltreatment (either a report of child maltreatment or placement of the Target Child in out-of-home care) after the CAP assessment, the median time until the event occurred was about 283 days.

Table 5.11
Rate of New Incident: New Report or Placement of Target Child after CAP Assessment

	Frequency	Percent
New report on the family after CAP	152	64.7%
No new report on the family after CAP	83	35.3%
Placement of Target Child after CAP	46	19.6%
No placement of Target Child after CAP	189	80.4%
Either new report or placement after CAP	156	66.4%
No new report or placement after CAP	79	33.6%
For those that had a new report/Placement episode after CAP: Median days till event was 283 Std. Error 63.9		

5.4 Survival Analysis

A series of survival analysis was done using the Cox proportional hazard regression program in SPSS-19. Time to either first new report or placement of the Target Child was entered as the time variable and event either occurring (coded as 1) or not occurring (coded as 0) was entered as the outcome variable. The first covariate entered was the Non-Defensive/Defensive group. Table 5.12 shows the results of the model and indicates no significant differences in survival time (time to next event) between the two groups. In Table 5.12, the Exp(B) indicated a hazard ratio of 1.276. An Exp(B) of 1.0 would indicate that someone above the cut score on this factor would be just as likely as someone below the cut score on this factor to experience an event while an Exp(B) of 2.0 would indicate that someone 'indicated' on this factor would be two times as likely to experience the event over the course of the study. An Exp(B) of less than 1.0 would indicate a decreased likelihood of an event occurring.

The survival curve for the Non-Defensive compared to the Defensive group is also shown in Figure 5.1. This is a graphical representation of the results shown in Table 5.12, the Y axis represents the percentage of families for which there is no new incident (cumulative survival) at a given time, and the X axis represents the number of days elapsed in the study. The study date for each family ended when a new report of child maltreatment was accepted or the Target Child was placed in out-of-home care. If neither of these events occurred, the study date ended when the data for the family was obtained from CWS. Consequently, each family had a unique number of days in the study, with the time variable either ending with an event or no event. The families were included in the study for an average of 422 days, with a range of 0 to 1517 days. In Figure 5.1, the two survival curves are fairly close to each other, indicating no significant difference in survival time between the two groups. Next, those who were indicated on the SDM were compared to those not indicated on the SDM. Table 5.12 shows that this variable was a significant predictor in the model (significance level of .001). Further, Exp(B) in Table 5.12 indicates that the hazard ratio for those with indicated SDM scores was about 1.92 times greater than

for the group with low or moderate SDM scores. Figure 5.2 shows the comparison between the two survival curves.

Table 5.12
Cox Proportional Hazard Models

Variable Entered into Model	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Lower
Non-Defensive /Defensive Group	.244	.194	1.573	1	.210	1.276	.872	1.867
Indicated/Not Indicated on SDM	.653	.200	10.724	1	.001	1.922	1.300	2.842
Non-Defensive Group: Indicated/Not Indicated on SDM	.787	.242	10.587	1	.001	2.196	1.367	3.528
Defensive Group: Indicated/Not Indicated on SDM	.165	.371	.196	1	.658	1.179	.569	2.441

Figure 5.1

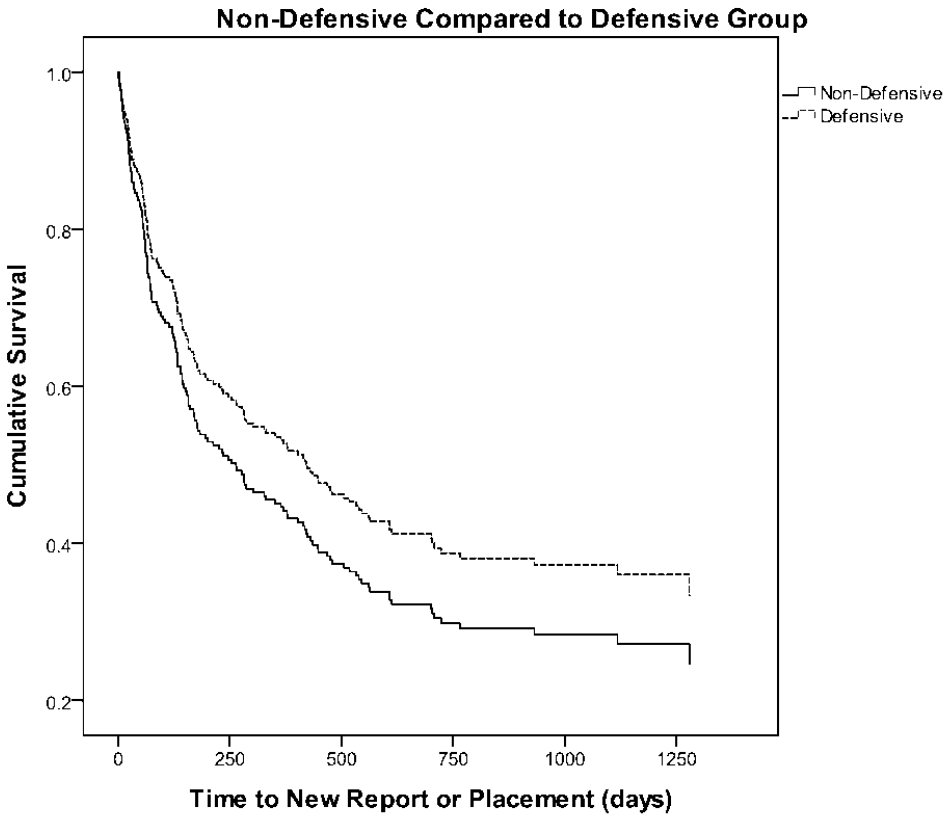
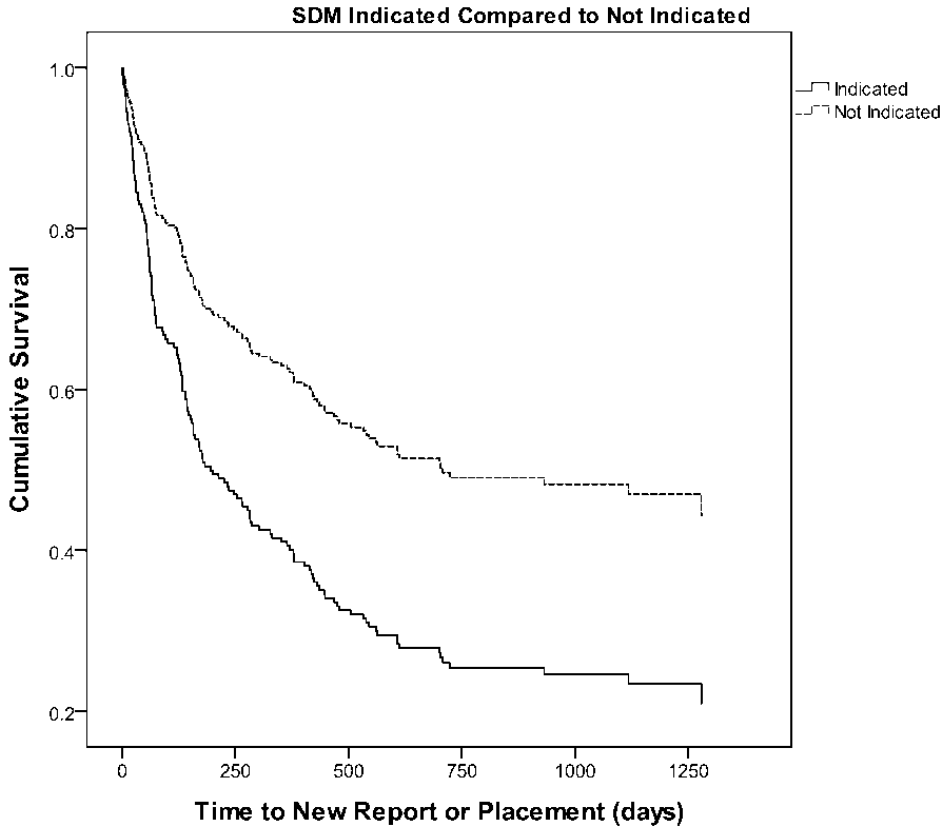


Figure 5.2



The subsequent analyses involved examining the Cox proportional hazard model for just the Non-Defensive group using the indicated versus not indicated category on the SDM as a covariate. The same analysis was then done for the Defensive group. Table 5.12 shows the results for these two analyses. The SDM was again a significant predictor for the Non-Defensive group (significant at .001) in which the Exp(B) increased to 2.2, indicating that for the Non-Defensive group, the hazard ratio for the indicated classification on the SDM was over double compared to the ratio for not indicated classification in the SDM (see Figure 5.3). This is contrary to the findings of the Defensive group. The caseworkers' classification on the SDM as either indicated or not indicated was not predictive of the time to next event in the Defensive group (see Table 5.12 and Figure 5.4).

Figure 5.3.

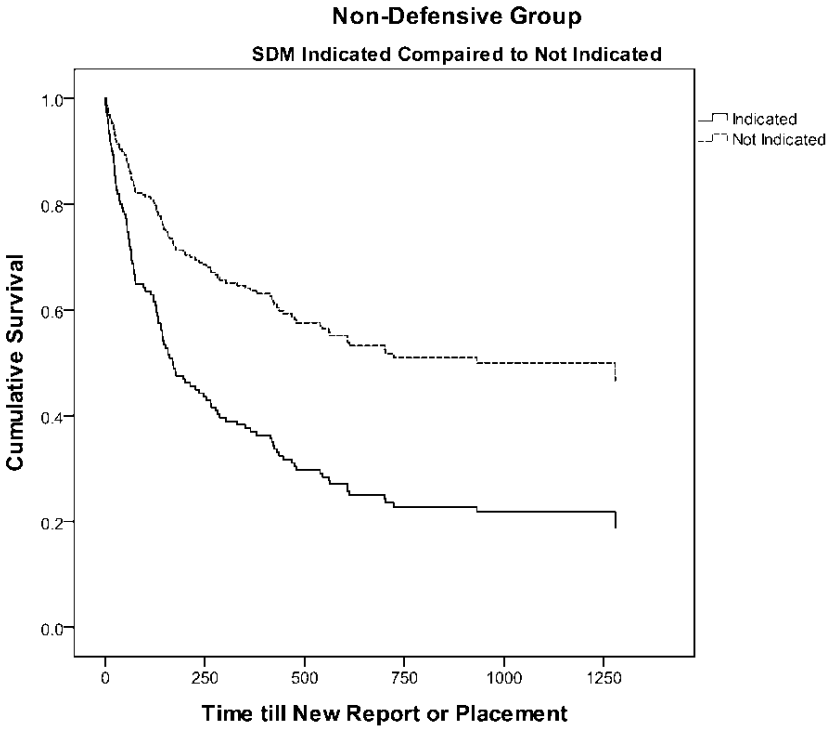
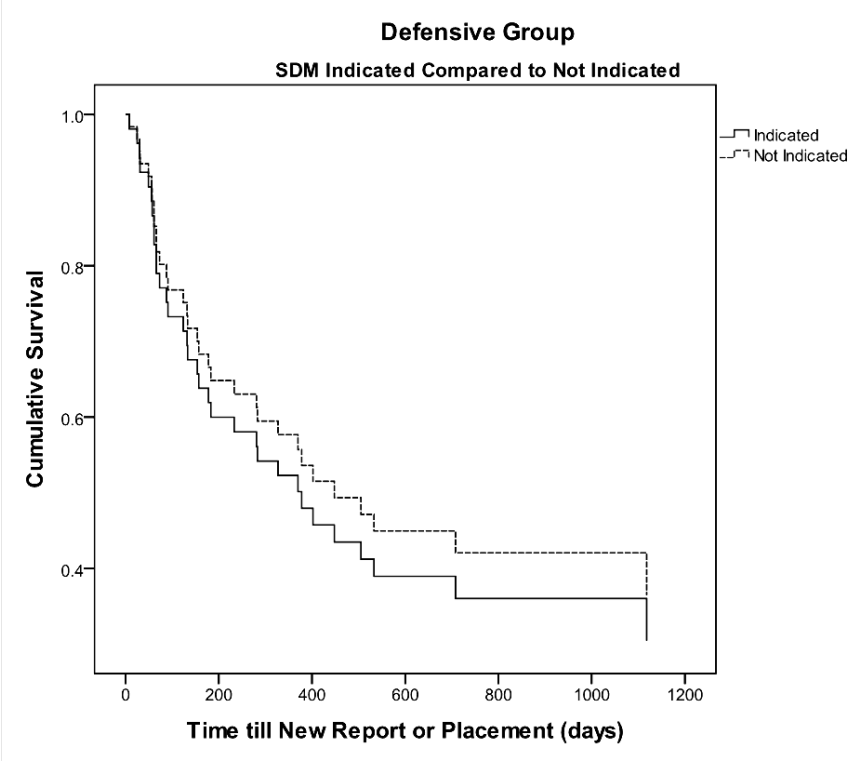


Figure 5.4



The next phase of the analysis involved incorporating the measures used during the CAP assessment to see whether any of the assessed theoretically modifiable variables would influence the model. The dichotomized (indicated or not indicated) results from the PHQ-9, PSI-SF, ASSIST 3.0, and the CTS-2 were then included in the Cox proportional hazard model for the Non-Defensive and Defensive groups separately while also keeping the SDM score as a covariate. This was done to see whether any of the data from the standardized instruments used in the CAP assessment could add to the predictive validity of the SDM. The Backward Likelihood Ratio (LR) function in SPSS-19 was used to determine, which of the covariates, if any, should be included in the model. Separate analysis was conducted with the PSC-17 scales by entering various subscales together as a covariate with the SDM. Similarly, the PEDS scale was entered separately as a covariate with the SDM using the procedure outlined above. This was done as the number of participants receiving the PSC-17 or the PEDS assessment was about half the total population. To maximize the use of the available sample size, it made sense to input these separately. Moreover, the PEDs and PSC-17 could not be entered together because families only received one of these assessments, depending on the age of their child. The Valid BCAP scores were not included, as the primary caregiver obtained valid scores on only 41 BCAP assessments, which resulted in very low cell sizes when split into the 4 categories (Defensive/Non-Defensive and SDM Indicated/SDM Not Indicated).

For the Defensive group, none of the additional covariates added significantly to the predictive validity of the model; consequently, no model was created for the Defensive group. However, for the Non-Defensive group, the resulting model included the SDM, PSI-Parent/Child Dysfunction subscale, parent report of Domestic Violence in the previous year, and parent report of Drug/Alcohol use (see Table 5.13 for the final model). None of the other measures, including the PSC-17 and the PEDS, significantly affected the model when entered individually with the SDM. The proportional hazard (PH) model assumptions were tested using a procedure outlined in Kleinbaum and Klein (2012). This

procedure uses the Schoenfeld residuals for each participant who had an event. If the PH assumption holds for each particular covariate, than the Shoenfeld residuals for that covariate would not be related to survival times. No covariate residuals were related to survival time in the final model, indicating that the model did not violate the PH model assumptions. Z-scores for each of the covariates were created using the obtained continuous measures. Interactions between the covariates were then entered into the Cox proportional hazard model. None of the interactions approached significance.

There was also a check to see if prior involvement with CWS may have impacted caregivers defensive responding. This was examined to determine if prior involvement was a potential confounding variable, as it may have been that the more defensive group had more history and consequently represented a higher risk group even though there scores on the standardized measures were lower (e.g., there was an unmeasured variable that increased risk in the defensive group). The results indicated that the Non-Defensive responders had a mean of 6.27 prior reports of child maltreatment (SD 5.36) while the Defensive responders had a mean of 3.82 prior reports (SD 2.86).

Table 5.13
Resulting Covariates in the Equation for the Non-Defensive Group

Covariates Entered into Model	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Lower
Indicated/Not Indicated on SDM	.914	.257	12.641	1	.000	2.494	1.507	4.128
Domestic violence/ Previous year	.405	.226	3.209	1	.073	1.500	.963	2.337
PSI-Parent/Child Dysfunction	.558	.191	8.533	1	.003	1.747	1.202	2.541
Drug/Alcohol	.501	.204	6.022	1	.014	1.650	1.106	2.462

Further analyses were conducted to assess how the non-modifiable variables influenced the model created thus far. As the SDM includes measures of the age of children in the home, number of children in the home, and number of previous reports, these variables were not included. This resulted in age of caregiver, child ethnicity, and prior placement of the Target Child being included in the

analysis. The procedure of using the Backward LR function in the Cox proportional hazard function in SPSS was repeated with these additional covariates. The resulting model continued to include the SDM, PSI Parent/Child Dysfunction subscale, parent report of domestic violence in the previous year, and parent report of drug/alcohol abuse while none of the additional non-modifiable variables added significantly to the model.

As one of the goals of this study was to explore how the data from the CAP might be used to create an assessment that not only better predicts future incidents of child maltreatment, but also provides guidance as to service provision, each tool's unique contribution was also examined. Using the cut scores for each instrument, each measure from the CAP assessment was entered into a Cox hazard model independently for the Non-Defensive responders. Table 5.14 shows the significance and hazard ratio for each of the CAP measures. As can be seen, a number of measures that did not remain in the final model when using the Backward LR function reached significance or approached significance when entered into the model independently. This is likely due to the high correlation between these measures, with stronger predictors suppressing other predictors. However, when entered independently, other areas in which significant levels of concern were associated with higher rates of future maltreatment were highlighted, providing insight into domains were intervention might focus.

Table 5.14

CAP Measures Entered Separately into the Cox Proportional Hazard Model for the Non-Defensive Group

Measure	Number	Significance	Hazard Ratio Exp(B)	95% Confidence Interval for Exp(B)	
				Lower	Upper
PSI-Parent Distress	177	.073	1.465	.965	2.224
PSI-Parent/Child Dysfunction	177	.001	1.849	1.287	2.656
PSI-Difficult Child	177	.016	1.565	1.087	2.252
PSI-Total Stress	177	.002	1.799	1.248	2.593
PHQ-9	177	.054	1.447	.994	2.107
CTS-2 Domestic Violence-Previous Year	176	.059	1.497	.984	2.276
CTS-2 Domestic Violence-Ever	176	.051	1.455	.998	2.120
ASSIST 3.0 (Drug/Alcohol)	172	.028	1.538	1.047	2.260
PSC17-Attention	113	.396	1.222	.769	1.941
PSC17-Internalizing	113	.451	1.195	.752	1.898
PSC17-Externalizing	113	.209	1.325	.854	2.054
PSC17-Total	113	.062	1.518	.980	2.352
PEDS	59	.452	1.284	.669	2.465
BCAP (Valid BCAP scores)	41	.007	2.883	1.341	6.199

All BCAP scores that were valid are included, whether or not the individual responded in a Defensive/Non-Defensive manner of the PSI.

5.5 Correlations between Caregiver and Caseworker Report

Subsequently, additional analysis investigated the modifiable data obtained from both the caseworkers' classification as alleged on the SDM and by the caregivers' self-classification as indicated by the corresponding CAP measure. This analysis included both current drug/alcohol use and domestic violence in the previous year. This was done to examine and illustrate how the discrepancies between caseworker reports and caregiver reports might account for caregivers' reports adding to the predictive validity of the model. Table 5.15 shows the cross tabulation table for Non-Defensive caregivers' reports and caseworkers' reports of the caregivers' use of drugs/alcohol. Forty-nine caregivers self-reported as being at risk for drug/alcohol abuse, but only about half of these were alleged to be substance abusing according to the caseworkers' assessment on the SDM. The Pearson Chi-Square test indicated an overall positive correlation between the caseworker assessment and the caregiver report (significant at less

than .001); however, the Phi Coefficient signified that the relationship is only modest (.279). Concerning domestic violence in the previous year, 35 of the Non-Defensive parents reported having experienced domestic violence in the previous year, while caseworkers alleged that only 11 of these 35 had encountered domestic violence in the previous year (see Table 5.16). Again, the Pearson Chi-Square test signified an overall positive correlation between the caseworker assessment and the caregiver report (significant at less than .001); however, the Phi Coefficient was again only modest (.325).

Table 5.15
Cross Tabulation for the Non-Defensive Group – Substance Abuse

		Current use of Drugs/Alcohol as Indicated by the Caseworker on the SDM		Total
		Indicated	Not Indicated	
Client Report of Substance Abuse Problem[1]	Indicated	25 51%	24 49%	49 100%
	Not Indicated	26 22.8%	88 77.2%	114 100%
Total	Count % of Total	51 31.3%	112 68.7%	163 100%
Significant at less than .000 on the Pearson Chi Square test				
Phi coefficient (indicates strength of association) = .279				

[1]As measured on the ASSIST 3.0 screener, score of 2 or 3 is indicated and a 0 or 1 not indicated

Table 5.16
Cross Tabulation for the Non-Defensive Group – Domestic Violence

		Domestic Violence in the Previous Year as Indicated by the Caseworker on the SDM		Total
		Indicated	Not Indicated	
Client Report of Domestic Violence During the Previous Year[1]	Indicated	11 31.4%	24 68.6%	35 100%
	Not Indicated	8 6.1%	124 93.9%	132 100%
Total	Count % of Total	19 11.4%	148 88.6%	167 100%
Significant at less than .000 on the Pearson Chi Square test				
Phi coefficient (indicates strength of association) = .325				

[1]As measured on the CTS-2, score considered indicated if any event was reported in the previous year

Chapter 6

Discussion of Study 1

Study 1 examined the results of the standardized assessments used by the CAP program as well as the data maintained by CWS. The obtained results provide a wealth of information that can be used to guide future efforts to increase the reliability and validity of the assessment of CWS involved families.

6.1 Discussion as Related to Research Questions

1) What are the characteristics of the family members according to the various CAP assessment tools?

The results of the CAP assessments indicated that the referred families struggled with multiple challenges, and they were at a high risk for future incidence of child maltreatment. This is particularly evident when looking at the scores on the various standardized assessment tools for the Non-Defensive responders who reported high rates of struggles in nearly all examined domains. However, even though the Non-Defensive responders scored significantly higher than would be expected in the general population, this study was designed to examine complex, high-risk, families involved with CWS. Based on previous research it might be expected that the percentage of family members classified as indicated on the various standardized assessments should have been even higher. An important consideration is that the results obtained in this study are based on the assessment of parents with open CWS cases, and the assessments were completed for the purposes of case planning. This is in contrast to data obtained for most other research studies of CWS involved caregivers, which provide assurances to participants of confidentiality and are often done after the CWS case is closed. One example of this difference might be seen on the PSI-SF. The average Total Stress score of 79.8 obtained in this study for the Non-Defensive group was less than the average of 89.2 found by Haskett et al. (2006) when administered to parents with a history of physically abusing their children. There is some indication that the effect of the

assessment format of the CAP also influenced the caregivers' reports in the domains of domestic violence and substance abuse.

It is possible, but seemingly unlikely, that the non-significant differences between the Defensive and Non-Defensive responders on the measures of domestic violence and substance abuse were due to the Defensive responders reporting in a less defensive manner in these two domains. Another, perhaps more likely alternative is that the Non-Defensive responders reported in a more defensive manner in these two areas, as it may be that caregivers involved with CWS attach greater weight to the potential consequences of fully disclosing current domestic violence and substance abuse. The lower percentage of both the Defensive group (13%) and Non-Defensive group (19%) in reporting domestic violence in the last year compared to previous research appears to support this claim (Kohl et al., 2005; Marcenko, Lyons, & Courtney, 2011). In contrast, particularly for the Non-Defensive group, the percentages reporting domestic violence occurring over a year ago were more consistent with what would be expected (55% for those in the Non-Defensive group and 40% for those in the Defensive group). It is a little harder to examine the possible extent of underreporting in the area of substance abuse due in part to the large discrepancy in previous research on the prevalence of substance abuse among clients involved with CWS (Young, Boles, & Otero, 2007). However, based on the high-risk level of those referred to the CAP program, the lack of significant difference between the Defensive and Non-Defensive responders, and the potential significant consequences of admitting substance abuse, it seems plausible that the study population underreported their substance abuse. As domestic violence and substance abuse are two areas of significant concern regarding caregivers involved with CWS, they are worthy of further investigation, as accurate assessment in these domains can influence both risk assessment and service planning.

In addition to the standardized assessment tools that indicate family struggles in a number of different domains, the CWS history of the families prior to the CAP assessment also indicated that the

referred families were at high risk of future child maltreatment. This was demonstrated by the number of previous reports, an average of over 5 previous reports per family, as well as by 30% of the Target Children having been placed in out-of-home care at some time prior to the CAP assessment. These results would indicate that CWS caseworkers were identifying families at high-risk of future child maltreatment to refer to the CAP.

2) What percentage of families receives new reports of child maltreatment or placement of the Target Child in out-of-home care and when do these incidences occur.

The re-report rate and placement rate of Target Children in out-of-home care after the CAP assessment would further support the notion that the CAP was assessing a group of caregivers at high risk of future child maltreatment. The re-report rate of 64% found in this study is higher than that found in previous studies of the general population involved with CWS as well as higher compared to other studies that have focused on the higher risk families involved with CWS (English & Marshall, 1999; Fluke, Shusterman, Hollinshead, & Yuan 2008). The placement rate of almost 19% of the Target Children after the CAP assessment and 30% at some time prior to the CAP are consistent with the placement rate of 37% that Loman (2006) found when examining families frequently encountered by CWS. Furthermore, consistent with previous studies (English & Marshall, 1999; Fluke, Shusterman, Hollinshead, & Yuan 2008), the highest risk time for a new report of child maltreatment was in the first year after the CAP assessment.

3) Does the SDM score predict the rate and timing of new reports of child maltreatment or placement of the Target Child after the CAP assessment?

Overall, the SDM seemed to distinguish between families at a higher and lower risk of receiving a new event, as indicated by the results of the survival analysis showing that those indicated on the SDM had an increased odds ratio of 1.92 of having a new event occur. However, it is also noteworthy that

about 50% of those with an SDM classified as not indicated experienced a new event over the course of the study. This seems to indicate that caseworker judgment in certain situation may be more valid compared to the SDM. This is in contrast to previous research that has indicated that when caseworkers override the SDM, keeping open cases that were lower risk according the SDM score, more often than not, their override is inaccurate in predicting future maltreatment (Johnson, 2004). However, this study looked at a particularly high-risk portion of the families involved with CWS, and it is possible that caseworker judgment performs better for this subpopulation than it does for the general population involved with CWS.

The three research questions just reviewed are similar to questions that have been examined in other studies of families involved with CWS. For the most part, the findings in this study are consistent with previous research, which has reported similar findings regarding the functioning of the SDM and regarding the risk factors associated with child maltreatment. The remaining five questions are more exploratory in nature.

4) Is there a significant difference between the Defensive and Non-Defensive groups regarding a new event occurring (either re-report or placement of the Target Child)?

Even though those classified as defensive according to the PSI had significantly lower scores on many of the standardized assessment tools, the results indicated that the risk of future child maltreatment was similar for groups that are both defensive and non-defensive on the PSI-SF. Furthermore, the results indicated that a significant portion of caregivers (24%) were categorized as Defensive. It is noteworthy that on the BCAP validity scale, 54% percent provided invalid responses. Due to the lower number of caregivers responding to the BCAP, it was not possible to fully evaluate the difference between those who reported in a valid versus invalid manner; however, it indicates that defensive reporting might be greater than 24%. Additionally, the rates of both defensive responding on the PSI-SF and invalid responding on the BCAP were much higher than previous findings (Haskett et al.,

2006; Ondersma et al., 2005; Walker & Davis 2012), indicating a difference between the assessment for treatment planning and assessment for research purposes.

5) Does the SDM predict a future maltreatment differently for the Defensive and Non-Defensive groups?

The SDM is largely based on historical or situational factors, with some measures relying on the caseworkers' assessment/judgment. As it was assumed that the risk associated with the historical information would be the same for the two groups, and since caseworkers rely on multiple sources to gather information on family functioning, it was predicted that the SDM would function equally well for the Defensive and Non-Defensive caregivers. It is not clear why it did not perform as well with the Defensive group as it did with the Non-Defensive group. This study was not set up as an item analysis of the SDM tool, and many of the items on the SDM were not collected. It is possible that on those items where the caseworker had to rely on the caregivers report, for example, recent use of drugs or alcohol, the Defensive group disclosed less compared to the Non-Defensive group, resulting in a lower overall score. However, without a closer examination of the SDM, it is not possible to make a more comprehensive determination of the functioning of the SDM with defensive responders. This would appear to be an important issue to examine in future research, as the results of this study indicated that for a significant portion of clients involved with the CWS, the caseworker assessment of risk of future maltreatment, as indicated on the SDM, is compromised.

6) Does the predictive validity of the modifiable variables (e.g., depression, child behavior problems…) measured during the CAP assessment add to the predictive validity of the SDM?

As with other studies, this study indicate that various family risk factors, such as substance abuse, domestic violence, and child behavior problems, are related to future reports of child maltreatment (Dakil, Cox, Lin, & Flores, 2012; Fluke, Shusterman, Hollinshed, & Yuan, 2008). However, this is the first study that this researcher is aware of that has examined whether these various risk

factors, as measured by caregiver report on standardized assessments, strengthen an actuarial tool, such as the SDM, in predicting future child maltreatment. Furthermore, this is the first study to this researcher's knowledge that has differentiated between those clients involved with CWS who report defensively compared to those who report in a non-defensive manner. The first notable finding is that the problems, as self-reported by Non-Defensive caregivers on the various standardized assessment tools, is related to future risk of child maltreatment. Again, it is important to keep in mind that the clients were aware that the assessment tools were going to be used as part of case planning, and therefore, even with this knowledge, the clients were seemingly willing to complete the assessments in an honest enough manner to provide insights into their struggles and treatment needs. Furthermore, the results indicated that for the Non-Defensive caregivers, the addition of these self-report measures added significantly to the predictive validity of the SDM. This finding has important implications and can help guide the evolution of the assessment of clients involved with the CWS.

7) Do any of the non-modifiable variables (e.g., caregiver age, child gender…) improve the predictive validity of the SDM?

It has previously been found that non-modifiable variables, such as history with CWS and age of children in the family home, are the best predictors of future child maltreatment, and some have demonstrated that when non-modifiable variable are accounted for, the modifiable variables add little predictive validity (Thompson & Wiley, 2009). In this study, the addition of non-modifiable variables not already included in the SDM did not add significantly to the predictive validity of the SDM. This included an examination of prior placement of the Target Child, which was predicted to correlate with increased risk. Although not directly examined, it is likely that the non-modifiable variables included in the SDM significantly added to the predictive validity of the SDM.

8) What is the correlation between the caseworkers' assessment and the clients' report of the presence of current domestic violence and substance abuse in the clients' life?

Lastly, the study examined the relationship between the concern regarding recent domestic violence and substance abuse as reported by caseworker and caregivers. This was done to explore whether the client reporting on the standardized assessments was providing information that the caseworker may have been missing. Although the results indicated a significant correlation between caseworker and caregiver reports of concern, the strength of the association was not very strong. For instance, the caseworkers identified just 50% of the Non-Defensive caregivers as experiencing substance abuse who were classified as substance abusing on the ASSIST 3.0. These results are similar to the results found in other studies in which CWS caseworkers failed to accurately assess clients' needs (Burns et al., 2004; McCrae & Barth, 2008; Young, Boles, & Otero, 2007) Based on these results, the inclusion of caregiver reports on standardized assessments seems vital to increase assessment accuracy.

6.2 Study Limitations

This study has a number of limitations that should be acknowledged. The use of CWS workers to pull data directly from the computerized state files rather than having the data automatically pulled through the use of a computer program has the benefit of ensuring a match among the case, the caregiver, and the child; however, it may have resulted in some inaccuracies due to human error. Moreover, some reports of maltreatment may have been missed by focusing only on primary caregivers. It is possible that other cases related to the Target Child were not included, for example, there may have been a separate case being maintained on the father of the Target Child that was not connected to the mothers' case. However, it is also important to note that similar reliability issues can be observed with automated data pulls. Furthermore, it is likely that any errors that may have occurred were randomly distributed. Another limitation was related to the proximity of the SDM in relation to the CAP assessment, which varied significantly. This limits the strength of the findings concerning the SDM; however, of primary concern in this study was the functioning of the standardized assessment tools

used by the CAP. Consequently, a decision was made to use the date of the CAP assessment as opposed to the date of the SDM assessment as the start date for the survival analysis.

Other limitations relate to the generalizability of the findings. The data was collected from CWS involved clients from just one state in the Pacific Northwest. Furthermore, families were not referred to the CAP program randomly; instead, caseworkers selected the cases they felt were most appropriate for a CAP assessment, leading to a number of issues that limit the generalizability of the findings. For example, it is possible that caseworkers focused on certain types of high-risk families, perhaps those more likely to be actively engaged with CWS, which might have been confounded by the fact that only those who were referred and completed the CAP assessment were examined in this study. It is also important to highlight that this study was not an evaluation of the CAP program. The resulting CAP report was not examined, and no effort was made to see whether caseworkers referred families to the recommended services or whether families consequently engaged in these services. Therefore, the study findings provide evidence in support of the concurrent and predictive validity of the assessment tools and not the utility and consequences of resources such as the CAP. Lastly, it is important to note that the study sample size, and in many analysis the unequal sample sizes, limits the strength of the findings. This is particularly important, as many of the questions explored are more exploratory, which increases the possibility that the findings were due to chance.

6.3 Conclusion

The results of Study 1 indicated that in most cases, caregivers' reports on standardized assessment add to the predictive validity of assessment in CWS. Furthermore, the results of the assessment may add to the concurrent validity and utility of the assessment process by providing information about the areas with which the family is struggling. Consequently, the information obtained on the standardized assessments can be used in both service planning and risk assessment of children in the homes that are at the highest risk for child maltreatment. Although this study would indicate value

of incorporating self-report assessments in the CWS assessment process, the use of standardized self-report assessments is challenging because these tools add to the burden and complexity of the assessment process. Although many of the assessment tools examined in this study are brief, the sheer number of domains that need to be assessed adds to the burden placed on caregivers, and the processing and interpretation of the results would add considerably to the burden placed on caseworkers. An additional challenge concerns the finding that a portion of the families reported in a defensive manner and the standardized assessment tools are not helpful in either risk assessment or case planning for this group. Additionally, since caseworkers face apparent challenges when assessing these families, as demonstrated by the performance of the SDM for the Defensive responders, caseworkers need particular support when assessing these families. This would indicate the need to create assessment tools that can be efficiently administered, easily processed and interpreted, as well as provide reliable and valid information for caregivers who respond in either a non-defensive or a defensive manner.

Study 2 built on the findings of Study 1 and attempted to address some of the assessment needs highlighted in Study 1. The same data that was collected for Study 1 was analyzed in Study 2 but in a different manner. Study 2 explored how Item Response Theory (IRT) may be used to assist in the development of a self-report assessment for parents involved with the CWS, which in addition to showing high reliability and validity is also resistant to defensive responding. Furthermore, the subsequent study examined how IRT can assist in creating a multidimensional assessment tool that is shorter and easier to administer compared to a series of standardized assessments as well as providing the assessment results in a format that caseworkers and caregivers could understand and use more easily.

Chapter 7

Study 2

7.1 Background and Literature Review

Caseworkers in Child Welfare Services (CWS) provide protection and support for the most vulnerable children in society. To fulfill their responsibilities, caseworkers must assess family needs accurately. In child welfare, assessment needs include (a) assessing the immediate level of risk to the child(ren), (b) determining the likelihood of future maltreatment, (c) determining family service needs to promote the safety and well-being of the children in the home, (d) matching family members to appropriate services, (e) and monitoring their progress in these services. To assist caseworkers with these duties, CWS have implemented a number of assessment tools to guide caseworkers in gathering information and assist them in determining the level of risk to the children in the family home.

The first concern that the assigned caseworker must address is to determine whether the children are currently safe in the family home. If a determination is made that the children are not safe in the family home, then a decision must be made as to whether a plan can be put in place to ensure their safety at home and if this is not possible, the children are then placed in an out-of-home care. CWS in the state of Washington, as well as in many other states, have implemented a safety assessment tool to assist caseworkers in making these determinations. To complete the safety assessment, the caseworker assigned to the family gathers information from family members and from those who are aware of the families current functioning and then uses this information to answer a series of questions, which in turn helps guide decision-making. However, as outlined in Study 1 (section 3.1.1), research has not been done to determine if caseworkers fill out the safety assessment in a consistent manner or to establish the concurrent or predictive validity of the safety assessment. Consequently, the safety

assessment can best be viewed as a guide to assist caseworkers in decision-making, as opposed to being considered a reliable and valid assessment tool.

Following the safety assessment, CWS caseworkers complete the structured decision making tool (SDM) (section 3.1.2 and Appendix B). The SDM is used to determine, which families are at a heightened risk for abusing or neglecting their child(ren). To complete the SDM, the caseworkers again rely on information from various family members and others who are familiar with the families' functioning. Additionally, the SDM scoring matrix considers the caregiver's history of child maltreatment as well as the prevalence of various risk factors both past and present. For example, the SDM requires caseworkers to indicate the presence/absence of risks, such as caregivers' substance abuse, domestic violence, or child vulnerability, among others. The items included in the SDM, and the weight given to these various items, have been determined using standard statistical procedures to maximize the tool's ability to predict both the likelihood and the severity of future child maltreatment (Rycus & Hughes, 2003).

Research has indicated that the SDM predicts future child maltreatment more accurately compared to caseworker judgment without the assistance of valid and reliable tools (Baird & Wegner, 2000). However, studies have also indicated that the SDM is far from perfect in predicting future maltreatment; in particular, the SDM has been shown to be overly sensitive, resulting in a high level of false positives (Gambrill & Shlonsky, 2000). Additional concerns regarding the accuracy of the SDM involve caseworkers' assessment of the various risk factors. Previous studies have indicated that caseworkers are poor at identifying risk factors present in family homes (Burns et al., 2004; McCrae & Barth, 2008; Young, Boles, & Otero, 2007). Concerns about the accuracy of caseworkers' assessment of clients' needs were also observed in Study 1. For example, in Study 1, the caseworkers identified only 51% of the caregivers who were classified as substance abusers by the Alcohol, Smoking and Substance

Involvement Screening Test (ASSIST 3.0), as experiencing substance abuse. Furthermore, the SDM does not provide insight into the significance of the families' struggles; consequently, the SDM is not useful for providing guidance as to what services might be helpful for the family or for monitoring a family's progress (Rycus & Hughes, 2003; Shlonsky & Wagner, 2005). These concerns indicate that additional assessment tools are needed to assist in the determination of family functioning.

Self-report assessments are traditional methods used to assess individual and family needs. Self-report assessments with support for their reliability and validity are available for measuring various issues that are of concern to CWS. In the state of Washington a program called the Comprehensive Assessment Program (CAP) was implemented to administer standardized self-report assessments to caregivers with open CWS cases who had their child(ren) in their care and were believed to be at high-risk of future child maltreatment. However, a potential challenge in administering self-report assessments when assessing families involved with CWS is that the clients are often not engaging in the assessment process voluntarily. Caregivers are reported to CWS due to concern regarding their ability to appropriately care for their child(ren). Based on the caseworkers' assessment, caregivers may be referred to services, which they may not want to attend, and in some instances, the caseworkers' assessment results in the child(ren) being removed from the family home. Consequently, assessments administered to CWS involved caregivers are susceptible to caregivers underreporting their concerns in order to limit potential negative consequences. The results of Study 1 indicated that most caregivers who completed the CAP assessment appeared to disclose information in a non-defensive manner; however, a significant portion appeared to respond in a defensive manner to at least some of the assessment items. For those caregivers who reported in a defensive manner the assessment tools did not appear to provide insight into service need or predict future child maltreatment; however, for those who reported less defensively, the assessment tools provided information that was useful for both service planning and for the prediction of future child maltreatment. In particular, less defensive

caregiver disclosure on the assessments of domestic violence, substance abuse, and parent-child dysfunctional interaction were not only independently predictive of future incidence of child maltreatment, but also added significant predictive value when included as a covariate in survival analysis with the SDM.

Domestic Violence

Many studies have shown a link between domestic violence (DV) and Child Welfare involvement (Hamby, Finkelhor, Turner, & Ormrod, 2010; Kohl, Barth, Hazen, & Landverk, 2005; Marcenko, Lyons, & Courtney 2011). In one study, which examined over 3000 female caregivers who were a part of the National Assessment of Child and Adolescent Well-Being, CWS caseworkers indicated that DV was present in 12% of the families investigated for maltreatment. In contrast, when items from the violence portion of the Conflict Tactics Scale (CTS) were administered to the caregivers who had been recently involved with CWS, 31% of the female caregivers reported domestic violence in the past year while 45% reported lifetime DV on this measure (Kohl et al., 2005). Furthermore, 19% of female caregivers in this study reported severe physical DV in the past year while 33% reported a lifetime prevalence of severe physical DV. Similar numbers were found in a study done by Marcenko, Lyons, and Courtney (2011). Based on the interviews with 747 parents with open CWS cases, the researchers found that 35.8% of the female caregivers reported verbal threats, physical aggression, or physical injury in their most recent relationship. The rate was 31.8% for mothers whose children were in their care while the rate for mothers whose children had been placed in out-of-home care was 38.8%. Furthermore, it has been found that high scores on the physical assault subscale of the Conflict Tactics Scale are associated with a significantly higher odds ratio of children being placed in out-of-home care (Horwitzet al., 2011).

In Study 1, the physical abuse portion of the CTS-2 was administered by the CAP to caregivers involved with CWS who had their children in their care and were believed to be at high risk of future

child maltreatment. Overall, 17.3% of the caregivers reported domestic violence in the previous year on the CTS-2. When examining just those who scored non-defensively on the Parent Stress Inventory-Short Form (PSI-SF), the percentage went up to 19% (Table 5.7). For the non-defensive caregivers, those who indicated domestic violence in the previous year also had an increased odds ratio of 1.497 [.984, 2.276] (95% confidence intervals are reported in brackets) for future child maltreatment (Table 5.14). When entered into the Cox proportional hazard model with the other covariates, the reports of domestic violence in the previous year remained a significant predictor of future child maltreatment, with a hazard ratio of 1.500 [.936, 2.337] (Table 5.13).

Substance Abuse

Studies examining the rates of substance abuse among CWS involved parents have reported various findings. Young, Boles, and Otero (2007) conducted a literature review and found that reported rates of substance abuse ranged from 11% to 80% across studies. Based on their analysis of the various studies, they estimated that 11% of children who were victims of child maltreatment and received in-home services had parents who met criteria for substance use disorder. Further, they concluded that this rate was between 43% and 70% for the parents of children who were placed in out-of-home care. A recent study examining data from the National Assessment of Child and Adolescent Well-Being found that 12.5 % of parents indicated harmful use or dependence on alcohol and/or other drugs (Chuang, Wells, Belletiere, & Cross, 2013). Another study based on confidential interviews with caregivers who had been recently involved with CWS and assessed as high-risk for future child maltreatment by the CWS caseworker found that 33.3% of the parents reported alcohol abuse (Proctor et al., 2012). This is consistent with findings from Marcenko et al. (2011) who found that about 30% of caregivers involved with CWS met the criteria for alcohol or drug abuse/dependency in the past 12 months. Furthermore, when comparing the re-report rates for families with no new reports and families categorized as

continuously re-reported for child maltreatment, alcohol abuse was associated with an increased odds ratio of 4.86 for belonging to the continuous re-reporting group (Proctor et al., 2012).

In Study 1, the ASSIST 3.0 was administered by the CAP to caregivers involved with CWS who had their children in their care and were believed to be at high risk of future child maltreatment. On the ASSIST 3.0, 27.9% of the assessed caregivers were at a risk of substance abuse. When examining just those who scored non-defensively on the defensive sub-score of the PSI-SF, the percentage went up to 29.5% (Table 5.7). Among the non-defensive caregivers, those found to be at risk of substance abuse also had an increased odds ratio of 1.538 [1.047, 2.260] for future child maltreatment (Table 5.14). When entered into the Cox proportional hazard model with the other covariates, the score on the ASSIST 3.0 remained a significant predictor of future child maltreatment, with a hazard ratio of 1.650 [1.106, 2.462] (Table 5.13).

Parent/Child Dysfunctional Interaction

The Parenting Stress Index-Short Form (PSI-SF) comprises three subscales, each measuring an aspect of a parenting stress (parental distress, parent-child dysfunctional interaction, and difficult child), as well as a defensive responding subscale. Research on the PSI-SF has indicated that the Parent-Child Dysfunctional Interaction scale in combination with Difficult Child scale was a unique predictor of child abuse, whereas the Parent Distress scale was not found to be a unique predictor (Haskett, Ahern, Ward, & Allaire, 2006). The Haskett et al.'s (2006) study, which examined both parents with a documented history of physically abusing their child(ren) and parents with no known history of abuse, found that parents with histories of abusing their child(ren) had a mean total score on the PSI-SF of 89.2 while the comparison group had a mean total score of 79.0. The researchers also found that 7.5% of all parents responded defensively, and these subjects were removed from further analysis in their study. There is also indication that Parent-Child Dysfunctional Interaction and Difficult Child scales correlate strongly

with measures of child externalizing and internalizing behavior problems while Parent Distress correlates more strongly with measures of depression (Costa, Weems, Pellerin, & Dalton 2006; Huth-Bocks & Hughes 2008).

In Study 1, the Parent-Child Dysfunctional Interaction scale (PSI-PCDI) was a significant predictor of future child maltreatment. Overall, 27% of the caregivers scored at or above the 85th percentile on the PSI-PCDI scale. When examining just those who scored non-defensively on the defensive scale of the PSI-SF, the percentage went up to 33.9% (Table 5.7). For the non-defensive caregivers, those scoring at or above the established cut score for the PSI-PCDI had an increased odds ratio of 1.849 [1.287, 2.656] for future child maltreatment (Table 5.14). When entered into the Cox proportional hazard model with the other covariates, the score on the PSI-PCDI remained a significant predictor of future child maltreatment, with a hazard ratio of 1.747 [1.202, 2.541] (Table 5.13).

Although previous research, as well as the results of Study 1, emphasized the value of incorporating self-report assessment tools into the CWS assessment process, a number of concerns remain. There are many different factors associated with child maltreatment, creating a burden for caregivers to complete the various assessment tools and for caseworkers to score and interpret the results of the various tools. Additionally, the apparent defensive responding by a significant number of caregivers presents problems that are difficult to address when assessment tools are scored using classical test theory. However, the use of Item Response Theory (IRT) may facilitate both the development and the scoring of assessment tools that address these concerns.

7.2 Item Response Theory

Assumptions of Classical Test Theory

Before examining IRT, it is helpful to review a few assumptions of classical test theory (CTT) that are particularly relevant to Study 2 and outline the ways in which the IRT procedures might address

challenges to these assumptions. The first assumption concerns the error surrounding the assessment items. In CTT, the assumption is that the error surrounding each item is random and unrelated to the true score and that in the long run, the sum of the error terms for all the items will equal zero (Streiner, 2010) (also see Chapter 2 for a review). A concern regarding the assumption that the error surrounding the items is random is that certain populations in certain circumstances may answer some items with greater error, which may not be random. An example of this may be evident in Study 1, were more error may have been seen in the reporting of domestic violence in the previous year than in the reporting of domestic violence ever due to possible caregiver underreporting. Another assumption regarding the error surrounding the items in CTT is that error is distributed equally across all trait levels (i.e., the confidence intervals around the scores are the same across the trait continuum), which often is not the case. An additional challenge is that according to CTT, longer tests are more reliable (Embertson & Reise, 2000). This may not be problematic if only one trait is being measured; however, when measuring multiple traits, this can create a significant burden and consequently reduce the utility of the assessment.

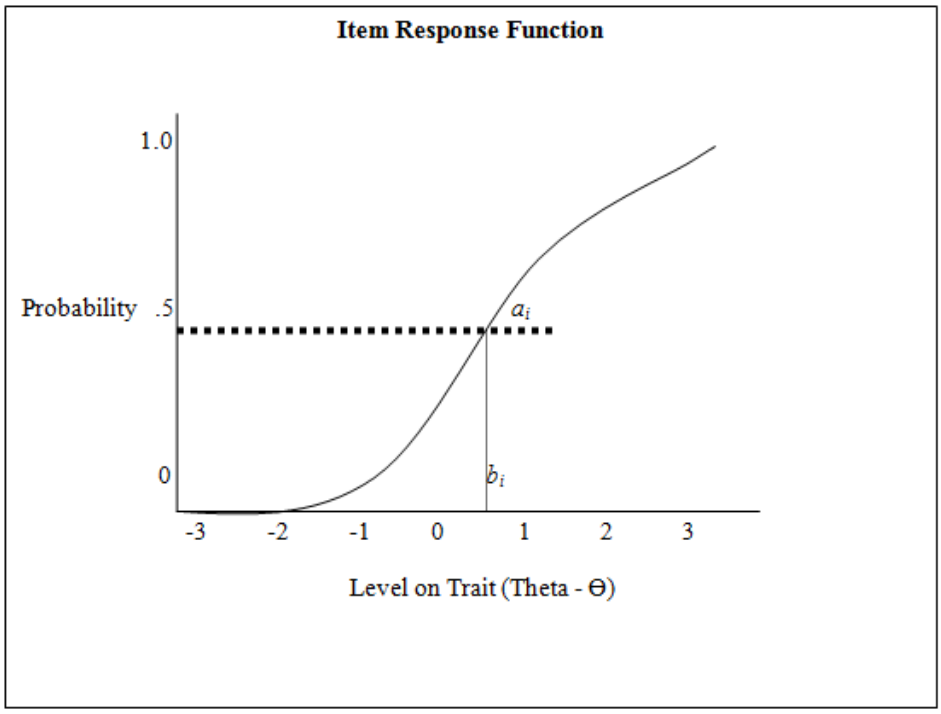
Another potential limitation is that items on the assessments developed with CTT are typically scored as if each of the items contributed equally to the total score (Streiner, 2010). Therefore, individual items are summed to create scale level scores based on the total number of items endorsed. The assumption of item equivalence can be seen in all standardized assessment tools administered during the CAP. On each of the administered assessments, the cut scores were established based on the count of the total number of items endorsed. For example, CTS-2 was considered indicated for domestic violence being present if any item on the CTS-2 was scored as having occurred in the previous year. Consequently, someone who reported that his or her "partner had pushed or shoved me" was scored the same as someone who reported that she or he had "A broken bone because of a fight with my partner". The concern related to item equivalence becomes confounded when an assessment contains

polytomous items (items with more than two alternatives). An example of this can be seen on the Parent Stress Index-SF, where a caregiver's response of *Not Sure* to the question, "My child turned out to be more of a problem than I had expected" is scored as 3 while the response *Agree* receives a 4 and response *Disagree* receives a 2. However, it is possible that the response distance (and potential risk to the child, or need for services) between *Not Sure* and *Agree* is different from that between *Not Sure* and *Disagree*.

Item Response Theory

IRT models rest on two basic postulates, "(a) The performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and (b) the relationship between examinees' item performance and the set of traits underlying items can be described by a monotonically increasing function called an *item response function* or *item characteristic curve"* (Hembleton, Swaminathan, & Rogers, 1991, p. 7). These two postulates can most easily be grasped when looking at a diagram of an item response function for a particular item (see Figure 7.1).

Figure 7.1
Sample of an Item Response Function for One Item



Where:

Probability: equals the likelihood that an individual with a specified level on the trait (Θ) is going to answer the item correctly or agree with the statement.

b_i: Indicates the location of the item on the underlying trait where there is a 50% chance that an individual will agree with the statement. In Figure 7.1, this particular item would be selected 50% of the time by those at about .50 above the mean trait level of the group. b_i typically reflects the item difficulty or location parameter. In cognitive assessments, b_i indicates how difficult the item is. However, in personality assessments, b_i indicates how much of the personality trait the individual possess for individuals who respond favorably to the item.

Theta (Θ): Represents the individual's level or ability on a standardized scale of the trait.

a_i: Is an indicator of the item discrimination. The steeper the slope of a_i at b_i, the greater the item discriminates between those below and above a trait level of b_i. In essence, the a_i parameter defines the information function for each item. The higher the discrimination, the more information the item gives at its location on the underlying scale and the lower the error associated with each item at its maximum information point.

Each item on an assessment has its own unique item response function. Consequently, the test-takers' level of the trait (Θ) can be estimated based on their responses to multiple items. The monotonic curve (increases occur simultaneously on the *x* and *y* axis) representing this type of item response function is based on the assumption of a dominance process, suggesting that a person would tend to endorse an item when his or her position on the underlying latent trait is greater than that of the item (Stark, Chernyshenko, Drasgow, & Williams, 2006). The item response function presented in Figure 7.1 would indicate that individuals with Θ levels at .50 or higher would be expected to endorse this item at least 50% of the time, with higher levels of the test-takers Θ indicating increasing probability of endorsement. If items representing different levels of the trait are available, maximum information

and efficiency (e.g., shorter assessments) can be obtained from assessments by selecting items representing the difficulty parameters of items covering the area of the trait that one is interested in assessing. If the goal were to distinguish those who have a certain amount of the trait from those who do not poses this amount, the test items would be centered at the predetermined cut score. If the goal is to get a good understanding of the Θ levels for a diverse range of test-takers, then the item locations should represent the range of possible Θ.

Recall that in CTT, the error surrounding each item is seen as random - unrelated to the true score of the individual or population being tested. In contrast, "in IRT models, trait scores are estimated separately for each score or response pattern, controlling for the characteristics (e.g., difficulty) of the items that were administered" (Embretson & Reise, 2000, p. 18). Many researchers advise to select items with higher *a*-parameters to improve item discrimination and consequently decrease measurement error (Hembleton, Swaminathan, & Rogers, 1991). This allows for a greater precision in interpreting a score's meaning, as the error around the scores is uniquely dependent on the test-takers response pattern and the difficulty and discrimination of items measuring that particular range of trait.

Another CTT assumption outlined earlier is that items are scored as if they are all equivalent. In contrast, the response to each IRT item can add a unique value to the assessment based on the items location (*b*-parameter) and the items discrimination (*a*-parameter). Consequently, an individual who answers *Agree* to three items that represent lower difficulty and *Disagree* to three items representing higher difficulty on the trait will receive a lower score compared to an individual who answers *Agree* to one of the lower and two of the higher difficulty items and *Disagree* to two of the lower and one of the higher difficulty level items. Furthermore, the weighting of the items is also influenced by the items' *a*-parameters, as items with greater discrimination estimate trait level with greater precision (lower standard errors) (Embretson & Reise, 2000).

These properties of IRT provide a number of advantages when assessing the functioning of an assessment and provide item level information that can be useful in the creation of new assessments. Regarding the assessment of existing tools, the use of IRT can provide insights into the functioning of individual items as well as the assessment as a whole. By examining the *a*- and *b*-parameters of each item, one can determine whether the items on the assessment are adequately assessing the trait levels that are of the greatest interest. This judgment would be based both on the assessment of whether an adequate number of items covers the trait level of interest and on the discrimination parameters of the items. Additionally, the overall test information function can show whether, and at what trait levels, the test as a whole provides accurate information. Furthermore, for polytomous items, an examination of each of the option's unique characteristics can assist in determining whether the options are spaced equally and whether all the options are useful (Streiner, 2010).

In addition to the benefits of IRT outlined thus far, IRT also provides a number of other advantages for test creation and use. First, IRT makes comparison of change scores possible even when initial score levels differ (Embreston & Reise, 2000). It has been shown that unless certain requirements are met, such as multiple measurement points, the use of CTT can be problematic when measuring change, particularly at the extremes, where an increase in one or two raw score points may mean something very different than a similar change in the middle of the scale (Embreston & Reise, 2000; McCoach, Rambo, & Welsh 2013). This is because instruments that are developed using CTT are often not true interval scales. In contrast, instruments developed using IRT, in which the measures of the items and the individual's trait level are standardized, provides a closer representation of interval scales and consequently facilitate measurement of change. A further benefit of using IRT is that it allows for relatively easy examination of differential responses to items by different groups (e.g., men, women, different ethnic groups, and others). This is very useful when the goal is to select items or create assessments that function similarly across different groups.

Additionally, a strength of IRT is that it provides information that is helpful for computer adaptive testing (CAT). In CAT, responses to earlier items guide the selection of future items as well as determine when a reasonable approximation of Θ has been found for the test-taker and the assessment should end. IRT allows for the ranking of items along the trait of interest, creating the opportunity to pull out subsets of items that cluster around the individual's level on the trait. If, for instance, an individual is responding in a manner that indicates she or he is at least average on the trait, then the remaining items administered can focus on determining how far above average the individual may be. It has been found that, when items are tailored to the individuals' trait levels with CAT, the scale can be reduced by as much as 50% without much loss of information (Embretson & Reise, 2000). Furthermore, if the *b*-parameter of the items is known, it may be possible to determine whether a person is above or below a cut score with just a few questions (Streiner, 2010). Consequently, the efficiency of assessing multiple domains simultaneously increases with CAT. Additionally, the utility of the assessment process can be further increased if the computer program assists with the scoring and interpretation of the results.

The advantages of IRT discussed thus far are significant and address a number of concerns related to the assessment of caregivers involved with CWS; however, how IRT might be able to help with the defensive responding (or faking) behavior found in Study 1 has not yet been addressed. As indicated in Study 1, a significant number of respondents appeared to be responding in a defensive manner. Further, the results of Study 1 suggested that those who appear to be responding in a defensive manner are not only harder to assess with self-report assessment tools, but also present assessment challenges for the caseworkers. The next section presents a measurement theory and assessment model that requires IRT analysis and may be of assistance in creating an assessment model that is more resistant to defensive responding. The outlined model does not yet have a significant amount of evidence supporting its validity, nor is an attempt made at this time to create an assessment tool based on this

model. However, the possibility of incorporating some of the ideas presented in the next section to assist in the development of an assessment tool that is resistant to defensive responding in a field like child welfare, where the potential consequences of inaccurate assessment are so severe, justifies a close examination.

7.3 IRT Consideration in a Coercive Assessment Environment

The assessment of families involved with CWS should be multi-source and multi-method. Caseworkers should acquire information from various people knowledgeable about the family's functioning as well as interview the family members and observe their interaction. Additionally, as demonstrated in Study 1, an important part of the assessment process should also be the inclusion of standardized questionnaires for caregivers. As noted earlier, when developing an assessment tool that would be completed by caregivers involved with CWS, one of the primary concerns is related to the reliability of the responses to the assessment items. In particular, caregivers might respond to assessment items in a manner that presents an overly positive view of either their overall functioning or their functioning in a particular domain (i.e., defensive reporting or faking their responses). When test-takers are motivated to fake their responses, it is thought that they respond in a way that is consistent with an "adopted schema" rather than responding according to a "self-schema". Test-takers who are responding to items using an adopted schema portray themselves in a way they would like to be seen rather than answering in a way they actually see themselves (Converse et al., 2010). This presents an element of systematic bias that cannot be controlled with statistical methods based on assumptions of random bias (Stark, Chernyshenko, Drasgrow, & White, 2012). In recent years, increased efforts have been made to address concerns related to faking by combining IRT with forced-choice (FC) questionnaires (Brown & Maydeu-Olivares, 2012; Chernyshenko et al., 2009; Christiansen, Burns, & Montgomery, 2005; Converse et al. 2010; Drasgow, Chernyshenko, and Stark, 2010; Jackson, Wroblewski, & Ashton, 2000; McCloy, Heggestad, & Reeve, 2005). The FC format is not new; however,

recent statistical models and the increasing power of computer programs have rekindled interest in this testing format.

Forced-choice questions can take on a number of formats. Questions can be presented as triplets (three items) or tetrads (four items), with individuals being asked to select the statements that are most and the least like themselves. A similar method used with tetrads involves presenting two positive and two negative statements for each question. The test-takers are instructed to identify the statement most like themselves and least like themselves. This format is assumed to reduce potential negative reaction from the test takers towards negatively worded statements (Dunnette, McCartney, Carlson, & Kirchner, 1962). Another commonly used method, and the format of focus in this dissertation, is to present statements as sets of dyads and the test-takers are instructed to select the statement in each dyad that best represents them. This format is also referred to as a pairwise preference questionnaire. Chernyshenko et al. (2009) indicated that this format provides a more straightforward way to construct and evaluate item quality and allows for the quality of different pairs of items to be more easily evaluated. The relevance of these advantages will become clearer as the theoretical background for the development of the assessment is further elaborated. Recent studies have shown that using various FC methods reduces faking (Christianson et al., 2005; Converse et al., 2010; Jackson et al., 2000). However, due to some contradictory findings and to methodological concerns, Cheryshenko et al. (2009) wrote that it is too soon to draw conclusions about the effectiveness of the FC format in reducing faking behavior. Even though the support for using the FC format to reduce faking is still tentative, it would seem that the context of assessment in CWS justifies the exploration of this format.

Traditionally, for assessments based on the FC format, the test-taker obtains a set number of points as the endorsement of one option results in the non-endorsement of another option, consequently in a multidimensional FC assessment a test-taker cannot score high or low on all assessed

domains. This limitation has led to a criticism of the FC format in that it produces ipsative[2] scores rather than normative data, consequently allowing for only intrapersonal and not interpersonal comparisons. The lack of normative data limits the inferences that can be made on the significance of the test-takers strengths and challenges, precludes the ability to compare individuals with different profiles, and presents difficulty in terms of determining change in the trait(s) of interest. These limitations have historically reduced the value of the FC assessment format. However, recent efforts to incorporate IRT into the FC format, and in particular IRT models based on the ideal point response process, have led a number of researchers to explore the creation of FC assessments, which would provide normative data (Brown & Maydeu-Olivares, 2012; Chernyshenko et al., 2009; McCloy, Heggestad, & Reeve, 2005).

The ideal point response model is theoretically different from the traditional monotonic model used in IRT. The monotonic model (an increasing progression of the item response function as outlined in section 7.1 and shown in Figure 7.1) works well for dominance/cumulative IRT models, which are built on the notion that the probability of an individual selecting the correct or favorable answer to an item increases as his or her latent trait increases (Bortolotti et al., 2013). This model works well with cognitive traits, as these are typically seen as monotonically increasing within a given domain. However, many have argued that, when trying to determine an individual's trait level on a personality measure, the ideal point response model fits the data better (Brown & Maydeu-Olivares, 2012; Drasgow, Chernyshenko, & Stark, 2010; McCloy, Heggestad, & Reeve, 2005; Roberts, Donoghue, & Laughlin, 2000)

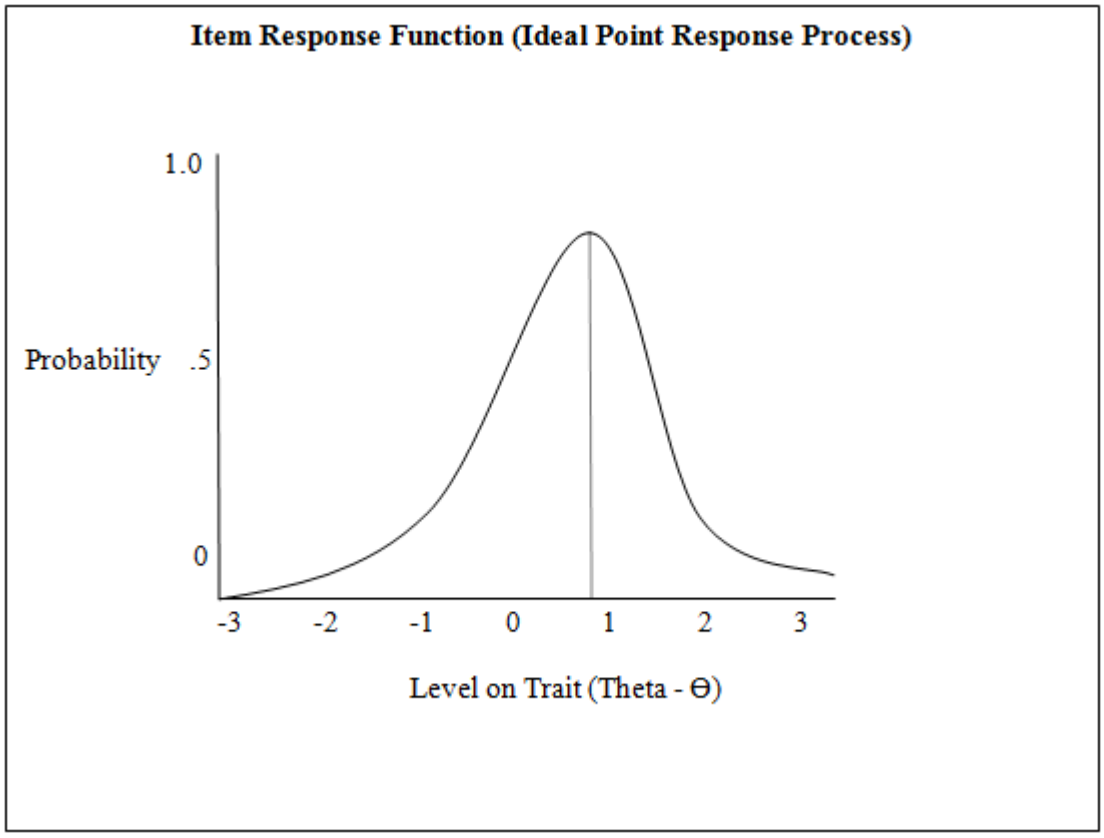
Much of the recent work that has used the ideal point response model has been based on Thurstone's (1927) proposed law of comparative judgment. The law of comparative judgment proposes that, when given a choice between statements, the maximum probability of endorsement occurs when the attitude level of the statement equals or is closest to the attribute level of the individual (Drasgrow

[2] A rank-ordered scale in which the various domains are ranked from most to least endorsed.

et al., 2010; Stark et al., 2006). This is referred to as the ideal point response process. An important implication of this model is that it emphasizes that a test-taker can respond negatively or that a test-taker may not select one of the statements if the individual's trait level is located far above or below the statement (Stark et al., 2006). One of the consequences of this is that information can be obtained about the individual's trait level by examining the statement she or he selected as well as the statement she or he did not select. As with the monotonic item response function, the item response function for an ideal point response process can be best be understood graphically (see Figure 7.2).

Figure 7.2

Sample of an Item Response Function for One Item (Ideal Point Response Process)



In Figure 7.2, the item response function shows the highest probability of the statement being endorsed by an individual with a Θ of about .9. Furthermore, this item response function indicates that

the statement may be rejected both if the individual's Θ is too far below (to the left) or above (to the right) the item location. Clearly, this model would not be expected to fit items in a cognitive domain. For example, it would not be expected that someone with an intermediate skill level in multiplication would have a higher probability of answering a multiplication question correctly compared to someone with very high multiplication skill. However, it has been argued that, when examining personality traits, which require test-takers to use introspection and reflect on their trait level, a non-monotonic curve often fits the test-takers response pattern better (Carter, Lake, & Zicker, 2010; Drasgow et al., 2010; Roberts et al., 2000). For example, someone who is severely depressed may answer "no" to the question, "Sometimes I feel alone", as he or she may always feel alone.

Item response functions tend to be somewhat more complicated to obtain when an ideal point model is used (Brown & Maydeu-Olivares, 2012; Roberts et al., 2000), and it may require a larger number of respondents to validate each items functioning. For example, Roberts et al. (2000) suggested 750 or more respondents to create models that truly fit the data. Given the need for alternative statistical methods and the added labor associated with using an ideal point model, the value of ideal point models has been questioned. Furthermore, it has been argued that, "there is no compelling evidence that either personality scales in general do not measure well in the 'intermediate' range of the construct or that dominance response process models are inadequate" (Reise, 2010, p. 486).

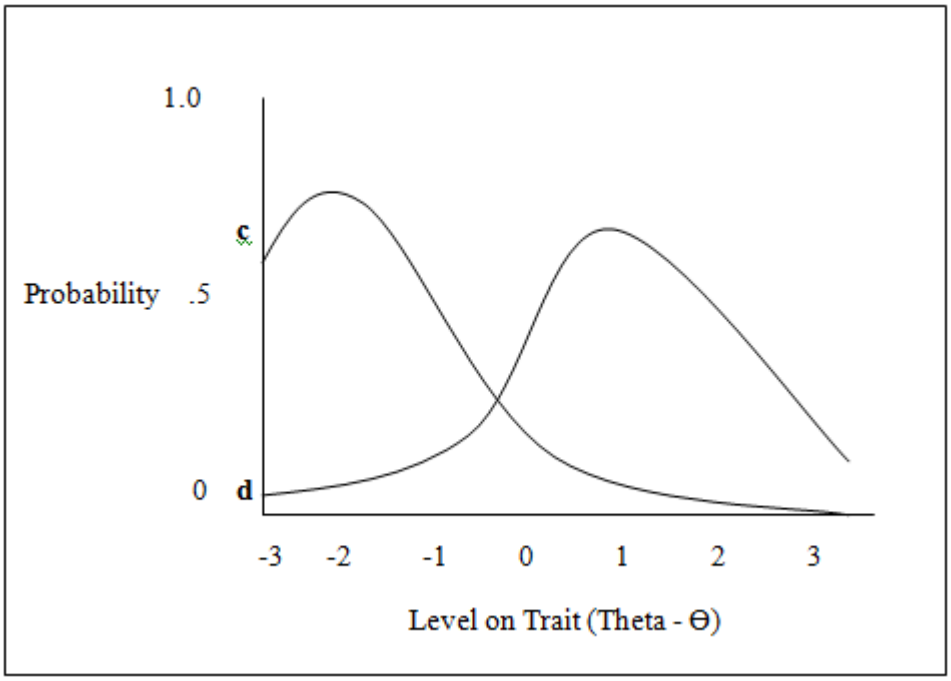
Due to the added complexity and labor needed to create assessments based on ideal point models as well as the abundance of reliable and valid personality measures, the burden to justify the development of assessments built on ideal point models lies with the test developer. Such justification should be based on the extent and severity of the problem as well as a reasonable level of concern about existing measurement instruments for the target population. As approximately 5.9 million children are reported to CWS every year for allegations of maltreatment (Child Maltreatment 2010, 2011) and as several reasons exist to question the value of existing assessment tools for this population,

it appears that assessment for CWS involved clients would warrant the additional effort. Furthermore, concerning the contention of the adequacy of the dominance-response model raised by Reise (2010), the assessment of clients involved with CWS presents a number of concerns that would indicate that the dominance-response model might not be adequate for this population. In particular, previous research has indicated that dominance response models are susceptible to faking (Chistiansen, Burns, & Montgomery, 2005; Dunnette, McCartney, Carlson, & Kirchner, 1962; Jackson, Wroblewski, & Ashton 2000) and as outlined earlier and indicated by the results of Study 1, clients involved with CWS have high motivation to provide misleading information, and it appears that many do in fact provide misleading information.

One of the values of incorporating the ideal point process is that it facilitates the selection of items that lend themselves to the creation of paired preference forced-choice questionnaires. When making a unidemensional FC assessment (measuring just one trait), it is important to know where the highest probability of each statement being accepted/rejected falls, as this allows for the appropriate matching of items. For example, if two statements with different item response functions were presented, then the test-takers selection of one statement over another would provide information regarding the individuals' underlying trait. A unidimensional FC question is illustrated in Figure 7.3. If the test-taker chooses statement **c** over statement **d**, this would indicate that the test-takers level on the trait is likely lower than 0 on the standardized trait continuum but may not be as low as -2 on **Ɵ**. For example, if statement **c** was, "I sometimes feel sad" and statement **d** was "I often feel sad", the selection of statement **c** indicates that the individuals perceive him/herself closer to **c** than **d,** but could still be above or below the maximum information level measured by statement **c.** By examining a series of forced-choice questions on this trait, an increasingly exact location of the test-takers level of the trait can be determined. Chernyshenko et al. (2009) indicated that when forced-choice questions are

measuring unidimensional traits, the ideal distance between the statements is about two standardized units on the underlying IRT scale.

Figure 7.3
Item Response Function for a Unidimensional Forced Choice Item



Although providing an interesting way to measure personality traits, the unidimensional forced-choice format does not do anything to reduce defensive responding, as an individual responding defensively would simply choose the option that represents a more favorable trait level. However, the potential to reduce defensive reporting using this method can be seen with forced-choice multidimensional assessments.

The process of examining multidimensional forced-choice items is a little more complicated. For multidimensional measures, statements addressing different traits are matched on desirability and should differ by less than 1.0 standardized unit on the underlying trait (Chernyshenko et al., 2009). It is believed that by matching items from different dimensions on desirability, the test-takers will be more likely to select the statements most like themselves, as the opportunity to respond in a socially desirable manner will be reduced. It is difficult to represent a figure illustrating an item response function of a

forced-choice multidimensional question because the item response function comparing statements from two traits must be represented in three dimensions. The *y*-axis (vertical axis) would indicate probability of the test-taker preferring one statement over the other given the test-takers trait scores on the respective dimensions and each statements parameters, the *x*-axis would represent one of the traits, and the *z*-axis would represent the other trait (Stark et al, 2012). The test-takers responses are then analyzed using the ideal point process and information is obtained both on the trait underlying the selected statement and on the trait underlying the statement that was not selected. Although this procedure may require longer tests, the increased test length does not appear significant. For example, Chernyshenko et al. (2009) found adequate measures of trait levels when using 32 multidimensional pairwise preference questions for a test assessing three different dimensions. The authors of this study also indicated that the 32 items included 6 unidimensional items, two for each dimension. The authors noted that although there is currently no mathematical proof of this theory, it is believed that it might not be possible to recover normative data without a small percentage of unidimensional items to anchor the metric. Stark et al. (2012) found that as few as 5% of unidimensional items could provide a sufficient anchor.

An important consideration in creating multidimensional pairwise preference questions is that the paired statements must be matched in terms of desirability. Dalal, Withrow, Gibby, and Zicker (2010) noted that guidelines for matching items in terms of their desirability have not been developed yet. However, test makers have used some general procedures that seem to be promising. One way is to match statements in different domains in terms of their ideal point response function. Additionally, researchers have used social desirability rating, having either test-takers or judges rate items on social disability and then match statements accordingly (Chernyshenko et al., 2009; Jackson et al., 2000). Chernyshenko et al. (2009) and Jackson et al. (2000) also indicated that they have used a combination of

the ideal point item response function and social desirability rating in the construction of their instruments; however, they did not provide details on how they combined the two measures.

Although this dissertation does not attempt to create a forced-choice multidimensional assessment, the results of Study 2 may provide insights into how statements from different domains might be paired. Furthermore, it is valuable to be aware of the various advantages of IRT in the creation of an assessment for clients involved with CWS.

Chapter 8

Methodology for Study 2

8.1 Participants

In 2009, a program was introduced in the state of Washington to assist with the assessment of children at high risk for future maltreatment who remained in their family home. The program called the Comprehensive Assessment Program (CAP) is administered to families with open CWS cases and is accessed by the caseworker assigned to these cases. The CAP program uses a number of standardized assessments to determine caregivers' functioning, including the CTS-2, the ASSIST 3.0, and the PSI-SF. The participants in this study were those caregivers who completed the CAP assessment.

Between 8/2009 and 7/2013, 235 families and 318 caregivers completed the CAP assessment. As mentioned, one of the assessment tools used during the CAP assessment was the PSI-SF. The PSI-SF includes a defensive responding scale. Based on the results of Study 1, which signified that the various assessment tools provided both concurrent and predictive validity for caregivers who reported in a non-defensive manner on the PSI-SF, only those caregivers were included in this study. This resulted in 177 families and 239 caregivers being included in the analysis. Seventy six percent of these caregivers were female, and the caregivers had a mean age of 32 years (SD = 9.4). Seventy four percent of the caregivers were the primary caregivers, and 26% were secondary caregivers.

When a family was referred for a CAP assessment, a Target Child was identified. The Target Child was the child in the family home identified as being most at risk for maltreatment and/or being most challenging for the caregiver(s). A greater percentage of target children were male (51.5% male and 48.5% female). The mean age of the children was 6.3 years (SD = 4.7). Table 8.1 summarizes the ethnicity of the children included in the study.

Table 8.1
Ethnicity of Children in the Study2

	Frequency	Percent
Caucasian	72	40.7
African American	34	19.2
Asian	4	2.3
Hispanic/Latino/Mexican	13	7.3
Multi Racial	18	10.2
Native American	19	10.7
Other/unable to determine	15	8.5
Nat Haw/Pac	2	1.1
Total	177	100.0

8.2 Measures

As the CAP program uses 7 measures (with 8 subscales) (Table 4.1), a decision was made to focus on just those measures that contributed unique significant predictive value to the final Cox proportional hazard model (Table 5.13). This included the Parent Stress Index-Parent/Child Dysfunctional Interaction scale (PSI-PCDI), the CTS-2 (domestic violence screener), and the ASSIST 3.0 (substance abuse screener). However, after examining the items on the ASSIST 3.0 and looking at the response patterns, a determination was made not to include the ASSIST 3.0. The format of the ASSIT 3.0 requires individuals to first indicate whether they have ever used each of the nine different substances. The remaining questions follow up on participants' affirmative responses. However, many of the substances had a very low response rate. Consequently, many of the questions had very few responses, making it impossible to do IRT analysis on many of the items. There had been one previous IRT study of the ASSIST 3.0 (Ali, Meena, Eastwood, Richards, & Marsden, 2013), but these researchers had a large sample size of over 2000 participants. Due to the limitations in the data available for this study, IRT analysis of the Assist 3.0 was not done. This resulted in IRT analysis being done with two of the measures included in the CAP assessment, the CTS-2 and the PSI-PCDI.

Conflict Tactics Scale-2 (CTS-2)

The Conflict Tactics Scale-2 (CTS-2) is used to identify domestic violence/intimate partner violence. The CTS-2 is the most widely used instrument for assessing intimate partner violence, with strong evidence supporting both the reliability and validity of the measure (Straus, Hamby, Boney-McCoy, & Sugarman, 1996). The CTS-2 is based on conflict theory, which assumes that conflict is an inevitable part of all human association; however, violence as a tactic to deal with conflict is not (Straus et al., 1996). The CTS-2 includes 5 subscales, physical assault, psychological aggression, negotiation, injury, and sexual coercion. The CAP assessment utilizes the physical assault and injury scales of the assessment. The developers of the CTS-2 indicated that it is acceptable to use selected scales when the length of assessment is an issue or when just particular types of conflict are of interest (Straus et al., 1996; Straus & Douglas, 2004). The abuse subscale consists of 12 physical assault questions. Five of these questions assess minor assault, such as "I pushed or shoved my partner", and 7 questions address severe assault, such as, "I beat my partner up". The injury scale includes 2 minor injury questions, such as, "I had a sprain, bruise or small cut because of a fight with my partner", and 4 severe injury items such as, "Had a broken bone from a fight with my partner". Each of the questions is also administered to the respondent as the perpetrator rather than the victim (e.g., "My partner has done this to me"). This resulted in 36 items scored on an 8-point scale indicating the frequency of behavior. Zero indicates never, 1 thru 6 indicate varying frequencies over the last year, and a 7 indicates that it has occurred but not in the previous year. There are various ways the CTS-2 can be scored (Straus et al., 1996; Straus & Douglas, 2004; Straus, 2004).

In Study 1, responses were dichotomized into occurred or did not occur, and if an event occurred, was it in the previous year or sometime prior? However, only few participants indicated more severe measures of aggression in the previous year. For example, 13 of the 22 severe violence questions

received 5 or less affirmative responses for the previous year, making IRT analysis for these items difficult. Hence, for this study a determination was made to split the responses into the categories of "ever victim" and "ever perpetrator", thus creating two data sets. In addition to maximizing the use of the available data to establish the trait level taped by the various items, separating the data in this way also has theoretical implications, as it is possible that the risk of child maltreatment is different for perpetrators than it is for victims of domestic violence.

Parenting Stress Index-Parent/Child Dysfunctional Interaction

Caregivers referred to the CAP were administered the Parenting Stress Index-Short Form (PSI-SF). Based on factor analyses of the Parent Stress Index, a shorter version, the PSI-SF, was developed. The PSI-SF is a brief, 36-item self-report measure of parenting stress. The index provides a total score and three subscale scores, Parental Distress, Parent-Child Dysfunctional Interaction, and Difficult Child. Each subscale contains 12 items measured on a 5-point scale: *Strongly Agree, Agree, Not Sure, Disagree,* and *Strongly Disagree*. As only the Parent-Child Dysfunctional Interaction (PSI-PCDI) contributed unique significant predictive value to the final Cox proportional hazard model, IRT analysis was only done with this subscale. The alpha reliability coefficients of the PSI-PCDI have been shown to be .80 (Kelley, 1998).

Scores at or above the 85th percentile on the PSI-PCDI are considered significant. This represents a score of 26 or higher on the PSI-PCDI. Eighty one (33.9%) of the caregivers classified as non-defensive scored at or above the 85th percentile, while 158 (66.1%) scored below this cut score. In Study 1 those scoring at or above the 85th percentile were considered indicated for the analysis using the Cox proportional hazard model. However, as one of the questions to be examined in this study is if a shorter version of the PSI-PCDI could be created that functioned as well as the original version, this prohibited the use of the instruments established cut score. Consequently, a determination was made

to set the cut score resulting from the IRT analysis at the same percentage as was found when using the established scoring matrix (about 34%).

Reports of Child Maltreatment after the CAP Assessment

The CWS case attached to the primary caregiver identified in the CAP database was also examined. The reports of child maltreatment linked to the caregivers' case were included for this study if the report alleged some form of child maltreatment by a caregiver that was accepted for some form of intervention. This includes reports that were accepted for an investigation of child maltreatment as well as those accepted for an alternative response, such as an intake worker calling the family to discuss the concerning information without initiating a formal investigation. All calls of concern received by CWS, regardless of whether they received a response, are documented in the computer system by the intake workers and attached to the family's case file; however, the cases that do not cross a pre-established threshold of maltreatment are entered into the system for information purposes only. Consequently, there are no criteria for the level of concern in the information only referrals; therefore, information only reports were not included in this study.

As the reports of child maltreatment included in this study were based on the primary caregiver's case file, this would include all reports on the family, regardless of whether or not the Target Child was involved in the allegations. Due to resource limitations, it was not possible to indicate which caregiver in the family home was alleged for what type of maltreatment (e.g., child abuse or child neglect).

When examining reports of child maltreatment, some researchers have focused on substantiated reports (Fluke, Yuan, & Edwards 1999; Solomon & Asberg, 2012) while others have used reports of maltreatment regardless of substantiation as the measure of interest. Kohl, Johnson-Reid, and Drake (2009) demonstrated that the risk of recidivism is similar for both substantiated and

unsubstantiated cases. Furthermore, English, Marshall, Brummel, and Orme (1999) argued that many factors unrelated to whether or not a child has been abused or neglected influence the substantiation of a report of child maltreatment. For example, work load, resources, office and individual worker practice, and standards of proof all influence the substantiation of allegations. Therefore, this study examined reports of maltreatment data regardless of substantiation. Using the described criteria, 66.7% (n=118) of the families had a new report of child maltreatment after the CAP assessment. Time till the first report received after the CAP assessment was used in the Cox proportional hazard model.

Placement of Target Child

As noted earlier, the CAP identified one of the children in the home as the Target Child. CWS case records were examined to determine the dates of placement in out-of-home care of this child. Placements were included if the removal occurred through a court order, which typically occurs when the CWS caseworker petitions the court requesting authority to place the child in out-of-home care. Additionally, placements were recorded if the placement occurred under a voluntary placement agreement, which are typically obtained when it is determined that a child is not safe in the family home and the parent and caseworker agree that the child should be temporarily placed in out-of-home care. Placements were also included if Law Enforcement placed the child into the protective custody of CWS, resulting in the child being removed from the family home. Since all three of these methods of removal are based on risk to the child in the family home, a decision was made to include all three in this study. Using this criteria, 21.5% (n=38) of the target children were placed in out-of-home care after the CAP assessment. Time till placement of the child in out-of-home care after the CAP assessment was used in the Cox proportional hazard model.

8.3 Analytic Strategy

CTS-2

The items on the CTS-2 were examined with a 2PL IRT model (two-parameter logistic model) using Xcalibre 4.2 (Assessment Systems Corporation, 2012)[3]. The 2PL model was chosen as the results of the model includes the Θ (trait levels) for the caregivers who completed the assessment, the *a*-parameters that represent each item's discrimination, and the *b*-parameters that represent each item's location on the underlying Θ scale. While the 1PL model could have been chosen, it would not have included the *a*-parameters which were of interest in this study. The 3PL model would have been inappropriate as it adjusts the item response functions for guessing; something that must be addressed in cognitive assessments, but is likely not an issue in personality assessments. Maximum Likelihood (MLE) was used to estimate Θ. The proficiency distribution was estimated empirically using 20 quadrature points, and the items were centered around Θ.

PSI-PCDI

The items on the PSI-PCDI were calibrated with the Samejima's Graded Response Model (SGRM) using Xcalibre 4.2 (Assessment System Corporation, 2012). A limitation of using this model is that best practice would indicate having at least double the number of participants as are included in this analysis (Demares, 2010). However, the partial credit model, which may require slightly fewer participants, assumes that each item is equally related to the underling trait. This was not an assumption that was being made during this analysis. Furthermore, as the PSI included five options per item, it is not clear that a partial credit model would produce more accurate results with the available sample size (Embretson & Reise, 2000). Consequently, the SGRM model was selected. Maximum Likelihood was used to estimate Θ. The score distribution was estimated using 20 quadrature points, and the items were centered around Θ.

[3]I would like to thank Applied Psychological Measurement Inc. for their generous grant allowing for the purchase of Xcalibre 4.2.

Survival Analysis

As in Study 1, Survival analysis was done using the Cox proportional hazard regression program in SPSS-19. Time to either first new report of maltreatment or placement of the Target Child after the CAP assessment was entered in the time function. 68.4% (n=121) of the families had either a new report of child maltreatment or placement of the Target Child in out-of-home care. The time to event represented the outcome variable in the model (Kleinbaum & Klein, 2012). The event was also entered as either occurring (coded as 1) or not occurring (coded as 0). Time in the study was determined for each participant, with the study time ending either when an event occurred (i.e., new report or placement) or when the observation period ended (in this case, when the data was pulled from the CWS system in October 2013). Thus, a strength of survival analysis is that it maximizes the use of available information. The covariates included in the Cox proportional hazard model were scores (rankings of individuals on the underlying trait) obtained from IRT analysis.

8.4 Study 2 Research Questions

The following questions, most of which are questions typically examined during IRT analysis (DeMars, 2010), were examined for each of the assessment tools.

1) What is the spread of item locations (and category locations for polytomous items)?

   Examining the location (*b*-parameter) of the items on the assessment provides information regarding Θ levels at which the assessment provides information. If the purpose of the assessment is to accurately measure the full range of the trait, then there should be items positioned along the entire range of the Θs. On the other hand, if the purpose of the assessment is to separate examinees above and below a particular cut point, then the items should cluster around this pre-determined score. With polytomous items (items with multiple response options), the distance between each of the response options for each item can be examined. This allows information to be

gathered for each response category, and based on the distance between the categories, to assist in determining whether some categories that provide less information could potentially be eliminated from the assessment. Additionally, in the creation of a forced-choice assessment, the item difficultly can be used as one way to inform the pairing of items from different domains to ensure the matching of items with similar locations on the underlying traits.

2) How discriminating is each item?

The item discrimination parameter (*a*-parameter) indicates the probability of selecting one response over another response at the *b*-parameter. The greater the slope at the *b*-parameter, the better the item distinguishes between those that are above the *b*-parameter and those that are below the *b*-parameter. Consequently, items with steep slopes provide more information than do those with shallow slopes. A very shallow slope would indicate an item that does not distinguish well between different Θs, and may indicate an item that should be removed from the assessment. For polytomous items, the item's discrimination can also refer to the maximum information obtained for the various response options, again with steeper curves indicating greater discrimination between the options.

3) How much information does the test provide over the trait range?

The test information curve is a representation of the assessment's coverage of the range of potential trait levels. The test information curve depends on the item spread (question 1) and item discrimination (question 2). Again, if the goal is to determine whether people are above or below a certain level on a trait (e.g., a cut score), then the test information should peak around this cut score. However, if the goal is to have information available for as much of the trait as possible, then the test information curve should be more spread-out, covering more of the trait. Additionally, test information distributed across the trait range is helpful when measuring change scores, and if

interactions between traits were suspected, then it would be helpful to have items representative across a wide range of Θ.

4) For the population examined, how reliable are the trait estimates?

This is represented as the inverse of the test information curve (referred to as the conditional standard error of measurement), in that the standard error is lower where the test provides greater information. The lower standard errors are typically somewhere near the middle of the trait level while moving towards the extremes, the error usually becomes greater. An examination of the conditional standard error of measurement is useful for interpreting scores, as it indicates how much confidence can be placed in resulting scores at different levels of the trait.

5) Do the results of the IRT analysis allow for the creation of a shorter assessment that does not adversely affect the assessment's ability to predict new reports of child maltreatment/placement of the Target Child into alternative care?

One of the benefits of IRT is that it provides a way to flag poorer performing items. Additionally, by examining the results of the analysis, items that appear to be providing redundant information may be removed. By identifying these poorer performing and redundant items, it is possible that a shorter assessment can be created which performs as well as or perhaps even better compared to the original assessment. Study 2 includes an examination of how the removal of selected items influences the assessment's ability to predict future reports of child maltreatment/placement of the Target Child into out-of-home care.

Chapter 9

Results – Study 2

Conflict Tactics Scale – 2

9.1 Item Unidimensionality

An important assumption that must be tested prior to undertaking IRT analysis is that of unidimensionality. To test for unidimensionality, the 18 items from the victim and the perpetrator data set were entered separately into the DIMTEST (Stout, 2005) program, which is designed to test for unidimensionality. The test for unidimensionality was run in DIMTEST by creating a subset of items that appear dimensionally similar to each other and somewhat distinct from the other items. As the CTS-2 items are split into mild incidence and severe incidence of violence, it was determined that the subset would consist of the 7 mild incidence items for both the victim and perpetrator data set. The DIMTEST analysis then checks whether participants respond to the subset of items in a manner indicative of a different dimension, thus violating the assumption of unidimensionality. A p-value of 0.5621 was found for the victimization data and a p-value of 0.3596 was found for the perpetrator data, indicating that the assumption of unidimensionality was not violated.

9.2 Results for the Victim Data

For the victim data, the estimations converged after 11 loops. Table 9.1 provides the summary statistics for the total scores. The Alpha of 0.958 indicates that the assessment has high internal consistency. Table 9.2 shows that the overall model fit is non-significant at a p-value of 0.95, indicating that the model fits the data well. Table 9.3 gives an overview of the functioning of each of the 18 items. The p_i values shown in the third column indicate the percentage of respondents who indicted that this behavior had occurred. Not surprisingly, the incidence of caregivers reporting victimization of mild violence/injury was greater than that of severe violence/injury. The $r_{i\Theta}$ in the fourth column represents

the Pearson's r correlation of item responses with full test Θ. The a-parameter represents the item's discrimination, with a larger value indicating that the item differentiates between examinees above and below the item's location on the Θ scale. The item discriminations ranged from 0.788 (item 15) to 1.791 (item 3), indicating that the items provided adequate discrimination. The standard error for the a-parameter is shown in the column to the right of the a-parameter. Table 9.4 provides summary statistics for the a-parameter and b-parameter for all calibrated items.

Table 9.1
Summary Statistics (Victim Scores)

Test	Items	Alpha	Mean	SD
Full Test	18	0.958	3.954	5.533

Table 9.2
Overall Model Fit (Victim Scores)

Test	Items	Chi-square	Df	P	-2LL
Full Test	18	283.365	324	0.950	1302

Table 9.3
Item Parameters (Victim Scores)

Item #	Item Type	p_i	$r_{i\Theta}$	a	(SE)	b	(SE)	Flag(s)
1	Mild Violence	0.321	0.777	1.414	(0.194)	-0.539	(0.113)	
2	Mild Violence	0.245	0.793	1.472	(0.194)	-0.077	(0.109)	
3	Mild Violence	0.414	0.766	1.791	(0.210)	-1.094	(0.110)	F
4	Mild Violence	0.342	0.822	1.615	(0.193)	-0.677	(0.107)	
5	Mild Violence	0.278	0.775	1.195	(0.196)	-0.262	(0.125)	
6	Mild Injury	0.308	0.835	1.548	(0.194)	-0.474	(0.106)	
7	Mild Injury	0.308	0.847	1.612	(0.194)	-0.477	(0.104)	
8	Severe Violence	0.131	0.678	1.087	(0.183)	0.869	(0.153)	
9	Severe Violence	0.287	0.811	1.272	(0.195)	-0.325	(0.120)	
10	Severe Violence	0.165	0.723	1.071	(0.185)	0.574	(0.145)	
11	Severe Violence	0.245	0.858	1.604	(0.195)	-0.091	(0.103)	
12	Severe Violence	0.219	0.819	1.350	(0.190)	0.100	(0.117)	
13	Severe Violence	0.046	0.489	1.105	(0.197)	1.959	(0.225)	
14	Severe Violence	0.143	0.700	1.097	(0.184)	0.747	(0.148)	
15	Severe Injury	0.114	0.568	0.788	(0.179)	1.237	(0.203)	
16	Severe Injury	0.160	0.640	0.855	(0.189)	0.718	(0.172)	
17	Severe Injury	0.165	0.718	1.073	(0.185)	0.573	(0.145)	
18	Severe Injury	0.064	0.473	0.848	(0.176)	1.860	(0.237)	

Table 9.4
Summary Statistics for All Calibrated Items (Victim Scores)

Parameter	Items	Mean	SD	Min	Max
a	18	1.267	0.299	0.788	1.791
b	18	0.257	0.865	-1.094	1.959

The b-parameters in Table 9.3 indicate the items location on the Θ continuum, with the location of b representing where there is a 50% likelihood of the respondent indicating that the statement had occurred. A higher b-parameter (>1.0) indicates the item is representative of a higher trait level while a lower b-parameter (<-1.0) indicates an item lower on the trait. The standard error for the b-parameter is in the next column to the right. Item 3, which had a b of -1.094, measured the lowest level of the trait. This mild violence item is asking the respondents to indicate whether they had ever been pushed or shoved by their partners. Item 13, with a value of 1.959, had the highest b. This severe violence question is asking the respondents to indicate whether they had ever been burned or scalded by their partners on purpose. The pattern of the b-parameters being centered near a Θ of 0 can be seen in the test information function, which indicates that maximum information was provided at a Θ of -0.30 (Figure 9.1). The test information function also indicates that slightly more information is present in the upper ends of the scale than in the lower ends of the scale. The conditional standard error of measurement is the inverse of the test information function and is shown in Figure 9.2. The conditional standard error of measurement indicates that the standard error is lowest at Θ -0.30 and increases at the extremes.

Figure 9.1

Test Information Function (Victim Scores)

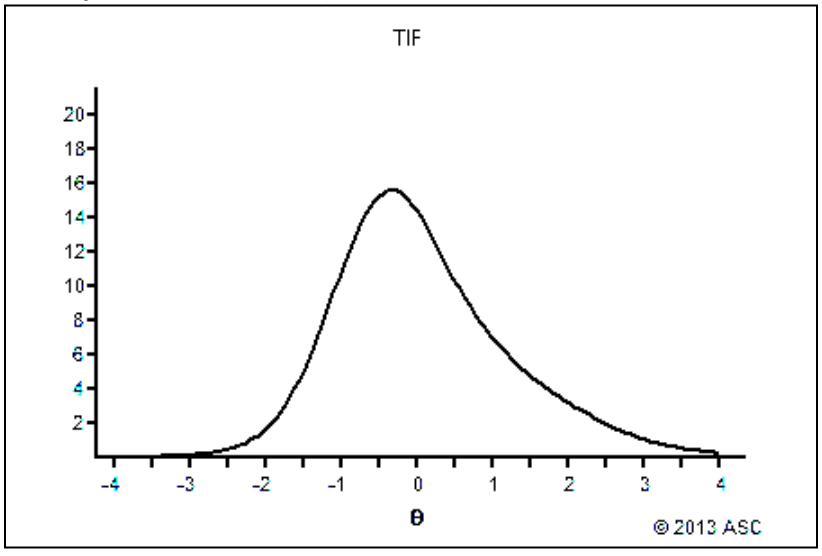
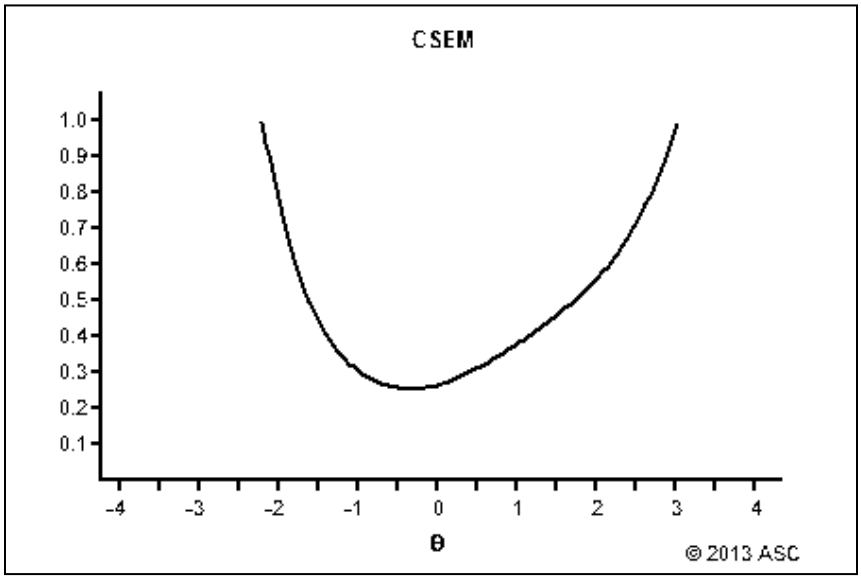


Figure 9.2
Conditional Standard Error of Measurement Function (Victim Scores)



The "F" in the last column of Table 9.3 indicates that the item fit statistic (z residual) for item 3 was significant (0.037). The z residual is a standardized chi-square statistic and is used to evaluate the significance of item misfit for dichotomous items (Guyer, & Thompson, 2013). This suggests that the item did not fit the IRT model and perhaps, it should be removed from the assessment. Appendix D provides the item questions, the complete test statistics, the item information curves, and the item statistics for each of the items on the CTS-2 victimization scale.

9.3 Results for the Perpetrator Data

For the perpetrator data, the estimations converged after 22 loops, indicating that the data did not fit the 2PL model quite as well as did the victim data which converged after 11 loops. Table 9.5 provides the summary statistics for the perpetrator scores. The Alpha of 0.890 again indicates that the assessment had high internal consistency. The total score mean of 1.722 indicates that less than half the number of items were reported as having been perpetrated by the test-takers as compared to those that reported being a victim of domestic violence which had a mean of 3.954 (Table 9.1). Table 9.6

shows the overall model fit (p-value 0.488), indicating that the model fit the data well. Table 9.7 gives an overview of the functioning of each of the 18 items. Consistent with the victim scores, more caregivers reported being perpetrators of mild violence/injury rather than severe violence/injury. Moreover, it is clear from the b-parameters that severe violence items had a much higher level of Θ compared to items measuring milder forms of violence. The a-parameters indicated that in general, the item discrimination was lower for the perpetrator items than it was for the victim items (also see Table 9.4 compared to Table 9.8). The "Hb" flags in Table 9.6 highlight items with high (>3.0) *b*-parameters, indicating that these items may not add much information to the assessment, as these items only distinguish between test-takers that are 3 standard deviations from the mean score. Additionally, item 3 did not fit the model well, as the z residual was significant (0.009).

Table 9.5
Summary Statistics (Perpetrator Scores)

Test	Items	Alpha	Mean	SD
Full Test	18	0.890	1.722	2.993

Table 9.6
Overall Model Fit (Perpetrator Scores)

Test	Items	Chi-square	df	P	-2LL
Full Test	18	324.115	324	0.488	1051

Table 9.7
Item Parameters (Perpetrator Scores)

Item #	Item Type	p_i	$r_{i\Theta}$	a	(SE)	b	(SE)	Flag(s)
1	Mild Violence	0.194	0.600	0.813	(0.207)	0.248	(0.174)	
2	Mild Violence	0.080	0.514	0.724	(0.178)	1.615	(0.243)	
3	Mild Violence	0.316	0.612	1.052	(0.684)	-0.707	(0.316)	F
4	Mild Violence	0.232	0.637	0.935	(0.768)	-0.119	(0.232)	
5	Mild Violence	0.169	0.665	0.927	(0.831)	0.383	(0.169)	
6	Mild Injury	0.114	0.584	0.743	(0.886)	1.095	(0.114)	
7	Mild Injury	0.093	0.630	0.802	(0.907)	1.293	(0.093)	
8	Severe Violence	0.042	0.497	0.673	(0.174)	2.456	(0.327)	
9	Severe Violence	0.156	0.609	0.729	(0.200)	0.651	(0.198)	
10	Severe Violence	0.038	0.472	0.682	(0.175)	2.552	(0.338)	
11	Severe Violence	0.063	0.555	0.709	(0.176)	1.899	(0.267)	
12	Severe Violence	0.046	0.608	0.840	(0.184)	2.049	(0.263)	
13	Severe Violence	0.008	0.479	0.790	(0.232)	3.717	(0.588)	Hb
14	Severe Violence	0.076	0.565	0.712	(0.178)	1.668	(0.249)	
15	Severe Injury	0.017	0.450	0.717	(0.195)	3.285	(0.459)	Hb
16	Severe Injury	0.030	0.377	0.614	(0.173)	3.004	(0.407)	Hb
17	Severe Injury	0.025	0.545	0.771	(0.190)	2.782	(0.364)	
18	Severe Injury	0.021	0.437	0.685	(0.186)	3.146	(0.429)	Hb

Table 9.8
Summary Statistics for all Calibrated Items (Perpetrator Scores)

Parameter	Items	Mean	SD	Min	Max
a	18	0.773	0.109	0.614	1.052
b	18	1.723	1.273	-0.707	3.717

Examining the item information functions separately for each item provides information that can help determine how an item is performing; an example of this can be seen when comparing Figure 9.3 with Figure 9.4. Figure 9.3 shows the item information function for item 2 on the victim scale and Figure 9.4 shows the item information function for item 8 on the perpetrator scale. When comparing these two figures, it is apparent that the curve in Figure 9.3 is steeper than that in Figure 9.4, indicating the greater item discrimination of item 2. Moreover, the b-parameter for item 8 on the perpetrator scale is significantly further to the right than the b-parameter for item 2 on the victim scale, indicating that this item measures a higher level of the trait.

Figure 9.3
Item Information Function for Item 2 on the Victim Scale
"My partner threw something at me that could hurt"

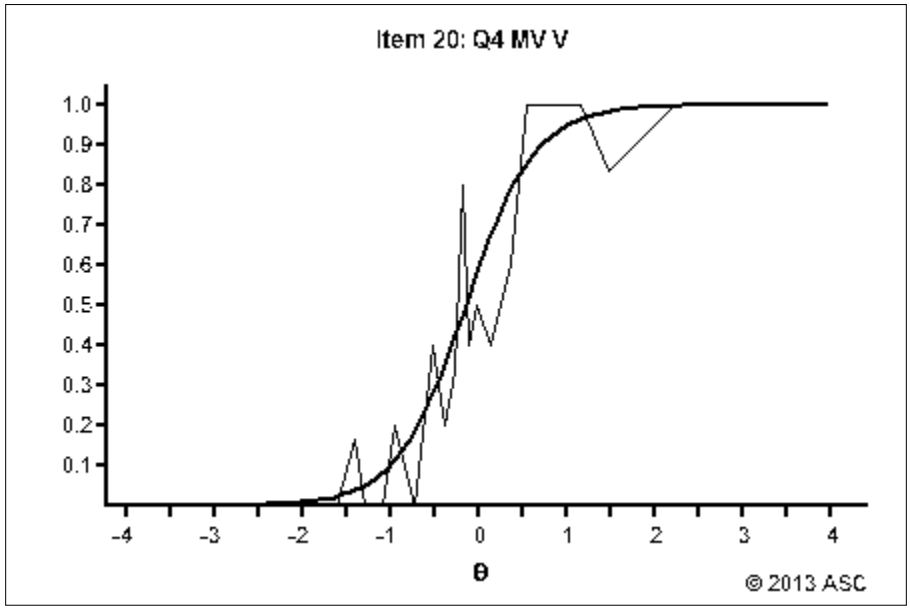
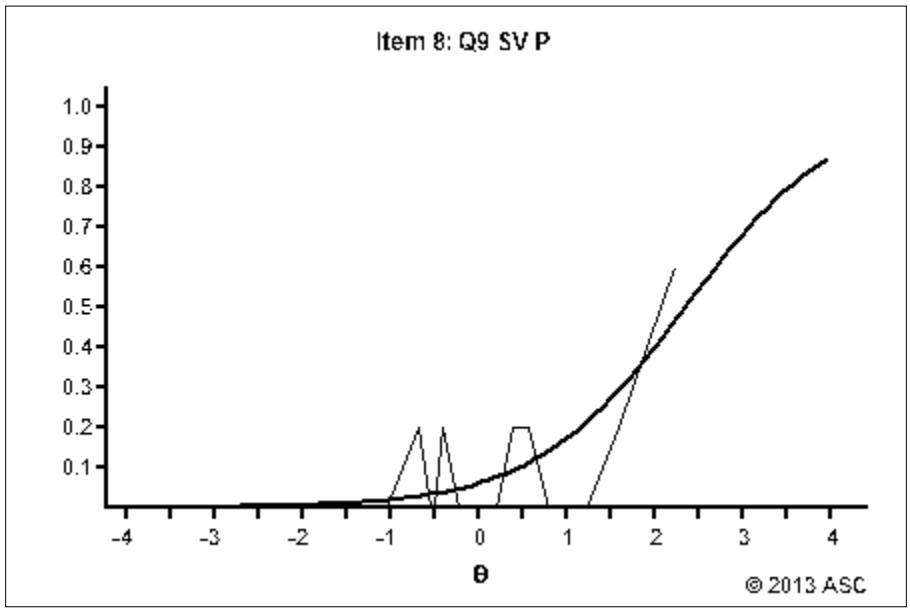


Figure 9.4
Item Information Function for Item 8 on the Perpetrator Scale "I used a knife or gun on my partner"



The *b*- parameters in general are centered at a higher Θ for the perpetrator items than for the victimization data. This can also be visually observed from the test information function, which indicates that maximum information is provided at a Θ of 1.60 (Figure 9.5). The test information function also indicates that more information is present in the upper end of the scale than in the lower end of the scale. Figure 9.6 shows the conditional standard error of measurement. When compared to the conditional standard error of measurement for the victim data, it can be seen that scores on the perpetrator scale should be interpreted with less certainty, as the standard errors for the perpetrator data are greater. Appendix E provides the item questions, test statistics, item information curves, and the item statistics for each of the items on the perpetrator scale.

Figure 9.5
Test Information Function (Perpetrator Scores)

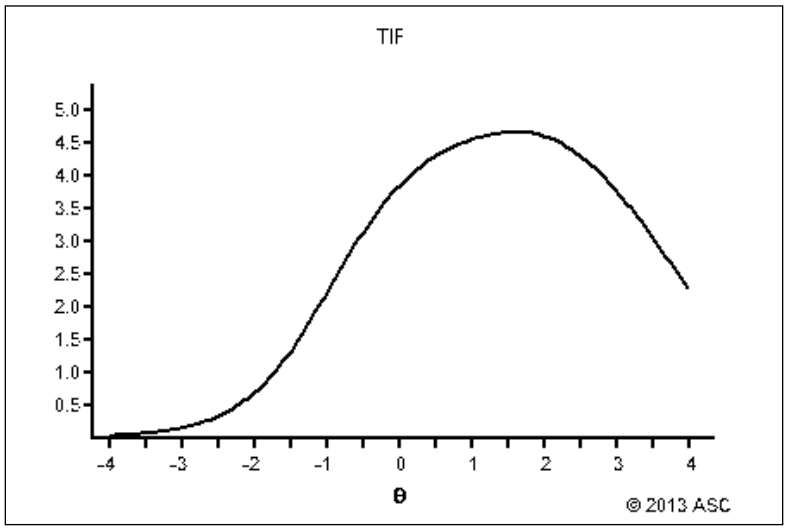
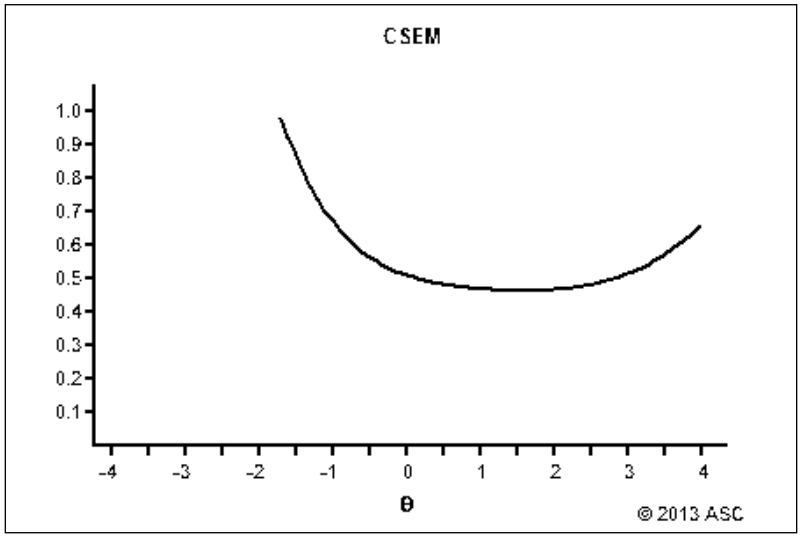


Figure 9.6
Conditional Standard Error of Measurement (Perpetrator Scores)



In summary, the model fit the victim data better compared to the perpetrator data, although the data fit was adequate in both instances. Further, the results indicated that, item 3 for the victim data and item 3 for the perpetrator data should be considered for deletion as they did not fit the 2PL model well based on the z residuals. The results also indicated that items 13, 15, 16, and 18 on the perpetrator scale might be considered for deletion, as they potentially added little information because they were positioned so high on the underlying Θ scale. The next phase of the analysis reports on the

results obtained after removing items from the scales. This was done to examine the functioning of the scales as predictor variables of future child maltreatment after modifying the scales based on IRT principals.

9.4 Scale Reduction and Functioning

The number of items on the victim and the perpetrator scales were reduced from 18 to 10. Although the decision to reduce the scales to 10 items was somewhat arbitrary, it was determined that 10 items represented a reasonable balance between scale length and coverage of the trait. Items were considered for removal if they were flagged during IRT analysis for poor model fit or if their b-parameters were significantly high or low. Also, the determination of which items to remove was based on the item statistics, including the a-parameter (higher a-parameters were given more weight) and the impact on the scales alpha level if the item was removed. As the item statistics were similar for a number of items, decisions were made to remove the redundant items that were not functioning quite as well. Additionally, as the scores were going to be used to predict future child maltreatment, a decision was made to ensure adequate coverage of the upper range of Θ on both the victim and perpetrator scale.

For the victim scale, item 3 was removed, as it appeared to not fit the model well. This resulted in seven more items needing to be selected for removal in order to reach the goal of testing a 10 item scale. Based on the guidelines outlined earlier, the seven additional items selected for removal from the victim scale were items 1, 2, 5, 6, 10, 16, and 18 (see Appendix D for the items). The final set included 1 mild violence item, 1 mild injury item, 6 severe violence items, and 2 severe injury items.

On the perpetrator scale, item 3 was removed, as it did not fit the model well. Three of the four items measuring severe injury (items 15, 16, and 18) on the perpetrator scale had very high locations on the underlying Θ scale (> 3). A determination was made to keep the best performing of these items

based on the test statistics to ensure coverage of the perpetration of sever violence. This resulted in item 15 being included in the final scale and items 16 and 18 being removed. Next, three of the 6 severe violence items were removed. Item 13 was removed as it had a very high location on the underlying Θ scale (3.717), and there were a number of other severe violence items that appeared to cover upper ranges of Θ. Items 8 and 10 were also removed from the severe violence scale as these were the poorest performing of the remaining items. The poorer performing of the two mild violence items (item 6) was then removed. Lastly, item 2 was removed as based on the selection criteria this was the poorest performing of the remaining mild violence items. The resulting 10 items on the perpetrator scale included 3 mild violence, 1 mild injury, 4 severe violence, and 2 severe injury items.

The same IRT model outlined in section 9.2 was used to analyze the new victim and perpetrator data sets. Table 9.9 provides the summary statistics for the total scores for the victim data. The alpha of 0.928 (Table 9.9) is close to the alpha of 0.958 found prior to the removal of 8 items. Table 9.10 shows the overall model fit; although the model fit fell from 0.950 (Table 9.2) to 0.237, the model still fit the data reasonably well. The results for the perpetrator data indicated that the alpha fell slightly from 0.890 to 0.849 (Table 9.11). Additionally, after removing the 8 items, the model fit became significant (p-value < 0.00) (Table 9.12), dropping from 0.488 (Table 9.6). This was likely a result of the reduction in the number of items on the scale, and indicates that the results should be interpreted with caution.

Table 9.9
Summary Statistics (Victim Scores-10 Items)

Test	Items	Alpha	Mean	SD
Full Test	10	0.928	2.000	3.038

Table 9.10
Overall Model Fit (Victim Scores- 10 Items)

Test	Items	Chi-square	df	P	-2LL
Full Test	10	141.208	130	0.237	604

Table 9.11
Summary Statistics (Perpetrator Scores-10 Items)

Test	Items	Alpha	Mean	SD
Full Test	10	0.849	1.072	1.961

Table 9.12
Overall Model Fit (Perpetrator Scores- 10 Items)

Test	Items	Chi-square	df	P	-2LL
Full Test	10	198.514	130	0.000	597

The next step of the analysis required examining the obtained *theta* score from the IRT analysis for each of the assessed caregivers. Based on their *theta* scores, the caregivers were given a percentage rank from 0-100, with 0 indicating low on the trait and 100 indicating high in the trait. A 70% threshold was established, and those with *theta* scores in the top 30% were classified as 'indicated'. Survival analysis was then conducted again, as in the first study, with the event occurring or not occurring as a new report of child maltreatment or placement of the Target Child into out-of-home care after the CAP assessment. Time was entered as the families' length of time in the study. The results for this analysis are displayed in Table 9.13. Even though 45% of the items were removed from both the victim scale and the perpetrator scale, the reduced scales functioned similar to the full scales. Additionally, it is also interesting to note that the CTS-2 perpetrator measure used in this analysis had an Exp B of 1.738 (and 1.783 when the scale was reduced to 10 items), which was a stronger predictor of future maltreatment/placement in out-of-home care than was found in Study 1 using domestic violence in the previous year (Exp B = 1.497) and domestic violence ever (Exp B = 1.455) as the variables of interest (From Table 5.16).

Table 9.13
Survival Analysis of CTS-2 Scores Based on IRT Analysis with all Items and with 10 Items

Measure	Number	Significance	Hazard Ratio - Exp(B)	95% Confidence Interval for Exp (B)	
				Lower	Upper
CTS-2 Victim (All Items)	175	.045	1.475	1.008	2.104
CTS-2 Victim (10 Items)	175	.054	1.434	.991	2.075
CTS-2 Perpetrator (All Items)	175	.004	1.738	1.199	2.519
CTS-2 Perpetrator (10 Items)	175	.003	1.783	1.225	2.595

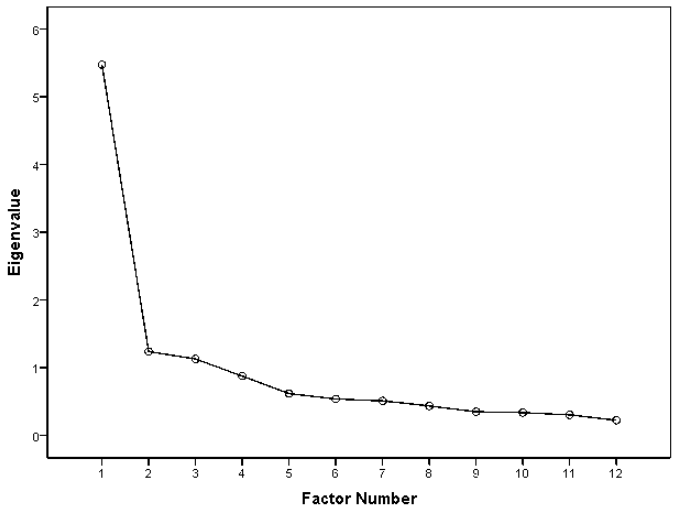
9.5 IRT Results of the PSI-Parent/Child Dysfunctional Interaction

Next, the Parent Stress Inventory-Parent/Child Dysfunctional Interaction (PSI-PDCI) scale was examined using IRT. The PSI-PDCI comprises 12 items measured on a 5-point scale: *Strongly Agree, Agree, Not Sure, Disagree,* and *Strongly Disagree.* As this scale contained only 12 items, it created a number of challenges.

9.6 Item Unidimensionality

As the PSI-PCDI scale comprised only 12 items, it was not possible to run DIMTEST to test for unidimensionality, because there were not enough items to create a subtest. Furthermore, the population of 240 non-defensive respondents that completed the PSI-PCDI was not large enough to be split in half to allow for both an exploratory and confirmatory factor analysis. Consequently, only an exploratory factor analysis was run, which is considered a weak method to use when checking for unidimensionality (DeMars, 2010)**.** The order of the first 5 eigenvalues was 5.473, 1.235, 1.127, .873, and .615. The big drop between the first and second eigenvalues followed by a leveling off of the remaining eigenvalues, suggest there was one dominant dimension (Scree plot shown in Figure 9.7).

Figue 9.7
Scree Plot of the Eigenvalues for the PSI-PCDI



9.7 IRT Analysis of the PSI-PCDI

Overall, the assessment had an alpha of 0.876, indicating that the assessment has high internal consistency (Table 9.14). Table 9.15 shows the overall model fit, indicating that the model fit the data well, although the p-value was approaching significance (0.091). The test information function indicates that maximum information was provided at a Θ of 1.450 (Figure 9.8) and that more information is present in the upper end of the scale than in the lower end of the scale. The inverse of the test information function, the conditional standard error of measurement, indicated that the minimum standard error was 0.227 at a Θ of 1.450, with significantly more error present at the lower extremes of the scale (Figure 9.9).

Table 9.14
Summary Statistics for the PSI-PCDI

Test	Items	Alpha	Mean	SD
Full Test	12	0.876	23.321	8.670

Table 9.15
Overall Model Fit for the PSI-PCDI

Test	Items	Chi-square	df	P	-2LL
Full Test	12	957.146	900	0.091	5015

Figure 9.8
Test Information Function for the PSI-PCDI

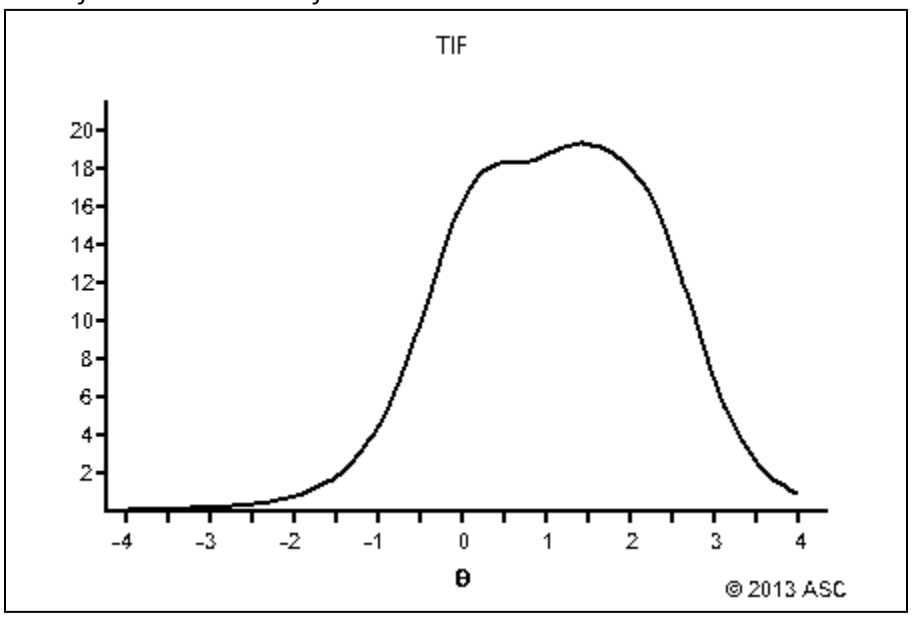
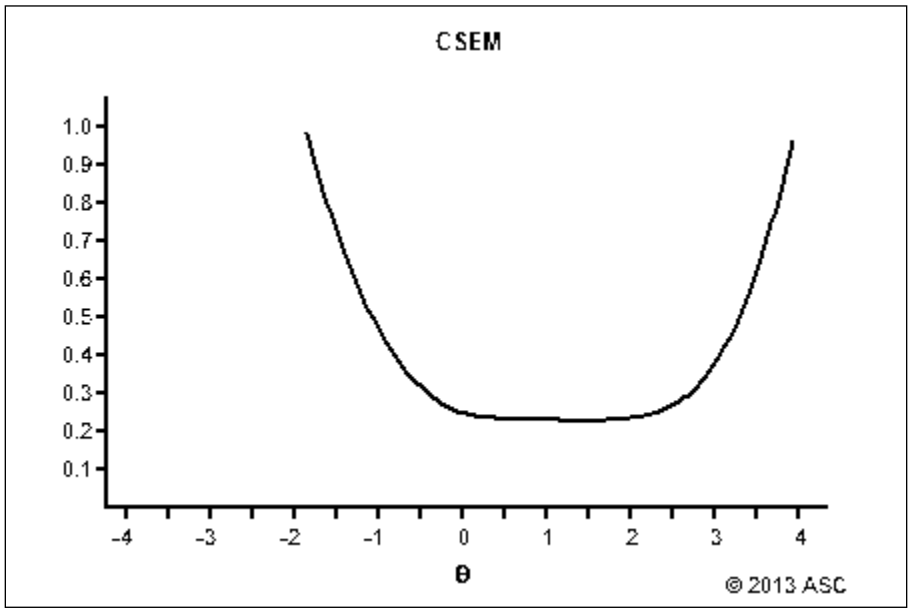


Figure 9.9
Conditional Standard Error of Measurement (PSI-PCDI)



As might be anticipated for an assessment with only 12 polytomous items, a number of concerning items demonstrated poor the model fit (Table 9.16) (Demars, 2010). The $r_{i\Theta}$ value in Table 9.16 is the Pearson product-moment correlation of the item response with the Θ estimate computed using all scored items. Most items had moderate $r_{i\Theta}$ values, with the exception of item 22, which had a low R-value (0.274). The a-parameter, as in the dichotomous model, reflects the second order derivative of the item response function at its point of inflection. However, in polytomous models, the item discrimination for examinees with different Θs depends on both the a-parameter and the relative locations of the b-parameters for the thresholds between polytomous item responses (DeMars, 2010). The items' a-parameters varied from 0.387 (item 22) to 1.728 (item 15). Items with a-parameters below 0.4 may indicate an item with low discrimination value (Demars, 2010), and might be considered for removal. The standard errors for the a-parameters are also shown in Table 9.16. There are four values for the b-parameters for each item. This is because the PSI-PCDI has five options for each item and when analyzed using a graded response model, the b-parameter indicates the boundary location (or threshold) between the various options. In other words, the boundary parameter indicates the

threshold at which a respondent has a 50% probability of choosing a lower option or a response greater than the lower option. For example, for item 13, a respondent with a Θ of 0.18 would have a 50% probability of choosing either *Strongly Disagree* or one of the other options (i.e., *Disagree, Not Sure, Agree,* or *Strongly Agree)*. Likewise, an individual with a Θ of 1.39 would have a 50% probability of choosing *Disagree* or one of the other options (i.e., *Not Sure, Agree,* or *Strongly Agree).* The standard errors for each of the b-parameters are shown in Appendix F along with the items, the complete test statistics, the item information curve, and the item statistics for each of the items. Three items (18, 22, and 24) were flagged for poor model fit (Table 9.16) based on a significant value of the chi-square statistic (p-values of 0.000, 0.018, and 0.008 respectively), indicating these items might not fit the SGRM model.

Table 9.16
Item Parameters for PSI-PCDI

Item Number	Item Mean	$r_{i\Theta}$	a	(SE)	b_1, b_2, b_3, b_4	Flag(s)
PSI Item 13	1.650	0.502	1.527	(0.217)	0.18, 1.39, 1.56, 2.43	
PSI Item 14	1.975	0.541	1.595	(0.214)	0.03, 0.78, 1.04, 1.96	
PSI Item 15	1.700	0.508	1.728	(0.248)	0.26, 1.18, 1.32, 2.11	
PSI Item 16	2.154	0.580	1.612	(0.212)	-0.23, 0.71, 0.91, 1.61	
PSI Item 17	1.442	0.470	1.572	(0.245)	0.51, 1.69, 1.93, 2.53	
PSI Item 18	2.142	0.441	1.003	(0.132)	-0.11, 0.77, 1.11, 2.01	F
PSI Item 19	1.575	0.527	1.700	(0.246)	0.30, 1.35, 1.70, 2.35	
PSI Item 20	1.846	0.541	1.532	(0.207)	-0.03, 1.06, 1.30, 2.37	
PSI Item 21	2.258	0.500	1.017	(0.126)	-0.48, 0.69, 1.06, 2.10	
PSI Item 22	2.421	0.274	0.387	(0.044)	-2.05, -0.12, 3.29, 6.40	F
PSI Item 23	1.704	0.467	1.284	(0.177)	0.29, 1.18, 1.57, 2.34	
PSI Item 24	2.454	0.518	1.062	(0.135)	-0.44, 0.44, 0.66, 1.72	F

When examining the category response function curves for the items, item 22 appeared to have the poorest fit (see Figure 9.10). The category response function for item 22 can be compared to one of the better performing items, item 20 (Figure 9.11). As can be seen in Figure 9.10, the thresholds between categories did a poor job of distinguishing between different Θs. Additionally, this item had significantly more responses in category three than did any of the other items. For all of the items on

the PSI-PCDI scale, except for item 22, category three is *Not Sure*. The categories for item 22 are presented in a different format, asking parents to indicate how good of a parent they believe they are, with category three stating that they feel they are *An Average Parent.* Figure 9.11 shows an item functioning in a manner that is more typical of a polytomous item (item 20), and for the most part, provides information between the various response categories. However, the third response category for this item (*Not Sure)* indicates that it may not be providing much information, and this pattern for the *Not Sure* response was consistent with the many of the items on the PSI-PCDI scale indicating that the *Not Sure* option may not be providing helpful information (see Appendix F).

Figure 9.10
Category Response Function for Item 22 on the PSI-PCDI

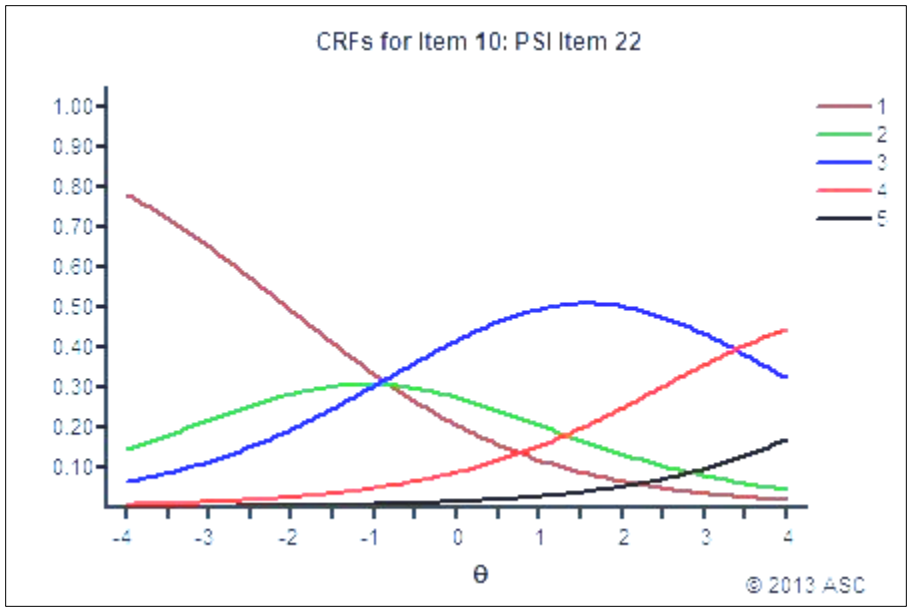
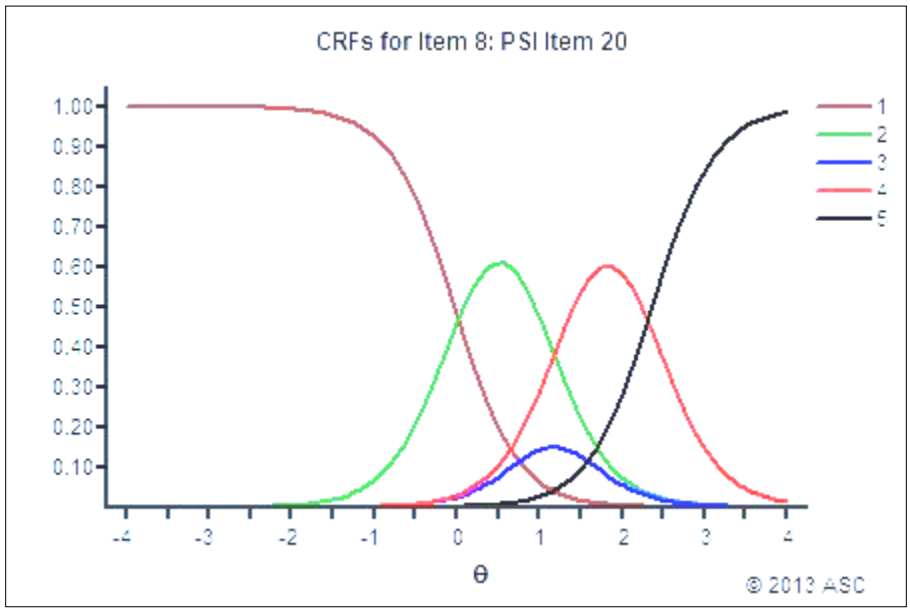


Figure 9.11

Category Response Function for Item 20 on the PSI-PCDI



9.8 Scale Reduction

As with the CTS-2, an attempt was made to see how removing some of the items would influence the functioning of the PSI-PCDI scale in predicting future child maltreatment. In addition to being flagged during IRT analysis as items that did not adequately fit the SGRM model (Table 9.16), items 18, 22, and 24 also appeared to be the poorest performing items when their category response function curves were examined. Item 22 (Figure 9.10) clearly did not distinguish well between all response categories; only categories 1, 3, and 4 appear to function. Although the concerns with items 18 and 24 were more subtle, the maximum information for *Strongly Disagree* and *Disagree*, and for A*gree* and *Strongly Agree* options overlapped considerably, indicating that these items may have performed better if scored dichotomously (i.e., *agree* or *disagree)*. Consequently, these items were removed from the scale and IRT analysis was again run.

IRT analysis was run with the remaining nine items in the same manner as outlined in Section 9.5. The alpha of the reduced PSI-PCDI was 0.886 (Table 9.17), slightly higher compared to the alpha for

the scale when all 12 items were included (0.876). After these three items were removed, the new model fit statistic became significant (p-value < 0.00), indicating that the model did not fit the data well (Table 9.18). This may be the result of having only 9 polytomously scored items. Although the results of the analysis are presented, this limitation should be kept in mind.

Table 9.17
Summary Statistics for PSI-PCDI Reduced Scale

Test	Items	Alpha	Mean	SD
Full Test	9	0.886	16.304	6.909

Table 9.18
Overall Model Fit for PSI-PCDI Reduced Scale

Test	Items	Chi-square	df	p	-2LL
Full Test	9	648.225	495	0.000	3221

As described in the methods section (8.2), since 3 items had been removed from the scale it was not possible to use the original PSI-PCDI cut score of 26. Rather, a threshold of 66% was set so the same number of caregivers would be 'indicated' on the new scale as were 'indicated' using the original PSI-PCDI scoring procedure. Those with a *theta* score below 66% were coded as 'not indicated' and those with a *theta* score above 66% were coded as 'indicated'. This dichotomized IRT PSI-PCDI variable was then entered into the Cox proportional hazard model. As can be seen in Table 9.19, running the Cox proportional hazard model with the reduced scale resulted in a similar outcome (Exp B = 1.858) to the original scale (Exp B = 1.849) in regards to predictive validity. It is noteworthy that, by using the item fit indices provided by the IRT analysis, 25% of the items were able to be removed without any loss of predictive validity. The correlation between the reduced IRT model and the original PSI-PCDI scale was 0.739 when the cut scores were examined, indicating a strong correlation as well as differences between two scoring procedures in classifying caregivers as 'indicated' versus 'not indicated'.

Table 9.19

Survival Analysis for PSI-PCDI Scores based on IRT Analysis with and without Poorly Functioning Items

Measure	Number	Number indicated	Significance	Hazard Ratio - Exp(B)	95% Confidence Interval for Exp (B)	
					Lower	Upper
PSI-PCDI using Established Assessment Cut score	177	62*	.001	1.849	1.287	2.656
PSI-PCDI using ITR *Theta* ranking with all Items Included	177	61	.002	1.793	1.246	2.580
PSI-PCDI using ITR *Theta* ranking Minus Items 18, 22 & 24	177	62*	.001	1.858	1.293	2.671

*52 of the same caregivers were categorized as indicated with both scoring procedures.

Chapter 10

Discussion for Study 2

Assessment tools can have various functions in CWS, including (a) screening and general disposition; (b) definitions, which may include diagnosis, labeling, or quantification of problem severity; (c) planning or matching treatment; (d) monitoring treatment progress; and (e) evaluating treatment outcome (Halverson, 1995). Due to the complexity of decisions made by CWS caseworkers, there is a need for assessments with support for their reliability and validity that can assist decision-making in all aspects of the assessment process. Study 2 was carried out to illustrate how the use of Item Response Theory (IRT) could provide information that might aid in the development of assessment tools with greater support for their reliability and validity in addressing these various needs.

The results of the IRT analysis indicated that for the scales examined, the administered items were representative of different levels of the trait, allowing for the classification of those higher and lower on the underlying trait. Additionally, many of the items on the scales appeared to adequately distinguish between those who had trait levels below and above the given items difficulty level (i.e., adequate items discrimination parameters). Further, the reliability of the estimates of the trait levels seemed reasonable given the relatively small sample size. Although these various measures are important, similar measures are also available using classical test theory (CTT). However, the value of IRT analysis goes beyond these item and test level statistics as the considerable detail obtained by IRT analysis facilitate a diversity of test development applications that may not be available using CTT.

One application of IRT analysis, as demonstrated in Study 2, involved an examination of each item's functioning, which facilitated scale reduction resulting in shorter assessments that functioned similarly as the full-scale assessments in predicting future child maltreatment. In the case of the CTS-2, 45% of the items were removed without reducing the tools predictive validity. In the case of the PSI-

PCDI, 25% of the items were removed without loss of predictive validity, which is a significant percentage considering that the PSI-PCDI only has 12 items. The value of creating shorter versions of the scales is particularly relevant when multiple dimensions are being measured, as fewer items per dimension may create an opportunity for the inclusion of standardized measures that otherwise would be too burdensome in the day to day practice of CWS. However, even though the utility of assessment administration may be increased, the shorter assessments are not helpful if caseworkers do not have the time to score or the expertise to interpret the assessment results. This is another area were the use of IRT analysis may be of assistance.

IRT provides information that is readily applicable for Computer Adaptive Testing (CAT). As demonstrated in Study 2, IRT analysis provides information on each item, indicating its level on the underlying trait. If a computer-administered assessment is programmed to start with items at a particular trait level, the individuals' responses can be used to select future items. This not only allows for further reduction of the number of items needed to establish an individual's trait level, it also significantly reduces or even eliminates the burden of scoring the assessment, as this could be done by the computer program. Furthermore, the CAT program could provide profile level information to the caseworker, not only indicating the areas with which the family appears to be struggling, but also providing information on the overall risk associated with particular profiles. Additionally, based on research and service availability, the computer could be programmed to suggest treatment options that have been shown to be effective for individuals/families with particular profiles. This is not to suggest that the CAT program could replace caseworkers' expertise; in fact, the caseworker assessment of the family (e.g., the SDM) could be included in the CAT algorithm. Rather, CAT could become a significant resource that caseworkers could utilize in both their assessment and case planning. Further, since assessments based on IRT analysis provide an adequate account of change scores, the resulting

assessment system could provide information for all areas of assessment, from screening and disposition through evaluation of treatment outcome.

Clearly, assessment in an area as complex as CWS should be multi-method and multi-source, including, not only parental report, but also caseworker assessment, family history, and information from others familiar with the situation. However, the addition of an assessment component that includes parent report and provides evidence supporting its reliability and validity would contribute significantly to the overall value of the assessment system. As shown in Study 1, most caregivers responded to the assessment items in a manner that could be used to assist in determining the areas with which the family is struggling, consequently providing information that is not only helpful with regards to risk assessment, but also giving indication of the needs that should be targeted to improve the conditions in the family home. However, the results also indicated that a significant percentage of caregivers reported in a defensive manner on the various assessments, resulting in an underreporting of the level of challenge they are facing. As outlined in the literature review (Chapter 7), IRT may also provide a mechanism for increasing the accuracy and efficiency of the assessment of both the non-defensive and the defensive caregivers involved with CWS.

IRT analysis provides item level information, indicating how much of a trait someone is likely to have if he/she endorses a given item. By examining the trait levels measured by the items on the various assessments, it may be possible to create an assessment that would include pairs of statements matched on the underlying trait levels from different domains. The caregivers would then be "forced" to select the statement that is most representative of their current situation. The use of ideal point IRT analysis could then provide information of their trait levels in the various domains based both on the statements they selected as well as the statements they did not select. Although not explored in the

results section due to limitations with the data, an example of statements being matched on a forced-choice assessment might be:

Item 1 from the CTS-2 Perpetrator scale states:

"*At some time, I have thrown something at my partner that could have hurt*".

This item has a difficulty level just above 0, at 0.248. This item could be paired with an item from the PSI-PCDI that also seems to be measuring a similar level of an underlying trait, instructing participants to select the item most like oneself. For example, item 16:

"*When I do things for my child, I get the feeling that my efforts are not appreciated very much*".

A caregiver's responses to a series of forced-choice paired statements may not only facilitate the creation of an individual profile on the measured domains (e.g., highlighting the areas with which the caregivers believe they are struggling with the most), but may also provide normative data on the caregivers' functioning in various domains as a result of the IRT scoring. Furthermore, scales built in this manner may be resistant to defensive responding; consequently, maximizing the value of the information obtained for a greater percentage of CWS involved families.

In order for a tool like this to be developed, there must be enough statements measuring each assessed domain to allow for both full coverage of the trait levels in the various domains, as well as for multiple pairings of statements from the different domains. Additionally, having multiple statements measuring similar trait levels within each domain reduces the need to reuse the same statements in different forced-choice pairs, which is important as test-takers might become confused or suspicious seeing the same statement multiple times. As demonstrated during the scale reduction section of Study 2, items could be removed if they were performing poorly (e.g. low discrimination), or if they were redundant. When creating a forced-choice assessment, these redundant items can be included in the

pool of available statements, resulting in an increased number of available statements. Again, it is through the use of CAT that the length of a forced-choice multidimensional assessment can be kept in check. As Chernyshenko et al. (2009) point out, "to create a test for a particular examinee, while keeping test length as short as possible, one could use computer adaptive item selection. Specifically, items could be formed dynamically by pairing statements that provide high information at an examinee's estimated trait score at any particular point during the test" (p. 112). Chernyshenko et al. (2009) found that by using CAT in combination with the forced-choice format that 10 dimensions could be adequately measured with just 50 forced-choice paired statements.

This study has a number of limitations that precluded the ability to create a full draft of a forced-choice assessment, and many of these limitations also apply to the findings of the IRT analysis. First, although the findings from IRT analysis are typically invariant across populations and assessments, the data collected by the CAP were obtained from a very restricted sample, that is, CWS involved caregivers at high risk of future child maltreatment who still have their children in their care. This limits the generalizability of the findings and indicates that the study should be replicated with other CWS involved populations to strengthen the evidence in support of the various difficulty and discrimination parameters of the items. Similarly, the number of participants limited the ability to ensure unidimensionality with the PSI-PCDI. Additionally, after removing certain items, the PSI-PCDI and the CTS-2 perpetrator scale both suggested that there may be issues with overall model fit. The issues with model fit for personality measures have been reported by others (Stark et al., 2006). Based on concerns with model fit, these authors recommend ideal point models rather than dominance models to fit the data. However, ideal point models, such as the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 1998) requires 750 test-takers while nonparametric maximum likelihood formula models require samples of over 2000 (Stark et al., 2006). Hence, it was not possible to use either of these models in this study due to sample size limitations. The findings with the reduced scales based on

IRT analysis will also need to be tested with a new population, as the results of this study could have been obtained by chance.

However, even given these limitations, this study provides direction on how IRT may be used to create a multi-dimensional assessment for CWS involved caregivers that are resistant to defensive responding. Furthermore, this study is just scratching the surface on the potential for tools built on IRT principals to influence assessments in a field such as child welfare. For example, assessments based on IRT function well with items from different formats (Embertson & Reise, 2000), creating the possibility that the assessment could comprise multiple formats with some multi-dimensional forced-choice items combined with forced-choice items from a single domain as well as polytomous items. Furthermore, with the assistance of CAT, the assessment format could depend on responses to items, with those reporting more defensively receiving a greater ratio of multi-dimensional forced-choice items. A defensive reporting flag could also signify to caseworkers that they need to be more diligent in collecting information from multiple sources for particular clients. Additionally, by documenting which services family members were engaged in data could be collected to determine the efficacy of the various interventions for a given profile. Thus, the use of IRT with CAT would provide a foundation for an assessment system that not only provides insight into individuals current functioning in various domains, but could also potentially provide information that would inform practitioners and policy makers of the effectiveness of different intervention.

Chapter 11

Conclusion

The purpose of this dissertation was to assist in the development of an assessment system that supports caseworkers in Child Welfare Service (CWS) in making informed decision regarding the families they are serving. Caseworkers in CWS make decisions that have significant consequences for the most vulnerable children in the United States. To adequately make these decisions caseworkers need access to assessment tools to assist in the areas of, (a) screening and general disposition; (b) definitions, which may include diagnosis, labeling, or quantification of problem severity; (c) planning or matching treatment; (d) monitoring treatment progress; and (e) evaluating treatment outcome (Halverson, 1995). Additionally, to insure accuracy and consistency the assessment tools must reliably measure client functioning and risk of child maltreatment. The assessment tools must also have support for all aspects of validity including, content validity, concurrent validity, utility, and consequential validity (Chapter 2). As outlined in Chapter 3, and indicated by the results of Study 1, current assessments utilized by CWS fail to meet many of these requirements. However, the results of the two studies undertaken in this dissertation provide insight into how the assessment process in CWS may be improved.

Among other things, Study 1 demonstrated that the current Structured Decision Making tool used by caseworkers provided some predictive validity, but that caseworkers had difficulty determining family level risk factors, such as the presence of substance abuse or domestic violence. Furthermore, Study 1 indicated that for the caregivers included in this study, using standardized self-report assessments added content validity, concurrent validity, and predictive validity, as well as potentially increased the utility of the assessment process. However, the results also indicated that a significant portion of caregivers involved with CWS did not complete the standardized assessments in a reliable manner. Additionally, CWS caseworkers struggled to assess accurately those caregivers who tended to respond in a more defensive manner.

Study 2 built on the finding of Study 1. First, Study 2 outlined and illustrated potential advantages of using IRT to evaluate the functioning of assessment tools. Study 2 then suggested ways in which the incorporation of ideal point IRT procedures combined with computer adaptive testing could facilitate the development of a forced-choice multidimensional assessment. A theoretical model was then provided which indicated that assessments built using the outlined procedures could offer insight into client functioning that is not only more reliable, but also has greater utility and can better meet the various criteria needed to support the validity of assessment tools. Although the process for developing a forced-choice multidimensional tool is considerable, given the scope of the challenges facing Child Welfare Services involved caregivers and the potential benefits of more adequate assessments, there is clear justification for the additional effort.

References

Adoption and Safe Families Act. (1997). *http://www.acf.hhs.gov/programs/cb/laws_policies/cblaws/public_law/pl105_89/pl105_89.htm*.

Ali, R., Meena, S., Eastwood, B., Richards, I., & Marsden, J. (2013). Ultra-rapid screening for substance-use disorders: The Alcohol, Smoking and Substance Involvement Screening Test (ASSIST-Lite). [Article]. *Drug and Alcohol Dependence, 132*(1-2), 352-361. doi: 10.1016/j.drugalcdep.2013.03.001

Anda, R. F., Felitti, V. J., Bremner, J. D., Walker, J. D., Whitfield, C., Perry, B. D., et al. (2006). The enduring effects of abuse and related adverse experiences in childhood. [Article]. *European Archives of Psychiatry & Clinical Neuroscience, 256*(3), 174-186. doi: 10.1007/s00406-005-0624-4

Antle, B. F., Barbee, A. P., Christensen, D. N., & Sullivan, D. J. (2009). The prevention of child maltreatment recidivism through the Solution-Based Casework model of child welfare practice. [Article]. *Children & Youth Services Review, 31*(12), 1346-1351. doi: 10.1016/j.childyouth.2009.06.008

Antle, B. F., Christensen, D. N., van Zyl, M. A., & Barbee, A. P. (2012). The impact of the Solution Based Casework (SBC) practice model on federal outcomes in public child welfare. [Article]. *Child Abuse & Neglect, 36*(4), 342-353. doi: 10.1016/j.chiabu.2011.10.009

Bada, H. S., Langer, J., Twomey, J., Bursi, C., Lagasse, L., Bauer, C. R., et al. (2008). Importance of Stability of Early Living Arrangements on Behavior Outcomes of Children With and Without Prenatal Drug Exposure. *Journal of Developmental & Behavioral Pediatrics, 29*(3), 173-182 110.1097/DBP.1090b1013e3181644a3181679.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*(3), 263-272. doi: 10.1111/j.1468-2389.2007.00386.x

Baird, C., & Wagner, D. (2000). The relative validity of actuarial and consensus based risk assessment systems. *Children and Youth Services Review, 22*(11-12), 839-871. doi: 10.1016/s0190-7409(00)00122-5

Baird, C., Wagner, D., Healy, T., & Johnson, K. (1999). Risk Assessment in Child Protective Services: Consensus and Actuarial Model Reliability. [Article]. *Child Welfare, 78*(6), 723-748.

Bank, L., & Burraston, B. (2001). Abusive home environments as predictors of poor adjustment during adolescence and early adulthood. *Journal of Community Psychology, 29*(3), 195-217.

Barth, R. P. (2009). Preventing Child Abuse and Neglect with Parent Training: Evidence and Opportunities. [Article]. *Future of Children, 19*(2), 95-118.

Barth, R. P., Wildfire, J., & Green, R. L. (2006). Placement Into Foster Care and the Interplay of Urbanicity, Child Behavior Problems, and Poverty. [Article]. *American Journal of Orthopsychiatry, 76*(3), 358-366. doi: 10.1037/0002-9432.76.3.358

Bellamy, J. L. (2008). Behavioral problems following reunification of children in long-term foster care. [Article]. *Children & Youth Services Review, 30*(2), 216-228. doi: 10.1016/j.childyouth.2007.09.008

Ben-Arieh, A., & Frønes, I. (2011). Taxonomy for child well-being indicators: A framework for the analysis of the well-being of children. [Article]. *Childhood, 18*(4), 460-476. doi: 10.1177/0907568211398159

Bortolotti, S., Tezza, R., Andrade, D., Bornia, A., & Sousa Júnior, A. (2013). Relevance and advantages of using the item response theory. [Article]. *Quality & Quantity, 47*(4), 2341-2360. doi: 10.1007/s11135-012-9684-5

Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory. [Article]. *Applied Measurement in Education, 24*(1), 1-21. doi: 10.1080/08957347.2011.532417

Brown, A., & Maydeu-Olivares, A. (2013). How IRT Can Solve Problems of Ipsative Data in Forced-Choice Questionnaires. *Psychological Methods, 18*(1), 36-52. doi: 10.1037/a0030641

Budd, K. S. (2001). Assessing Parenting Competence in Child Protection Cases: A Clinical Practice Model. [Article]. *Clinical Child & Family Psychology Review, 4*(1), 1-18.

Burns, B. J., Phillips, S. D., Wagner, H. R., Barth, R. P., Kolko, D. J., Campbell, Y., et al. (2004). Mental Health Need and Access to Mental Health Services by Youths Involved With Child Welfare: A National Survey. [Article]. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(8), 960-970. doi: 10.1097/01.chi.0000127590.95585.65

California Evidence Based Clearing House: For Child Welfare. http://www.cebc4cw.org/

Carter, N. T., Lake, C. J., & Zickar, M. J. (2010). Toward Understanding the Psychology of Unfolding. *Industrial and Organizational Psychology-Perspectives on Science and Practice, 3*(4), 511-514. doi: 10.1111/j.1754-9434.2010.01283.x

Cash, S. J. (2001). Risk assessment in child welfare: The art and science. *Children and Youth Services Review, 23*(11), 811-830. doi: 10.1016/s0190-7409(01)00162-1

Chaffin, M., Bard, D., Hecht, D., & Silovsky, J. (2011). Change Trajectories During Home-Based Services With Chronic Child Welfare Cases. [Article]. *Child Maltreatment, 16*(2), 114-125. doi: 10.1177/1077559511402048

Chaffin, M., Silovsky, J. F., Funderburk, B., Valle, L. A., Brestan, E. V., Balachova, T., et al. (2004). Parent-Child Interaction Therapy With Physically Abusive Parents: Efficacy for Reducing Future Abuse Reports. [Article]. *Journal of Consulting & Clinical Psychology, 72*(3), 500-510. doi: 10.1037/0022-006x.72.3.500

Chaffin, M., Valle, L. A., Funderburk, B., Gurwitch, R., Silovsky, J., Bard, D., et al. (2009). A Motivational Intervention Can Improve Retention in PCIT for Low-Motivation Child Welfare Clients. [Article]. *Child Maltreatment, 14*(4), 356-368.

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative Scoring of Multidimensional Pairwise Preference Personality Scales Using IRT: Empirical Comparisons With Other Formats. [Article]. *Human Performance, 22*(2), 105-127. doi: 10.1080/08959280902743303

Child Abuse and Neglect Data System. U.S. Department of Health and Human Services: Administartion for Children and Families. from www.acf.hhs.gov/programs/cb/stats_research/index.htm#can

Children's Administartion Profile-Selected Demographic and Proformance Measures Spanning 2004-2009. (2010). *http://ca.dshs.wa.gov/intranet/ppt/leg/EL_%20CS_LegPresentation.ppt*.

Child Maltreatment 2010. (2011). U.S. Department of Health and Human Services, Administration for Children and Families, Administration on Children, Youth and Families, Children's Bureau.

Child Welfare Education and Research Programs. *Policies to Implement Developmental screening in Pennsylvania Child Welfare Services: Reports from Agency Perspectives.* (2011) (pp. 59): University of Pittsburgh, School of Social Work, Pittsburgh, Pennsylvania.

Christiansen, D. N., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment. [Article]. *Human Performance, 18*(3), 267-307. doi: 10.1207/s15327043hup1803_4

Christensen, D. N., Todahl, J., & Barrett, W. C. (1999). *Solution-based casework : an introduction to clinical and case management skills in casework practice*. New York: Aldine de Gruyter.

Cicchetti, D. (2010). Resilience under conditions of extreme stress: a multilevel perspective. *World Psychiatry, 9*(3), 145-154.

Children's Administration Practices and procedures Guide, 2013. http://www.dshs.wa.gov/ca/pubs/mnl_pnpg/chapter1.asp

Chuang, E., Wells, R., Bellettiere, J., & Cross, T. P. (2013). Identifying the substance abuse treatment needs of caregivers involved with child welfare. [Article]. *Journal of Substance Abuse Treatment, 45*(1), 118-125. doi: 10.1016/j.jsat.2013.01.007

Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010). Statement Desirability Ratings in Forced-Choice Personality Measure Development: Implications for Reducing Score Inflation and Providing Trait-Level Information. [Article]. *Human Performance, 23*(4), 323-342. doi: 10.1080/08959285.2010.501047

Costa, N., Weems, C., Pellerin, K., & Dalton, R. (2006). Parenting Stress and Childhood Psychopathology: An Examination of Specificity to Internalizing and Externalizing Symptoms. [Article]. *Journal of Psychopathology & Behavioral Assessment, 28*(2), 113-122. doi: 10.1007/s10862-006-7489-3

Courtney, M. E. (2010). Families in child welfare system struggle to meet basic needs. *Partners for Our Children, http://www.partnersforourchildren.org/*, 1-2.

Cronbach, L., and Meehl, P. (1955). Construct Validity in Psychology Tests. *Psychological Bulletin, 52*, 281-302.

Dakil, S. R., Cox, M., Lin, H., & Flores, G. (2012). Physical abuse in U.S. Children: Risk factors and deficiencies in referrals to support services. *Journal of Aggression, Maltreatment & Trauma, 21*(5), 555-569. doi: 10.1080/10926771.2012.680007

Dalal, D. K., Withrow, S., Gibby, R. E., & Zickar, M. J. (2010). Six Questions That Practitioners (Might) Have About Ideal Point Response Process Items. *Industrial and Organizational Psychology-Perspectives on Science and Practice, 3*(4), 498-501. doi: 10.1111/j.1754-9434.2010.01279.x

DeMars, C. (2010). *Item response theory*. Oxford; New York: Oxford University Press.

DePanfilis, D., & Zuravin, S. J. (2001). Assessing risk to determine the need for services. *Children and Youth Services Review, 23*(1), 3-20. doi: 10.1016/s0190-7409(00)00125-0

Dishion, T. J., Shaw, D., Connell, A., Gardner, F., Weaver, C., & Wilson, M. (2008). The Family Check-Up with high-risk indigent families: Preventing problem behavior by increasing parents' positive behavior support in early childhood. *Child Development, 79*(5), 1395-1414. doi: 10.1111/j.1467-8624.2008.01195.x

Dorsey, S., Kerns, S. E. U., Trupin, E. W., Conover, K. L., & Berliner, L. (2012). Child Welfare Caseworkers as Service Brokers for Youth in Foster Care: Findings From Project Focus. [Article]. *Child Maltreatment, 17*(1), 22-31. doi: 10.1177/1077559511429593

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*(4), 465-476. doi: 10.1111/j.1754-9434.2010.01273.x

Dunnette, M., McCartney, J., Carlson, H., & Kirchner, W. (1962). A Study of Faking Behavior on a Forced-Choice Self-Description Checklist. *Personnel Psychology, 15*(2).

Eyberg, S. M., Nelson, M. M., & Boggs, S. R. (2008). Evidence-Based Psychosocial Treatments for Children and Adolescents With Disruptive Behavior. [Article]. *Journal of Clinical Child & Adolescent Psychology, 37*(1), 215-237. doi: 10.1080/15374410701820117

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

English, D. J., Graham, C., Litrownik, A. J., Everson, M., & Bangdiwala, S. I. (2005). Defining maltreatment chronicity: Are there differences. [Article]. *Child Abuse & Neglect, 29*(5), 575-595. doi: 10.1016/j.chiabu.2004.08.009

English, D., Marshall, D. B., Brummel, S. C., & Orme, M. . (1999). Characteristics of Repeated Referrals to Child Protective Services in Washington State. *Child Maltreatment, 4*, 297-307.

English, D. J., & Pecora, P. J. (1994). Risk Assessment as a Practice Method in Child Protective Services. [Article]. *Child Welfare, 73*(5), 451-473.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement (3rd ed.).* (pp. 105-146). New York, NY England: Macmillan Publishing Co, Inc. American Council on Education.

Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., et al. (1998). Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults - The adverse childhood experiences (ACE) study. *American Journal of Preventive Medicine, 14*(4), 245-258.

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core Implementation Components. *Research on Social Work Practice, 19*(5), 531-540. doi: 10.1177/1049731509335549

Fluke, J. D., Shusterman, G. R., Hollinshead, D. M., & Yuan, Y. Y. T. (2008). Longitudinal analysis of repeated child abuse reporting and victimization: Multistate analysis of associated factors. *Child Maltreatment, 13*(1), 76-88. doi: 10.1177/1077559507311517

Fluke, J. D., Yuan, Y.-Y. T., & Edwards, M. (1999). Recurrence of Maltreatment: An Application of the National Child Abuse and Neglect Data System (NCANDS). [Article]. *Child Abuse & Neglect, 23*(7), 633-650.

Gambrill, E., & Shlonsky, A. (2000). Risk assessment in context. *Children and Youth Services Review, 22*(11-12), 813-837. doi: 10.1016/s0190-7409(00)00123-7

Gardner, W., Lucas, A., Kolko, D. J., & Campo, J. V. (2007). Comparison of the PSC- 17 and Alternative Mental Health Screens in an At-Risk Primary Care Sample. [Article]. *Journal of the American Academy of Child & Adolescent Psychiatry, 46*(5), 611-618. doi: 10.1097/chi.0b013e318032384b

Gardner, W., Murphy, M., Childs, G., Kelleher, K., Pagano, M., Jellinek, M., et al. (1999). The PSC-17: a brief pediatric symptom checklist with psychosocial problem subscales. A report from PROS and ASPN. [Article]. *Ambulatory Child Health, 5*(3), 225. Gardner, W., Murphy, M., Childs, G., Kelleher, K., Pagano, M., Jellinek, M., et al. (1999). The PSC-17: a brief pediatric symptom

checklist with psychosocial problem subscales. A report from PROS and ASPN. [Article]. *Ambulatory Child Health, 5*(3), 225.

Gill, A. M., Hyde, L. W., Shaw, D. S., Dishion, T. J., & Wilson, M. N. (2008). The Family Check-Up in Early Childhood: A Case Study of Intervention Process and Change. [Article]. *Journal of Clinical Child & Adolescent Psychology, 37*(4), 893-904. doi: 10.1080/15374410802359858

Glascoe, F. P. (2000). Evidence-based approach to developmental and behavioural surveillance using parents' concerns. [Article]. *Child: Care, Health & Development, 26*(2), 137-149. doi: 10.1046/j.1365-2214.2000.00173.x

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293-323. doi: 10.1037/1076-8971.2.2.293

Guyer, R., & Thompson, N.A. (2013). *User's Manual for Xcalibre item response theory calibration software, version 4.2*. Woodbury MN: Assessment Systems Corporation

Halle, T., Forry, N., Hair, E., Perper, K., Wander, L., Wessel, J., & Vick, J. (2009). Disparities in Early Learning and Development: Lessons from the Early Childhood Longitudinal Study - Birth Cohort (ECLS-B). *Child Trends*, 31.

Halverson, C. (1995). Measurement Beyond the Individual. In J. C. Conoley (Ed.), *Family Assessment* (pp. 3-19). Lincoln-Nebraska: Buros Institute of Mental Measurements.

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

Hamby, S., Finkelhor, D., Turner, H., & Ormrod, R. (2010). The overlap of witnessing partner violence with child maltreatment and other victimizations in a nationally representative survey of youth. [Article]. *Child Abuse & Neglect, 34*(10), 734-741. doi: 10.1016/j.chiabu.2010.03.001

Haskett, M. E., Ahern, L. S., Ward, C. S., & Allaire, J. C. (2006). Factor Structure and Validity of the Parenting Stress Index-Short Form. [Article]. *Journal of Clinical Child & Adolescent Psychology, 35*(2), 302-312. doi: 10.1207/s15374424jccp3502_14

Horwitz, S. M., Hurlburt, M. S., Cohen, S. D., Zhang, J., & Landsverk, J. (2011). Predictors of placement for children who initially remained in their homes after an investigation for abuse or neglect. [Article]. *Child Abuse & Neglect, 35*(3), 188-198. doi: 10.1016/j.chiabu.2010.12.002

Humeniuk, R., Ali, R., Babor, T. F., Farrell, M., Formigoni, M. L., Jittiwutikarn, J., et al. (2008). Validation of the alcohol, smoking and substance involvement screening test (ASSIST). [Article]. *Addiction, 103*(6), 1039-1047. doi: 10.1111/j.1360-0443.2007.02114.x

Hussey, J. M., Chang, J. J., & Kotch, J. B. (2006). Child maltreatment in the United States: Prevalence, risk factors, and adolescent health consequences. *Pediatrics, 118*(3), 933-942. doi: 10.1542/peds.2005-2452

Huth-Bocks, A., & Hughes, H. (2008). Parenting Stress, Parenting Behavior, and Children's Adjustment in Families Experiencing Intimate Partner Violence. [Article]. *Journal of Family Violence, 23*(4), 243-251. doi: 10.1007/s10896-007-9148-1

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The Impact of Faking on Employment Tests: Does Forced Choice Offer a Solution? [Article]. *Human Performance, 13*(4), 371-388.

Johnson, W. (2004). *Effectiveness of California's Child Welfare Structured Decision Making Model: A Prospective Study of the Validity of the California Family Risk Assessment*. Retrieved from http://www.nccdglobal.org/publications

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. [Article]. *Journal of Educational Measurement, 50*(1), 1-73. doi: 10.1111/jedm.12000

Kelley, S. J. (1998). Stress and coping behaviors in substance abusing mothers. *Journal of the Society of Pediatric Nurses, 3*, 103-110.

Kerlinger, F. (1986). *Foundations of Behavioral Research* (3rd ed.). New York: Holt, Rinehart and Winston

Kernic, M. A., Wolf, M. E., Holt, V. L., McKnight, B., Huebner, C. E., & Rivara, F. P. (2003). Behavioral problems among children whose mothers are abused by an intimate partner. [Article]. *Child Abuse & Neglect, 27*(11), 1231. doi: 10.1016/j.chiabu.2002.12.001

Kerns, S. E. U., Dorsey, S., Trupin, E. W., & Berliner, L. (2010). Project Focus: Promoting Emotional Health and Well-Being for Youth in Foster Care Through Connections to Evidence-Based Practices. *Emotional & Behavioral Disorders in Youth*(Spring), 9.

Klein, S., & Harden, B. J. (2011). Building the evidence-base regarding infants/toddlers in the child welfare system, Editorial, *Children & Youth Services Review,* pp. 1333-1336. Retrieved from http://offcampus.lib.washington.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=60920243&site=ehost-live

Kleinbaum D. G., & Klein M. (2012). *Survival analysis a self-learning text* (3rd edition ed.). New York, NY: Springer.

Knitzer, J., Suzanne Theberge, and Kay Johnson. (2008). Reducing Maternal Depression and Its Impact on Young Children: Toward a Responsive Early Childhood Policy Framework *National Center for Children in Poverty*, 24.

Kohl, P. L., Barth, R. P., Hazen, A. L., & Landsverk, J. A. (2005). Child welfare as a gateway to domestic violence services. [Article]. *Children & Youth Services Review, 27*(11), 1203-1221. doi: 10.1016/j.childyouth.2005.04.005

Kohl, P. L., Kagotho, J. N., & Dixon, D. (2011). Parenting Practices among Depressed Mothers in the Child Welfare System. *Social Work Research, 35*(4), 215-225.

Kotch, J. B., Lewis, T., Hussey, J. M., English, D., Thompson, R., Litrownik, A. J., et al. (2008). Importance of early neglect for childhood aggression. *Pediatrics, 121*(4), 725-731. doi: 10.1542/peds.2006-3622

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. [Article]. *JGIM: Journal of General Internal Medicine, 16*(9), 606-613. doi: 10.1046/j.1525-1497.2001.016009606.x

Landsverk, J., Garland, A., Reutz, J. R., & Davis, I. (2011). Bridging science and practice in child welfare and children's mental health service systems through a two-decade research center trajectory. *Journal of Social Work, 11*(1), 80-98. doi: 10.1177/1468017310381816

Limbos, M., Joyce, D. (2011). Comparison of the ASQ and PEDS in Screening for Developmental Delay in Children Presenting for Primary Care. *Journal of Developmental & Behavioral Pediatrics, 7*, 499-511.

Loman, A. (2006). Families Frequently Encountered by Child Protection Services: A report on Chronic Child Abuse and Neglect. *Institute of Applied Research*, 60.

Loman, L. A., Siegel, G. L., Research, I. o. A., & Services, M. D. o. H. (2004). *An Evaluation of the Minnesota SDM Family Risk Assessment*: Institute of Applied Research.

Lunkenheimer, E. S., Shaw, D. S., Gardner, F., Dishion, T. J., Connell, A. M., & Wilson, M. N. (2008). Collateral Benefits of the Family Check-Up on Early Childhood School Readiness: Indirect Effects of Parents' Positive Behavior Support. [Article]. *Developmental Psychology, 44*(6), 1737-1752. doi: 10.1073/a0013858

Lyons, P., & Doueck, H. J. (1996). Risk assessment for child protective services: A review of the empirical literature on instrument performance. [Article]. *Social Work Research, 20*(3), 143-156.

Mansell, J., Ota, R., Erasmus, R., & Marks, K. (2011). Reframing child protection: A response to a constant crisis of confidence in child protection. [Article]. *Children and Youth Services Review, 33*(11), 2076-2086. doi: 10.1016/j.childyouth.2011.04.019

Marcenko, M., Newby, M., Lee, J., Courtney, M., & Brennan, K. (2009). Evaluation of Washington's Solution Based Casework practice model. Interim Report Part IV: Baseline parent survey analysis

by state, region, and service context. *Partners for Our Children, http://www.partnersforourchildren.org/*, 39.

Marcenko, M. O., Lyons, S. J., & Courtney, M. (2011). Mothers' experiences, resources and needs: The context for reunification. *Children and Youth Service Review 33*(3), 431-438.

Marshall, D. B., & English, D. J. (1999). Survival Analysis of Risk Factors for Recidivism in Child Abuse and Neglect. [Article]. *Child Maltreatment, 4*(4), 287.

Martinez-Torteya, C., Anne Bogat, G., von Eye, A., & Levendosky, A. A. (2009). Resilience Among Children Exposed to Domestic Violence: The Role of Risk and Protective Factors. [Article]. *Child Development, 80*(2), 562-577. doi: 10.1111/j.1467-8624.2009.01279.x

McCoach, D. B., Rambo, K. E., & Welsh, M. (2013). Assessing the Growth of Gifted Students. [Article]. *Gifted Child Quarterly, 57*(1), 56-67. doi: 10.1177/0016986212463873

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A Silk Purse From the Sow's Ear: Retrieving Normative Information From Multidimensional Forced-Choice Items. *Organizational Research Methods, 8*(2), 222-248. doi: 10.1177/1094428105275374

McCrae, J. S., & Barth, R. P. (2008). Using Cumulative Risk to Screen for Mental Health Problems in Child Welfare. [Article]. *Research on Social Work Practice, 18*(2), 144-159. doi: 10.1177/1049731507305394

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.).* (pp. 13-103). New York, NY England: Macmillan Publishing Co, Inc. American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037/0003-066x.50.9.741

Miller, M., Linn, R., Gronlund, N. (2009). *Measurement and assessment in teaching* (10th ed. ed.): Upper Saddle River, N.J. : Merrill/Pearson

Munro, E. (2004). A simpler way to understand the results of risk assessment instruments. [Article]. *Children and Youth Services Review, 26*(9), 873-883. doi: 10.1016/j.childyouth.2004.02.026

Mustillo, S. A., Dorsey, S., Conover, K., & Burns, B. J. (2011). Parental Depression and Child Outcomes: The Mediating Effects of Abuse and Neglect. [Article]. *Journal of Marriage & Family, 73*(1), 164-180. doi: 10.1111/j.1741-3737.2010.00796.x

National Council on Crime and Delinquency. http://www.nccdglobal.org/

National Resource Center for Child Protective Services (2011). http://nrccps.org/

Nicklas, E., & Mackenzie, M. (2013). Intimate Partner Violence and Risk for Child Neglect during Early Childhood in a Community Sample of Fragile Families. [Article]. *Journal of Family Violence, 28*(1), 17-29. doi: 10.1007/s10896-012-9491-8

Nisbett, R., & Ross, L. . (1980). Generalizing from instances to population: Representativeness versus sampling theory *Human inference: Strategies and shortcomings of social judgment* (pp. 77-101). Englewood Cliffs, NJ: Prentice-Hill, Inc.

Nock, M. K., & Kazdin, A. E. (2005). Randomized Controlled Trial of a Brief Intervention for Increasing Participation in Parent Management Training. [Article]. *Journal of Consulting & Clinical Psychology, 73*(5), 872-879. doi: 10.1037/0022-006x.73.5.872

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. [Article]. *Journal of Clinical Child & Adolescent Psychology, 34*(2), 301-311. doi: 10.1207/s15374424jccp3402_9

Oskamp, S. (1965). Overconfidence in case-study judgments. *The Journal of Consulting Psychology, 29*, 261-265.

Perry, B. D., Pollard, R. A., Blakley, T. L., Baker, W. L., & Vigilante, D. (1995). Childhood Trauma, the Neurobiology of Adaptation, and "Use-dependent" Development of the Brain: How "States" Become "Traits". [Article]. *Infant Mental Health Journal, 16*(4), 271-291.

Pithouse, A., Broadhurst, K., Hall, C., Peckover, S., Wastell, D., & White, S. (2012). Trust, risk and the (mis)management of contingency and discretion through new information technologies in children's services. [Article]. *Journal of Social Work, 12*(2), 158-178. doi: 10.1177/1468017310382151

Proctor, L. J., Aarons, G. A., Dubowitz, H., English, D. J., Lewis, T., Thompson, R., et al. (2012). Trajectories of Maltreatment Re-Reports From Ages 4 to 12: Evidence for Persistent Risk After Early Exposure. [Article]. *Child Maltreatment, 17*(3), 207-217. doi: 10.1177/1077559512448472

Reise, S. P. (2010). Thurstone Might Have Been Right About Attitudes, but Drasgow, Chernyshenko, and Stark Fail to Make the Case for Personality. *Industrial and Organizational Psychology-Perspectives on Science and Practice, 3*(4), 485-488. doi: 10.1111/j.1754-9434.2010.01276.x

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A General Item Response Theory Model for Unfolding Unidimensional Polytomous Responses. [Article]. *Applied Psychological Measurement, 24*(1), 3.

Rohland, B. M., Rohrer, J. E., & Tzou, H. (1998). Broker model of case management for persons with serious mental illness in rural areas. *Administration and Policy in Mental Health, 25*(5), 549-553. doi: 10.1023/a:1022345500739

Rycus, J. H., R. (2003). Issues in Risk Assessment in Child Protective Services (Policy White Paper) (pp. 55). Columbus, Ohio: North American Resource Center for Child Welfare.

Samuels, B. (2012). Information Memorandum: Promoting Social and Emotional Well-Being for Children and Youth Receiving Child Welfare Services. *U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES: Administration on Children, Youth and Families*, 21.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shepard, L. (1993). Evaluating Test Validity. *Review of Research in Education, 19*, (pp. 405-450).

Shlonsky, A., Saini, M., Wu, M. (2007). The recurrence of child maltreatment: Predictive validity of risk assessment instruments. (pp. 1-33): THe Campbell Collaboration Library of Systematic Reviews.

Shlonsky, A., & Wagner, D. (2005). The next step: Integrating actuarial risk assessment and clinical judgment into an evidence-based practice framework in CPS case management. [Article]. *Children & Youth Services Review, 27*(4), 409-427. doi: 10.1016/j.childyouth.2004.11.007

Solomon, D., & Åsberg, K. (2012). Effectiveness of child protective services interventions as indicated by rates of recidivism. [Article]. *Children & Youth Services Review, 34*(12), 2311-2318. doi: 10.1016/j.childyouth.2012.08.014

Stark, S., & Chernyshenko, O. S. (2011). Computerized Adaptive Testing with the Zinnes and Griggs Pairwise Preference Ideal Point Model. [Article]. *International Journal of Testing, 11*(3), 231-247. doi: 10.1080/15305058.2011.561459

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15*(3), 463-487. doi: 10.1177/1094428112444611

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? [Article]. *Journal of Applied Psychology, 91*(1), 25-39. doi: 10.1037/0021-9010.91.1.25

Stiffman, A. R., Pescosolido, B., & Cabassa, L. J. (2004). Building a Model to Understand Youth Service Access: The Gateway Provider Model. [Article]. *Mental Health Services Research, 6*(4), 189-198.

Stiffman, A. R., Striley, C., Horvath, V. E., Hadley-Ives, E., Polgar, M., Elze, D., et al. (2001). Organizational Context and Provider Perception as Determinants of Mental Health Service Use. [Article]. *Journal of Behavioral Health Services & Research, 28*(2), 188.

Stout , W. (2006). DIMTEST (Version 2.1) [Computer Software]. Champaign, IL: The William Stout Institute for Measurement.

Straus, M. A. (2004). Scoring the CTS2 and CTSPC (pp. 1-12). http://pubpages.enh.edu

Straus, M. A., & Douglas, Emily M. (2004). A Short Form of the Revised Conflict Tactics Scales, and Typologies for Severity and Mutuality. *Violence and Victims, 19*(5), 507-520.

Straus, M. A., Hamby, S. L., Boney-McCoy, S., & Sugarman, D. B. (1996). The Revised Conflict Tactics Scales (CTS2). [Article]. *Journal of Family Issues, 17*(3), 283-316.

Streiner, D. L. (2010). Measure for Measure: New Developments in Measurement and Item Response Theory. [Article]. *Mesure pour mesure : nouveaux développements de la mesure et de la théorie de la réponse à l'item., 55*(3), 180-186.

Taylor, C. (2013). *Validity and Validation*. Oxford.: Oxford University Press

Thompson, R., & Wiley, T.R. (2009). Predictors of Re-Referral to Child Protective Services: A Longitudinal Follow-Up of an Urban Cohort Maltreated as Infants. *Child Maltreatment, 14*(1), 89-99

Thurstone, L. L. (1929). A law of comparative judgment. *Psychological Review, 34,* 273-286

Titov, N., Dear, B. F., McMillan, D., Anderson, T., Zou, J., & Sunderland, M. (2011). Psychometric Comparison of the PHQ-9 and BDI-II for Measuring Response during Treatment of Depression. [Article]. *Cognitive Behaviour Therapy, 40*(2), 126-136. doi: 10.1080/16506073.2010.550059

Tversky, A., & Kahneman, D. . (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185*, 8.

Vance, J. E., Bowen, N. K., Fernandez, G., & Thompson, S. (2002). Risk and Protective Factors as Predictors of Outcome in Adolescents With Psychiatric Disorder and Aggression. [Article]. *Journal of the American Academy of Child & Adolescent Psychiatry, 41*(1), 36.

Walker, C., & Davies, J. (2010). A Critical Review of the Psychometric Evidence Base of the Child Abuse Potential Inventory. [Article]. *Journal of Family Violence, 25*(2), 215-227. doi: 10.1007/s10896-009-9285-9

Wilson, D., & Homer, W. (1995). Chronic Child Neglect: Needed Developments in Theory and Practice. *Families in Society, 86*(4), 11.

Young, N. K., Boles, S. M., & Otero, C. (2007). Parental Substance Use Disorders and Child Maltreatment: Overlap, Gaps, and Opportunities. [Article]. *Child Maltreatment, 12*(2), 137-149.

Appendix A

17 Safety Threats

1. The family/facility *situation* results in no adults in the home/facility performing parenting/child care duties and responsibilities that assure child's safety.
2. The family/facility *situation* is that the living/child care arrangement(s) seriously endanger the child's physical health.
3. Caregiver(s) are acting (*behaving*) violently or dangerously and the behaviors affect child safety.
4. There has been an incident of domestic violence that affects child safety.

   If "Yes", complete the questions below:
   a. The domestic violence perpetrator has caused serious harm or threats of harm against the adult victim/caregiver of the child.
   b. The domestic violence perpetrator has seriously harmed or threatened serious harm to the child.
   c. The level of violence and/or threats towards either the adult victim or child is increasing so that serious harm is likely to occur.
   d. There are other indications of increased dangers from the domestic violence perpetrator such as suicide threats or attempts, substance abuse or threats with weapons
5. Caregiver(s) will not or cannot control their *behavior* and their *behavior* affects child safety.
6. Caregiver(s) perceives child in *extremely* negative terms
7. Caregiver(s) do not have or do not use resources necessary to meet the child's immediate basic needs which present an immediate threat of serious harm to a child.
8. Caregivers' *attitudes*, *emotions* and *behavior* are such that they are threatening to severely harm a child or are fearful they will maltreat the child or request placement.
9. Caregiver(s) intend(ed) to seriously hurt the child.
10. Caregiver(s) lack the parenting knowledge, skills, or motivation necessary to assure a child's safety.
11. Caregiver(s) overtly rejects CA intervention, refuses access to a child, or there is some indication that the caregiver(s) will flee.
12. Caregiver(s) are not meeting, cannot meet or will not meet the child's exceptional physical, emotional, medical, or behavioral needs.
13. Caregiver(s) cannot or will not explain child's injuries or maltreating condition(s) or explanation is not consistent with the facts.
14. A child has serious physical injuries or serious physical conditions resulting from maltreatment.
15. A child demonstrates serious emotional symptoms, self-destructive behavior and/or lack of behavioral control that results in provoking dangerous reactions in caregivers.
16. A child is extremely fearful of the home/facility situation or people within the home/facility.
17. Child sexual abuse is suspected has occurred or circumstances suggest sexual abuse is likely to occur.

Appendix B – Structured Decision Making Tool

EXAMPLE ONLY – NOT FOR USE

SDM® FAMILY RISK ASSESSMENT OF ABUSE/NEGELCT r: 05-07

Referral Name: **Referral #:** **Referral Date:** / / **Office:**

Worker Name: **WorkerID:** **Assessment Date:** / /

NEGLECT	Score
N1. Current complaint is for neglect	
a. No	0
b. Yes	1
N2. Prior investigations (assign highest score that applies)	
a. None	0
b. One or more, abuse only	1
c. One or two for neglect	2
d. Three or more for neglect	3
Household has previously received CPS (voluntary/court-ordered)	
a. No	0
b. Yes	1
N4. Number of children involved in the CA/N incident	
a. One, two, or three	0
b. Four or more	1
N5. Age of youngest child in the home	
a. Two or older	0
b. Under two	1
Primary caregiver provides physical care inconsistent with child needs	
a. No	0
b. Yes	1
N7. Primary caregiver has a past or current mental health problem	
a. No	0
b. Yes	1
Primary caregiver has historic or current alcohol or drug problem (check applicable items and add for score)	
a. Not applicable	0
b. Alcohol (current or historic)	1
c. Drug (current or historic)	1
N9. Characteristics of children in household (check applicable items and add for score)	
a. Not applicable	0
b. Medically fragile/failure to thrive	1
c. Developmental or physical disability	1
d. Positive toxicology screen at birth	1
N10. Housing (check applicable items and add for score)	
a. Not applicable	0
b. Current housing is physically unsafe	1
c. Homeless at time of investigation	2
TOTAL NEGLECT RISK SCORE	

ABUSE	Score
A1. Current complaint is for abuse	
a. No	0
b. Yes	1
A2. Number of prior abuse investigations/assessments	
a. None	0
b. One	1
c. Two or more	2
(actual number:)	
A3. Household has previously received CPS (voluntary/court-ordered)	
a. No	0
b. Yes	1
A4. Prior injury to a child resulting from CA/N	
a. No	0
b. Yes	1
Primary caregiver's assessment of incident (check applicable items and add for score)	
a. Not applicable	0
b. Blames child	1
c. Justifies maltreatment of a child	2
A6. Domestic violence in the household in the past year	
a. No	0
b. Yes	2
Primary caregiver characteristics (check applicable items and add for score)	
a. Not applicable	0
b. Provides insufficient emotional/psychological support	1
c. Employs excessive/inappropriate discipline	1
d. Domineering caregiver(s)	1
A8. Primary caregiver has a history of abuse or neglect as a child	
a. No	0
b. Yes	1
Secondary caregiver has historic or current alcohol or drug problem	
a. No	0
b. Yes, alcohol and/or drug (check all applicable)	1
Alcohol Drug	
Characteristics of children in household (check appropriate items and add for score)	
a. Not applicable	0
b. Delinquency history	1
c. Developmental disability	1
d. Mental health/behavioral problem	1
TOTAL ABUSE RISK SCORE	

Appendix C - Family Assessment

Family Assessment, Case Plan And Safety Assessment

CASE NAME/CASE NUMBER		ASSESSMENT REASON
PARENT / CAREGIVER(S)	CHILD(REN)	
EFFECTIVE DATE	SOCIAL WORKER	PROGRAM TYPE ☐ FVS ☐ CFWS ☐ FRS

I. MALTREATMENT

Maltreatment/Presenting Problem:

Describe the nature and extent of the maltreatment. Include collateral and other witnesses' description of the threat to safety and maltreatment.

Describe the abuse or neglect, both reported and found, and the impact on the child in terms of physical or emotional impacts.

Sequence of Events:

Describe the sequence of events that led up to the maltreatment (this includes information gathered from victims, siblings/other household members, non-offending caregiver(s), perpetrator, and collateral sources).

DSHS 15-421 (REV. 11/2011) Page 1

II. FAMILY FUNCTIONING

Safety Assessment:

Yes	No	
☐	☐	The family *situation* results in no adults in the home performing parenting duties and responsibilities that assure child's safety.
☐	☐	The family *situation* is that the living arrangement(s) seriously endanger the child's physical health.
☐	☐	Caregiver(s) do not have or do not use resources necessary to meet the child's immediate basic needs which present an immediate threat of serious harm to a child.
☐	☐	Caregiver(s) lack the parenting knowledge, skills, or motivation necessary to assure a child's safety.
☐	☐	Caregiver(s) are not meeting, cannot meet, or will not meet the child's exceptional physical, emotional, medical or behavioral needs.
☐	☐	Child sexual abuse is suspected, has occurred, or circumstances suggest sexual abuse is likely to occur.

Developmental Stages and Tasks:

Describe the developmental stage(s) of the family and the overall tasks the family typically faces. Include information identifying the family's culture and how they accomplish their everyday life tasks.

Describe the specific task(s) that cause or contribute to the safety threats. Describe the family's interactions and difficulty in achieving the task(s).

Describe past exceptions in how the family has handled this difficult task. Include information and evidence of the family's parenting practices regarding other everyday life tasks (e.g., medical needs, morning/evening routines, supervision) and provide strengths and concerns.

Family Choice of Discipline:

Describe the disciplinary approaches used by the parents/caregivers. Include strengths (e.g., uses self control while disciplining child and is fair and consistent) and concerns (e.g., uses violence or threats, discipline is vengeful, physical discipline stems from frustration and/or anger).

Family Support:

Describe the family's support system. Include any negative or positive impacts these supports may have had while the family used them in the past. Describe how these support systems help or may help the family protect the child(ren). Describe areas in the family life where additional supports may benefit the family.

III. PARENT/CAREGIVER FUNCTIONING

Safety Assessment:

Yes	No	
☐	☐	Caregiver(s) are acting *(behaving)* violently or dangerously and the behaviors impact child safety.
☐	☐	There has been an incident of domestic violence that impacts child safety.

Yes	No	
☐	☐	The domestic violence perpetrator has caused serious harm or threats of harm against the adult victim/caregiver of the child.
☐	☐	The domestic violence perpetrator has seriously harmed or threatened serious harm to the child.
☐	☐	The level of violence and/or threats towards either the adult victim or child is increasing so that serious harm is likely to occur.
☐	☐	There are other indications of increased dangers from the domestic violence perpetrator such as suicide threats or attempts, substance abuse or threats with weapons.

☐	☐	Caregiver(s) will not or cannot control their *behavior* and their *behavior* impacts child safety.
☐	☐	Caregiver(s) perceives child in *extremely* negative terms.
☐	☐	Caregiver(s) *attitude, emotions* and *behavior* threaten severe harm to a child, or caregiver(s) fear they will maltreat the child and are requesting placement.
☐	☐	Caregiver(s) intend(ed) to seriously hurt the child.
☐	☐	Caregiver(s) overtly rejects CA intervention, refuses access to a child, or there is some indication that the caregiver(s) will flee.
☐	☐	Caregiver(s) cannot or will not explain child's injuries or maltreating condition(s) or explanation is not consistent with the facts.

PARENT/CAREGIVER NAME

Describe how the parent/caregiver loses control and exhibits behaviors (e.g., substance use/abuse, violent, depression, etc.) that led to a disruption in meeting specific everyday life tasks. Describe the individual's patterns for their loss of control.

Describe the information and evidence collected regarding the parent/caregiver that indicates prevention skills are needed or have been learned to manage the identified behaviors. Include behavioral strengths and exceptions to the problem. Evidence may include but is not limited to professionals (e.g., mental health, substance abuse, law enforcement, relatives, etc).

Describe how the parent/caregiver functions in respect to daily life management and general adaptation, independent of their parenting abilities. Include descriptions of strengths and concerns in adult functioning. Identify primary ways of coping with day-to-day life.

Describe the parent/caregiver's behavioral, cognitive, and emotional capacity to protect their children.

IV. CHILD FUNCTIONING

Safety Assessment:

Yes	No	
☐	☐	A child has serious physical injuries or serious physical conditions resulting from maltreatment.
☐	☐	A child demonstrates serious emotional symptoms, self-destructive behavior and/or lack of behavioral control that results in provoking dangerous reactions in caregivers.
☐	☐	A child is extremely fearful of the home situation or people within the home.

CHILD NAME

Describe how the child functions on a daily basis. Include behaviors, feelings, cognitive functioning, physical capacity, temperament, relationships, etc. Include information on their ability to accomplish developmentally appropriate tasks.

Identify strengths and concerns using behaviorally specific descriptors (i.e., if developmentally on target what is observed that indicates that) and any child related issues which may cause stress on the family (i.e., substance use, running away, health).

Describe the child's development; specifically including cultural, educational, vocational, health, mental health, independent living skills, vocational, and peer/community relationships.

V. SAFETY DECISIONS	
FINAL SAFETY DECISION ☐ Safe ☐ Unsafe	FINAL SAFETY PLAN DECISION ☐ No Plan Required ☐ In-Home Safety Plan ☐ Out-of-Home Safety Plan

SIGNATURES			
PARENT/GUARDIAN SIGNATURE	DATE	PARENT/GUARDIAN SIGNATURE	DATE
CHILD (OVER 12 YRS) SIGNATURE	DATE	OTHER SIGNATURE	DATE
SOCIAL WORKER SIGNATURE	DATE	SUPERVISOR SIGNATURE	DATE

Case Plan

☐ Initial Plan
☐ Follow-up Plan

The Case Plan specifies what must change to reduce or eliminate safety threats and increase the parent or caregiver's protective capacities to assure the child's safety and well being.

☐ In-Home Case Plan: This plan is designed to keep children in their home.
If sufficient progress is not made by the parent/caregiver, the planned arrangement for the child is placement out of the parent's home.

☐ Out-of-Home Case Plan: This plan is designed to assist in the child's timely and safe return home.
If sufficient progress is not made by the parent / caregiver, the case plan is used to help achieve a permanent plan other than return home.

CAREGIVER(S)	CHILD(REN)

Native American Heritage? ☐ Yes ☐ No (If Yes, Refer to ICW Manual for Policy Requirements Related to Voluntary Case Plan.	DATE PLAN BEGINS	DATE PLAN REVIEWED

FAMILY LEVEL OBJECTIVE

OBJECTIVE

OBJECTIVE START DATE		TARGET END DATE	

TASKS

SERVICES

SERVICE

PROVIDER

START DATE		END DATE	

SERVICE

PROVIDER

START DATE		END DATE	

SERVICE

PROVIDER

START DATE		END DATE	

SERVICE

PROVIDER			
START DATE		END DATE	
INDIVIDUAL LEVEL OBJECTIVE			
PARENT/CAREGIVER NAME			
OBJECTIVE			
OBJECTIVE START DATE		TARGET END DATE	
TASKS			
SERVICES			
SERVICE			
PROVIDER			
START DATE		END DATE	
SERVICE			
PROVIDER			
START DATE		END DATE	
SERVICE			
PROVIDER			
START DATE		END DATE	
SERVICE			
PROVIDER			
START DATE		END DATE	
CHILD ACTION PLAN			
CHILD NAME			
OBJECTIVE			
OBJECTIVE START DATE		TARGET END DATE	

TASKS			
SERVICES			
SERVICE			
PROVIDER			
START DATE		**END DATE**	
SERVICE			
PROVIDER			
START DATE		**END DATE**	
SERVICE			
PROVIDER			
START DATE		**END DATE**	
SERVICE			
PROVIDER			
START DATE		**END DATE**	
SIGNATURES			
PARENT/CAREGIVER SIGNATURE	DATE	PARENT/CAREGIVER SIGNATURE	DATE
CHILD (OVER 12 YEARS) SIGNATURE	DATE	OTHER SIGNATURE	DATE
SOCIAL WORKER SIGNATURE	DATE	SUPERVISOR SIGNATURE	DATE

Appendix D

IRT Item Parameter Calibration Report

CTS-2 All Victim Items

Report created on 12/9/2013

Introduction

This report provides the results of the IRT item parameter calibration by the computer program Xcalibre Version 4.2.0.1 (Assessment Systems Corporation, 2012) for CTS Vict all items. The output is divided into four sections:

1. Specifications
2. E-M Algorithm
3. Summary statistics
4. Item-by-item results.

The statistical output is also recorded in a comma-separated value (CSV) file of the same name.

Specifications

This section records the input/output specifications and settings for historical purposes.

The Windows paths for the input files used in this analysis were:

C:\Users\Douglas Klinman\Documents\IRT_CTSc+e.csv
C:\Users\Douglas Klinman\Documents\IRT_CTSc+eCon.csv

The Windows paths for the output files produced by this analysis were:

CTS_V_All.rtf
CTS_V_All.csv
CTS_V_All Scores.csv

Table 1 presents the file specifications. Table 2 presents the IRT specifications used to perform the IRT item parameter calibration. Table 3 presents the flag specifications.

Table 1: File Specifications

Specification	Value	Specification	Value
Number of examinees	237	Total Items	36
Calibrated Items	18	Pretest Items	0
Excluded Items	18	Number of domains	1
Classic Data Header	No	Delimited input	Yes
Delimiter for input	Comma	Number of ID columns	N/A
ID begins in column	N/A	Responses begin in column	N/A
Omit character	O	Not Admin character	-
Save item parameters	No	Item parameter format	N/A
Save data matrix	No	Omit codes are	N/A
Not Admin codes are	N/A	Score Not Admin as omits	No
Plot the IRFs	Yes	Save the IRFs and IIFs	No
Produce the fit line	Yes	# Groups for Plot	20
Type of score groups	Equally sized	# Groups for Chi-square	20
Perform classification	No	Classify using	N/A
Two-group cutpoint	N/A	Low group label	N/A
High group label	N/A	Merge empty poly categories	N/A

Table 2: IRT Calibration Specifications

Specification	Value	Specification	Value
IRT Specification	Dichotomous	Model constant	1.7
Polytomous IRT Model	N/A	Dichotomous IRT Model	2-parameter
Center the boundary locations	No	Centered value	N/A
Floating priors	Yes	a parameter prior mean (sd)	1.000 (0.250)
b parameter prior mean (sd)	0.000 (1.000)	c parameter prior mean (sd)	0.250 (0.025)
Theta estimation method	MLE	Bayesian prior mean (sd)	N/A
Maximum E-M loops	60	Convergence criterion	0.001
Quadrature points	20	Center dich item parameters on	theta
Acceptable P range	0.00 to 1.00	Acceptable item-corr range	-1.00 to 1.00
Acceptable item mean range	0.00 to 15.00	Correct for spuriousness	Yes
Fit statistic critical alpha	0.050	Minimum a	0.05
Maximum a	6.00	Minimum b	-4.00
Maximum b	4.00	Minimum c	0.00
Maximum c	0.70	Minimum theta	-7.00
Maximum theta	7.00	Treat scored items as poly	No
Center poly parameters on theta	No	Test for DIF	No
Group status column	N/A	Ability levels for DIF Test	N/A
Group 1 code	N/A	Group 2 code	N/A
Group 1 label	N/A	Group 2 label	N/A
Exclude items with low N	No	Minimum valid N	N/A
Compute scaled scores	No	Mean (SD) of scaled scores	N/A
Minimum scaled score	N/A	Maximum scaled score	N/A
Save statistics output	Yes	Delimiter	Comma
Save scores output	Yes	Delimiter	Comma
Save test information output	Yes	Delimiter	Comma
Save item information output	Yes	Delimiter	Comma

Table 3: Flag Specifications

Specification	Value	Specification	Value
Low a Flag Bound	0.30	High a Flag Bound	4.00
Low b Flag Bound	-3.00	High b Flag Bound	3.00
Low c Flag Bound	0.00	High c Flag Bound	0.40
Key Flag	K	Fit Flag	F
Low a Flag	La	High a Flag	Ha
Low b Flag	Lb	High b Flag	Hb
Low c Flag	Lc	High c Flag	Hc

E-M Algorithm

Xcalibre uses the expectation-maximization approach to calibrate item parameters. The estimation process is iterative, and repeated in loops until the convergence criterion is satisfied. The following list presents the item with the largest parameter change after each loop, and the value of the change.

The number of loops needed is evidence regarding the fit of the data; if many loops are required, or

convergence is never reached, it means that the data does not fit well with the selected IRT model.

Item 34 failed to converge on this loop
Maximum change after Loop 1 was 3.7165 for Item 13 for the b parameter
Maximum change after Loop 2 was 0.4144 for Item 20 for the a parameter
Maximum change after Loop 3 was 0.2761 for Item 21 for the a parameter
Maximum change after Loop 4 was 0.0936 for Item 20 for the a parameter
Maximum change after Loop 5 was -0.0226 for Item 36 for the a parameter
Maximum change after Loop 6 was 0.0120 for Item 36 for the b parameter
Maximum change after Loop 7 was 0.0035 for Item 19 for the a parameter
Maximum change after Loop 8 was 0.0023 for Item 19 for the a parameter
Maximum change after Loop 9 was 0.0016 for Item 19 for the a parameter
Maximum change after Loop 10 was 0.0011 for Item 19 for the a parameter
Maximum change after Loop 11 was 0.0008 for Item 19 for the a parameter

Summary statistics

Table 4 presents the summary statistics for the item parameters for all calibrated items. Table 5 summarizes the total scores for the full test for just the calibrated items. Table 6 summarizes the theta estimates for the full test. Table 7 provides the overall model fit chi-square(s) for the full test. Definitions of these statistics are found in the Xcalibre manual.

Table 4: Summary Statistics for All Calibrated Items

Parameter	Items	Mean	SD	Min	Max
a	18	1.267	0.299	0.788	1.791
b	18	0.257	0.865	-1.094	1.959

Table 5: Summary Statistics for the Total Scores

Test	Items	Alpha	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	18	0.958	3.954	5.533	1.159	0	0.00	0.0	8.00	18	8.00

Table 6: Summary Statistics for the Theta Estimates

Test	Examinees	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	237	-0.025	0.673	0.645	-7.000	-7.000	-7.000	-0.131	7.000	6.869

Table 7: Overall Model Fit

Test	Items	Chi-square	df	p	-2LL
Full Test	18	283.365	324	0.950	1302

Table 8 presents the item control information and item status for each item

Table 8: Item Control and Item Status for All Items

Seq.	Item ID	Key	Options	Domain	Inclusion	Item Type	Status
1	Q1 MV P	1	2	1	N	P	Not Included
2	Q3 MV P	1	2	1	N	P	Not Included
3	Q7 MV P	1	2	1	N	P	Not Included
4	Q25 MV P	1	2	1	N	P	Not Included
5	Q27 MV P	1	2	1	N	P	Not Included
6	Q6 MI P	1	2	1	N	P	Not Included
7	Q34 MI P	1	2	1	N	P	Not Included
8	Q9 SV P	1	2	1	N	P	Not Included
9	Q13 SV P	1	2	1	N	P	Not Included
10	Q17 SV P	1	2	1	N	P	Not Included
11	Q19 SV P	1	2	1	N	P	Not Included
12	Q23 SV P	1	2	1	N	P	Not Included
13	Q31 SV P	1	2	1	N	P	Not Included
14	Q35 SV P	1	2	1	N	P	Not Included
15	Q12 SI P	1	2	1	N	P	Not Included
16	Q16 SI P	1	2	1	N	P	Not Included
17	Q22 SI P	1	2	1	N	P	Not Included
18	Q30 SI P	1	2	1	N	P	Not Included
19	Q2 MV V	1	2	1	Y	P	Included
20	Q4 MV V	1	2	1	Y	P	Included
21	Q8 MV V	1	2	1	Y	P	Included
22	Q26 MV V	1	2	1	Y	P	Included
23	Q28 MV V	1	2	1	Y	P	Included
24	Q5 MI V	1	2	1	Y	P	Included
25	Q33 MI V	1	2	1	Y	P	Included
26	Q10 SV V	1	2	1	Y	P	Included
27	Q14 SV V	1	2	1	Y	P	Included
28	Q18 SV V	1	2	1	Y	P	Included
29	Q20 SV V	1	2	1	Y	P	Included
30	Q24 SV V	1	2	1	Y	P	Included
31	Q32 SV V	1	2	1	Y	P	Included
32	Q36 SV V	1	2	1	Y	P	Included
33	Q11 SI V	1	2	1	Y	P	Included
34	Q15 SI V	1	2	1	Y	P	Included
35	Q21 SI V	1	2	1	Y	P	Included
36	Q29 SI V	1	2	1	Y	P	Included

Table 9 presents the classical statistics, the item parameters, and any flags for each calibrated item. The K flag indicates that the keyed alternative did not have the highest correlation with total score. The F flag indicates that the item fit statistic (z Resid for dichotomous / chi-square for polytomous) was significant, and the item did not fit the IRT model.The La, Lb, and Lc flags indicate that the a/b/c parameters were lower than the minimum acceptable value.The Ha, Hb, and Hc flags indicate that the a/b/c parameters were higher than the maximum acceptable value

Table 9: Item Parameters for All Calibrated Items

Seq.	Item ID	P	R	a	b	Flag(s)
1	Q2 MV V	0.321	0.777	1.414	-0.539	
2	Q4 MV V	0.245	0.793	1.472	-0.077	
3	Q8 MV V	0.414	0.766	1.791	-1.094	F
4	Q26 MV V	0.342	0.822	1.615	-0.677	
5	Q28 MV V	0.278	0.775	1.195	-0.262	
6	Q5 MI V	0.308	0.835	1.548	-0.474	
7	Q33 MI V	0.308	0.847	1.612	-0.477	
8	Q10 SV V	0.131	0.678	1.087	0.869	
9	Q14 SV V	0.287	0.811	1.272	-0.325	
10	Q18 SV V	0.165	0.723	1.071	0.574	
11	Q20 SV V	0.245	0.858	1.604	-0.091	
12	Q24 SV V	0.219	0.819	1.350	0.100	
13	Q32 SV V	0.046	0.489	1.105	1.959	
14	Q36 SV V	0.143	0.700	1.097	0.747	
15	Q11 SI V	0.114	0.568	0.788	1.237	
16	Q15 SI V	0.160	0.640	0.855	0.718	
17	Q21 SI V	0.165	0.718	1.073	0.573	
18	Q29 SI V	0.064	0.473	0.848	1.860	

Figure 1 displays the distribution of the theta estimates for all calibrated items.
Table 10 displays the frequency distribution for the theta estimates.

Figure 1: Theta Estimates for All Calibrated Items

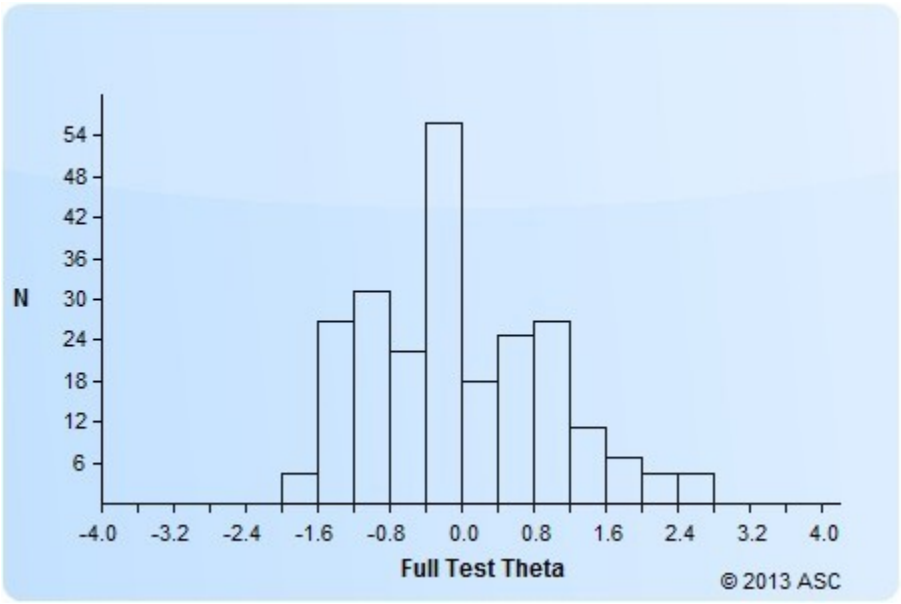


Table 10: Frequency Distribution for Full Test Theta

Range	Frequency
Below -4	125
-4.0 to -3.6	0
-3.6 to -3.2	0
-3.2 to -2.8	0
-2.8 to -2.4	0
-2.4 to -2.0	0
-2.0 to -1.6	2
-1.6 to -1.2	12
-1.2 to -0.8	14
-0.8 to -0.4	10
-0.4 to 0.0	25
0.0 to 0.4	8
0.4 to 0.8	11
0.8 to 1.2	12
1.2 to 1.6	5
1.6 to 2.0	3
2.0 to 2.4	2
2.4 to 2.8	2
2.8 to 3.2	0
3.2 to 3.6	0
3.6 to 4.0	0
Above +4	6

Figure 2 displays the distribution of the a parameters.
Table 11 displays the frequency distribution of the a parameters shown in Figure 2.

Figure 2: Histogram of the a Parameters

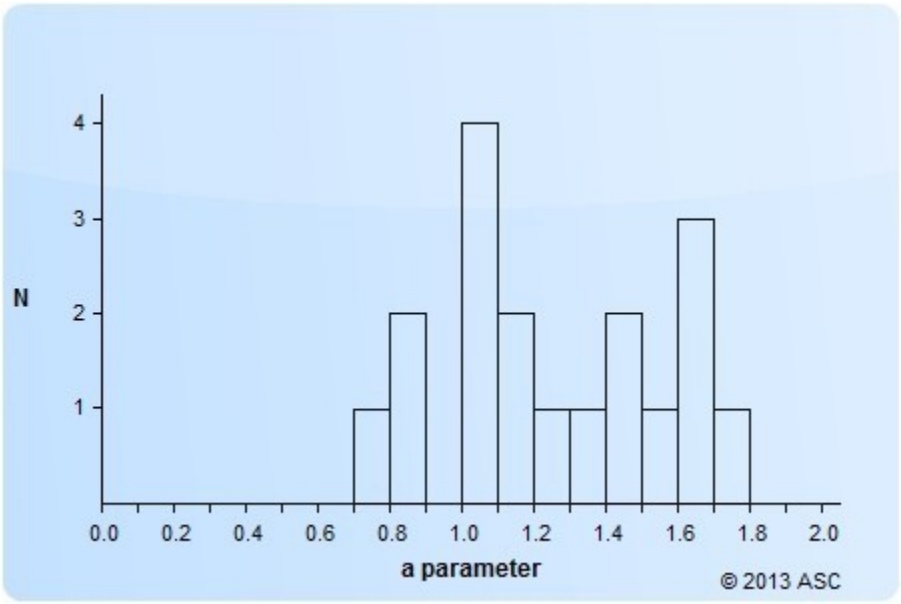


Table 11: Frequency Distribution for the a Parameters

Range	Frequency
0.00 to 0.10	0
0.10 to 0.20	0
0.20 to 0.30	0
0.30 to 0.40	0
0.40 to 0.50	0
0.50 to 0.60	0
0.60 to 0.70	0
0.70 to 0.80	1
0.80 to 0.90	2
0.90 to 1.00	0
1.00 to 1.10	4
1.10 to 1.20	2
1.20 to 1.30	1
1.30 to 1.40	1
1.40 to 1.50	2
1.50 to 1.60	1
1.60 to 1.70	3
1.70 to 1.80	1
1.80 to 1.90	0
1.90 to 2.00	0

Figure 3 displays the distribution of the b parameters.
Table 12 displays the frequency distribution of the b parameters shown in Figure 3.

Figure 3: Histogram of the b Parameters

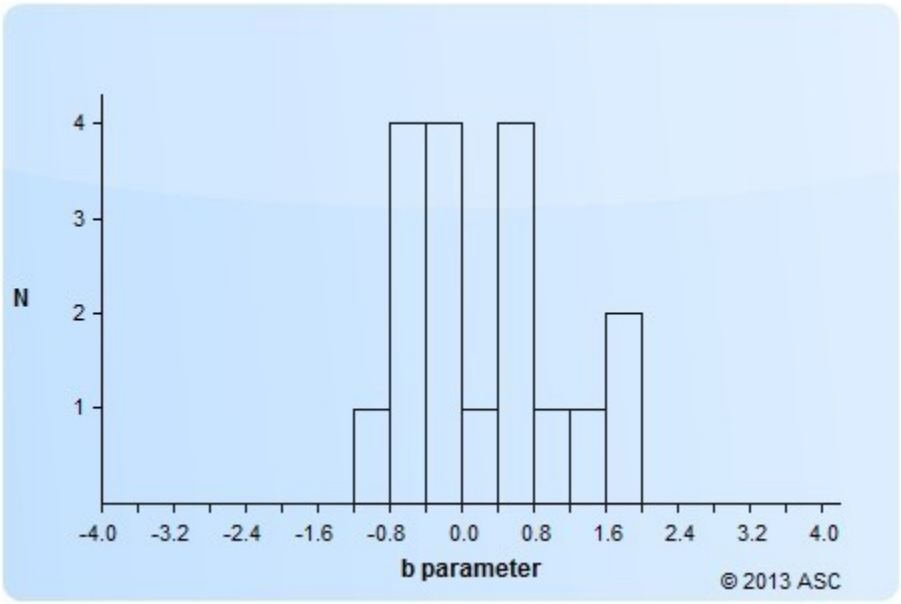


Table 12: Frequency Distribution for the b Parameters

Range	Frequency
-4.0 to -3.6	0
-3.6 to -3.2	0
-3.2 to -2.8	0
-2.8 to -2.4	0
-2.4 to -2.0	0
-2.0 to -1.6	0
-1.6 to -1.2	0
-1.2 to -0.8	1
-0.8 to -0.4	4
-0.4 to 0.0	4
0.0 to 0.4	1
0.4 to 0.8	4
0.8 to 1.2	1
1.2 to 1.6	1
1.6 to 2.0	2
2.0 to 2.4	0
2.4 to 2.8	0
2.8 to 3.2	0
3.2 to 3.6	0
3.6 to 4.0	0

Figure 4 displays the scatterplot of the b parameter (difficulty) by the a parameter (discrimination) for all calibrated items.

Figure 4: b Parameter by a Parameter

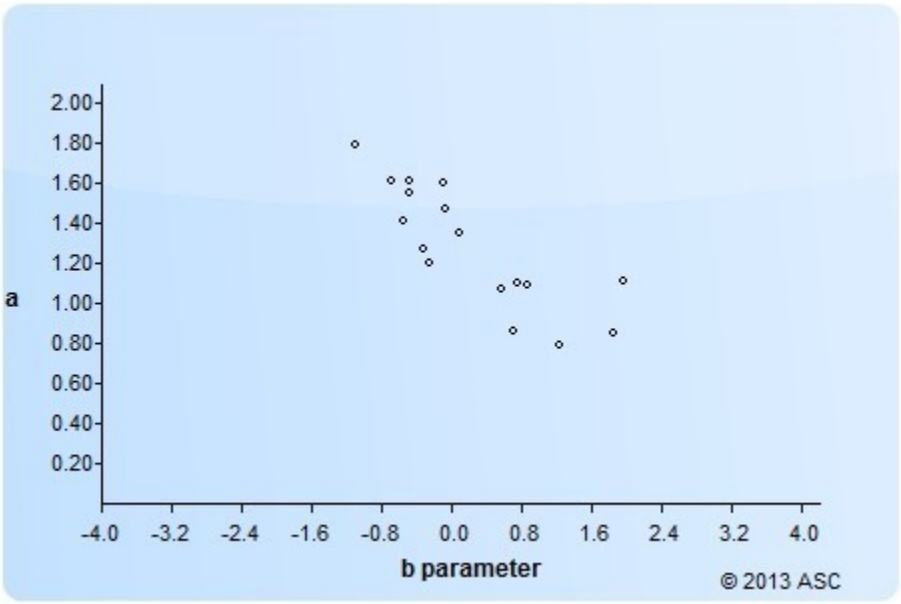


Figure 5 displays the joint distribution of the b parameter by Theta.

Figure 5: b parameter by Theta

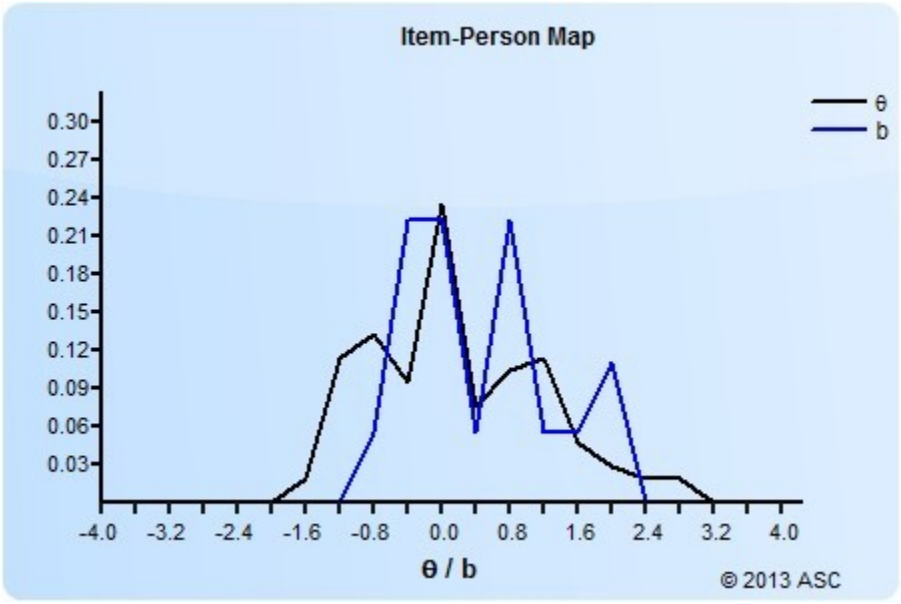


Figure 6 displays a graph of the Test Response Function (TRF) for all calibrated items. The TRF predicts the proportion or number of items that an examinee would answer correctly as a function of theta. The left Y-axis is in proportion correct units while the right Y-axis is in number-correct units.

Figure 6: Test Response Function

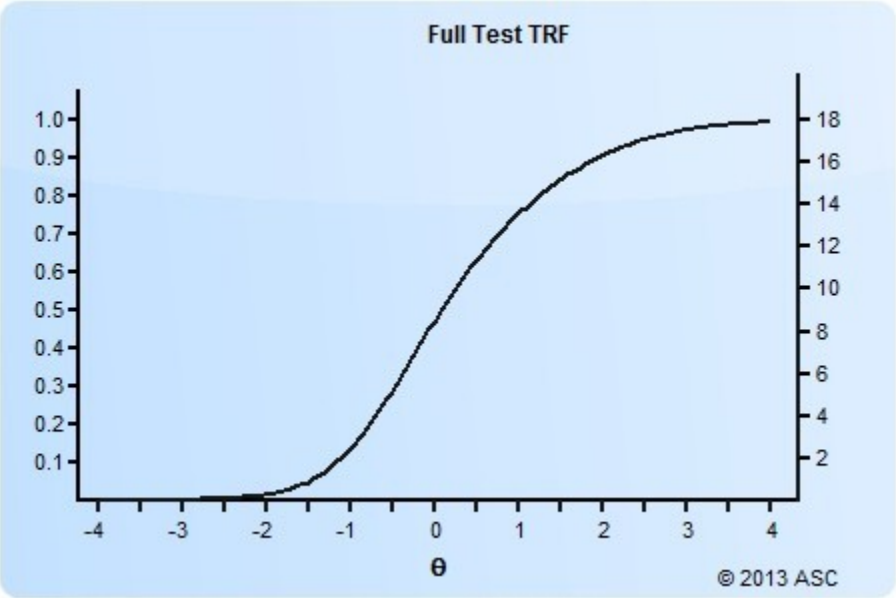


Figure 7 displays a graph of the Test Information Function for all calibrated items. The TIF is a graphical representation of how much information the test is providing at each level of theta. Maximum information was 15.596 at theta = -0.300.

Figure 7: Test Information Function

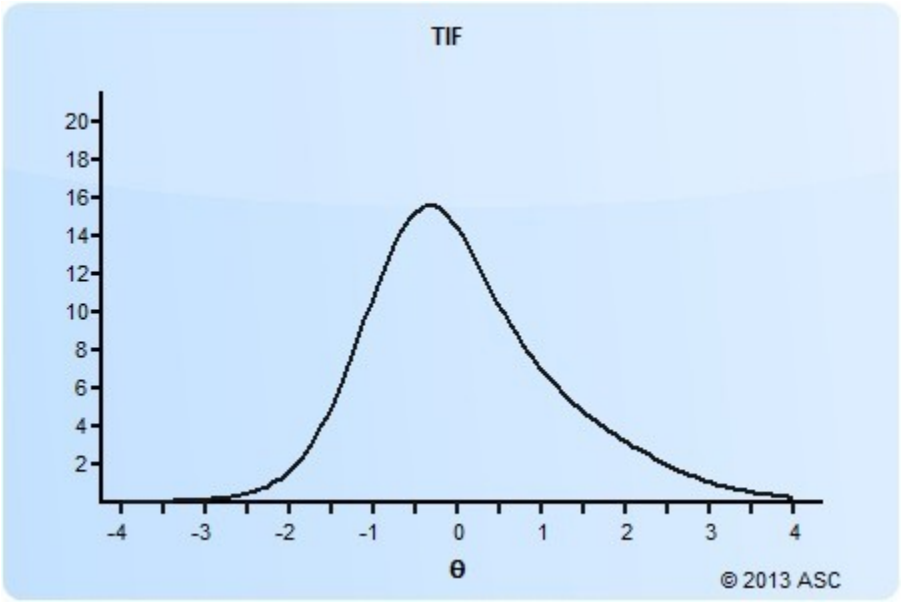
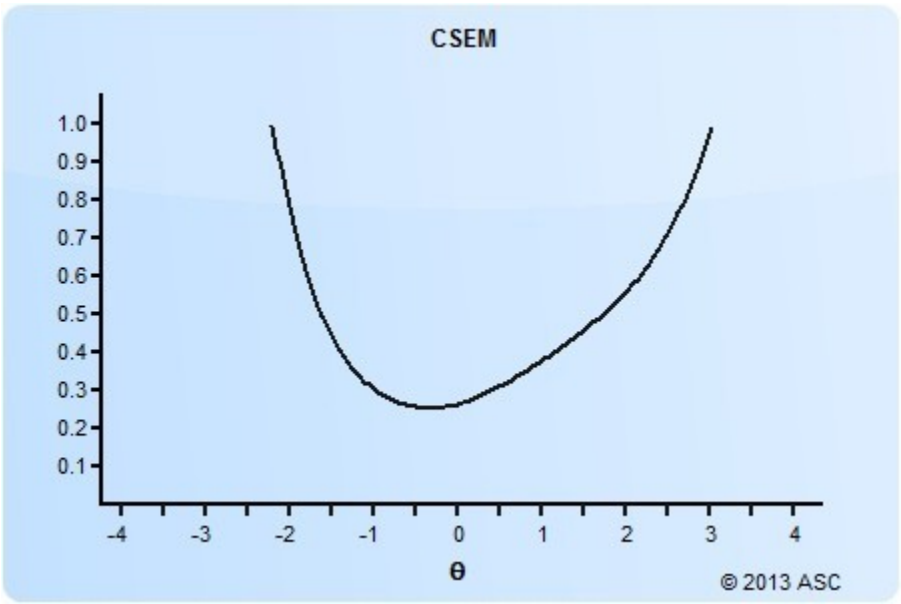


Figure 8 displays a graph of the Conditional Standard Error of Measurement (CSEM) Function. The CSEM is an inverted function of the TIF, and estimates the amount of error in theta estimation for each level of theta. The minimum CSEM was 0.253 at theta = -0.300.

Figure 8: CSEM Function



Item-by-item results

The following section presents the item-by-item results of the analysis. Each scored item has four tables and a plot of the item response function (IRF). The red line (fit line) represents the observed proportion correct conditional on theta. Large deviations of the red line from the IRF are suggestive of poor item fit. Thus, the fit line could be used to identify why items are not fitting the chosen IRT model.

There are four tables presented for each item.

1. Item information table: records the information supplied by the control file (or Classic Data Header) for this item.
2. Classical statistics table: classical statistics for the item.
3. IRT parameters table: item parameter estimates for the item.
4. Option/Category statistics: detailed statistics for each item, which helps diagnose issues in items with poor statistics.

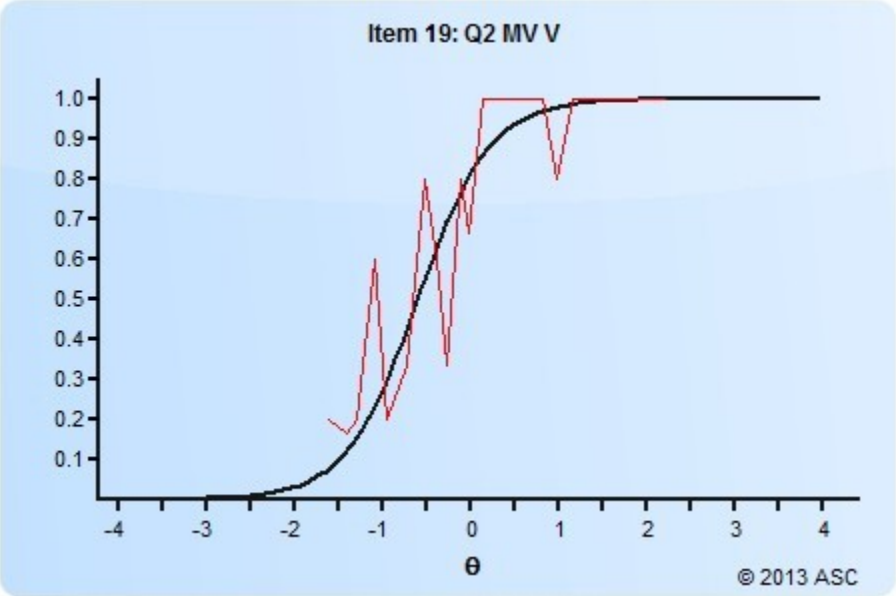
The classical statistics presents classical summary statistics for the item. For multiple choice items the P value and the point-biserial correlations are presented in the first three columns of the table. The P value is the proportion of examinees that answered an item in the keyed direction and ranges from 0 to 1. The S-Rpbis and T-Rpbis are the point-biserial correlations of an item with total score and theta, respectively. The Alpha w/o is Cronbach's alpha computed with the current item excluded. The item-total correlation is a measure of the discriminating power of the item and is related to the IRT discrimination parameter.

The IRT parameters table presents the IRT item parameters and the fit statistics. The latent trait theta is expressed on a standardized scale, so a one unit change equals a one standard deviation change. The "a" parameter indexes the discrimination of the item, as larger values for "a" will result in a greater slope of the IRF and indicate the item differentiates examinees well. The "b" parameter is the item difficulty parameter and equals the location on the theta continuum where the probability of a correct response equals .50. It follows that multiple choice items with more positive "b" parameters are more difficult for examinees, as a higher trait level is required to endorse the keyed response 50% of the time.

The standard errors (SE) for each item parameter estimate are also presented in the item parameter table. A large SE for an item parameter (compared to the other items) indicates that the item parameter was poorly estimated. The IRT standardized (z) residual is the last entry in the item parameter table. It indexes the fit of the data to the item response function. For dichotomous items, the p-value for rejecting the item as poor fit was computed using the z residual with the standard normal distribution as its sampling distribution. The chi-square fit statistic and its degrees of freedom are reported for each item. The chi-square fit statistic and its degrees of freedom are reported for each item.

The option statistics table presents statistics for each individual option (alternative). The key thing to examine in this portion of the table is that no distractors have a higher S-Rpbis or T-Rpbis than the correct answer. That indicates that higher scoring examinees are selecting the incorrect answer, which therefore might be arguably correct.

Item 1: My partner threw something at me that could hurt.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
19	Q2 MV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.321	0.777	0.783	0.955

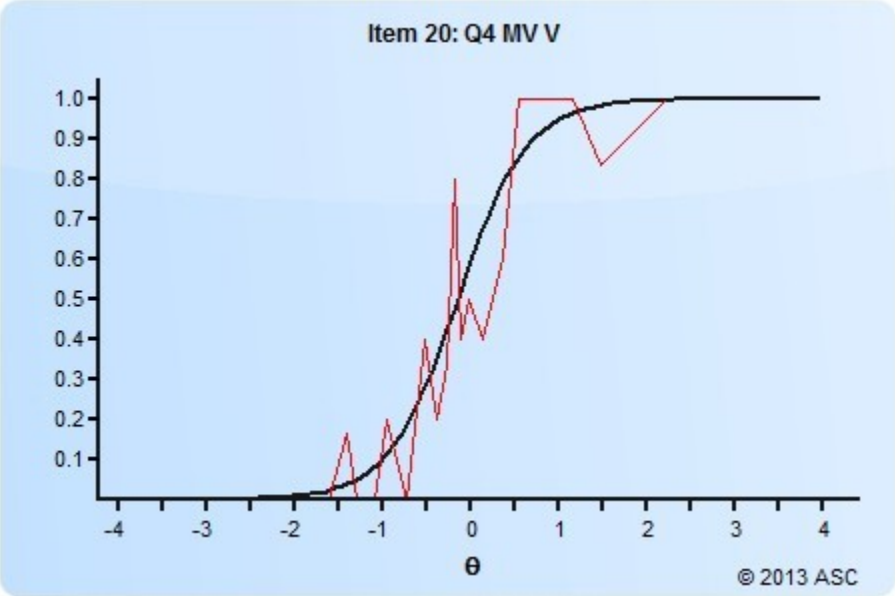
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.414	-0.539	0.194	0.113	20.040	18	0.331	1.066	0.286

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	161	0.679	-0.777	-0.783	-5.617	2.602	
1	76	0.321	0.777	0.783	0.903	2.006	**KEY**
Omit	0						
Not Admin	0						

Item 2: My partner twisted my arm or hair.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
20	Q4 MV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.245	0.793	0.709	0.955

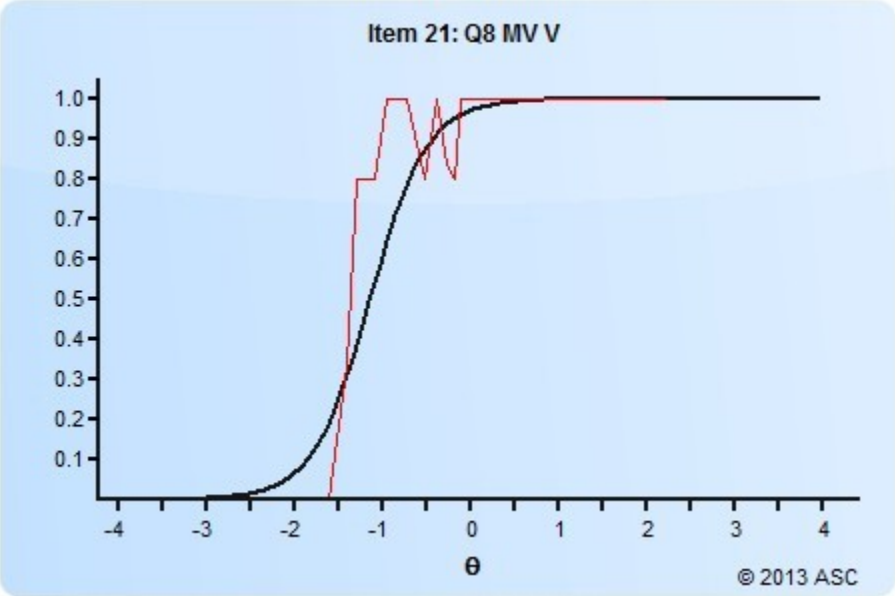
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.472	-0.077	0.194	0.109	20.545	18	0.303	0.664	0.506

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	179	0.755	-0.793	-0.709	-5.095	2.927	
1	58	0.245	0.793	0.709	1.318	2.108	**KEY**
Omit	0						
Not Admin	0						

Item 3: My Partner pushed or shoved me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
21	Q8 MV V	2PL	1	Yes	2	1	F

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.414	0.766	0.886	0.956

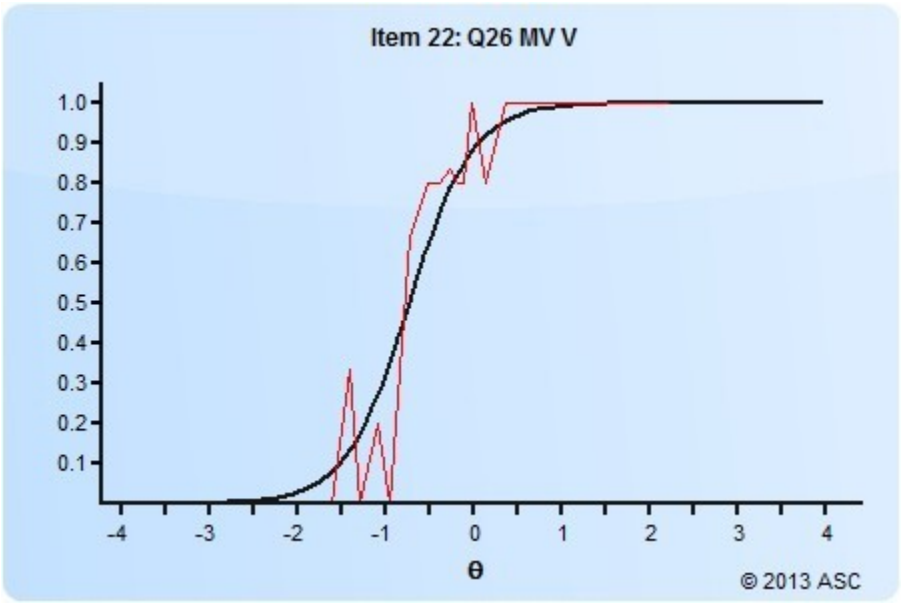
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.791	-1.094	0.210	0.110	16.182	18	0.580	2.086	0.037

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	139	0.586	-0.766	-0.886	-6.416	1.757	
1	98	0.414	0.766	0.886	0.574	1.886	**KEY**
Omit	0						
Not Admin	0						

Item 4: My partner grabbed me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
22	Q26 MV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.342	0.822	0.816	0.954

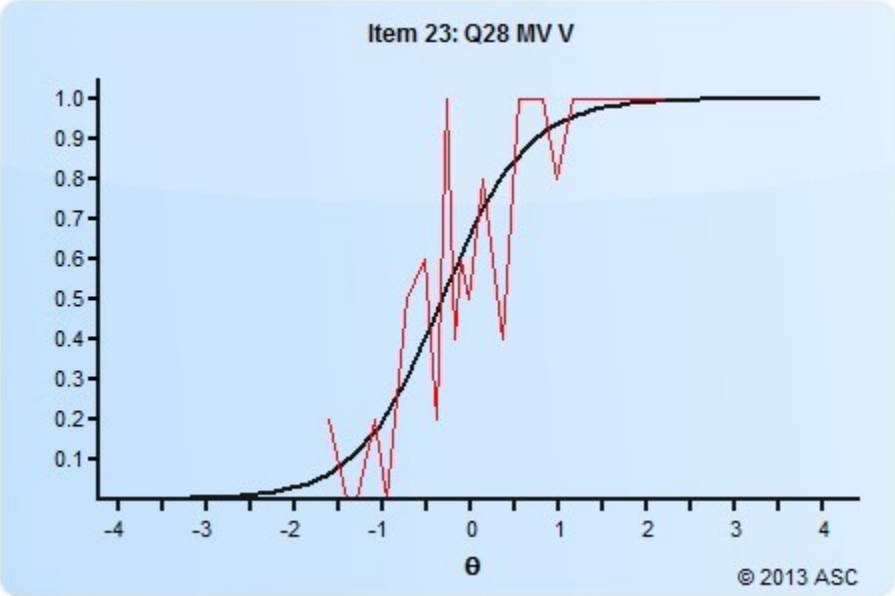
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.615	-0.677	0.193	0.107	10.631	18	0.909	1.359	0.174

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	156	0.658	-0.822	-0.816	-5.811	2.405	
1	81	0.342	0.822	0.816	0.875	1.936	**KEY**
Omit	0						
Not Admin	0						

Item 5: My partner slapped me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
23	Q28 MV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.278	0.775	0.739	0.955

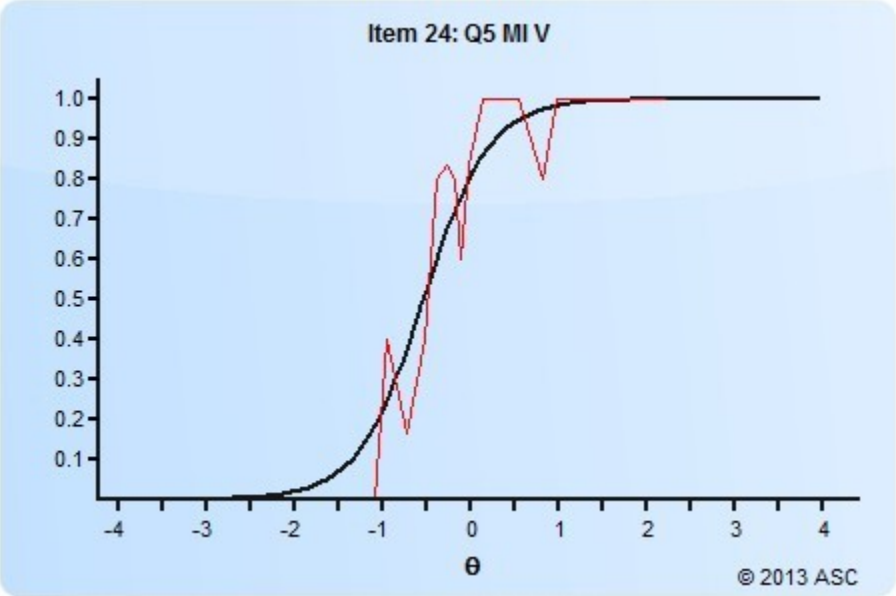
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.195	-0.262	0.196	0.125	22.628	18	0.205	1.170	0.242

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	171	0.722	-0.775	-0.739	-5.311	2.814	
1	66	0.278	0.775	0.739	1.098	2.068	**KEY**
Omit	0						
Not Admin	0						

Item 6: I had a sprain, bruise, or small cut because of a fight with my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
24	Q5 MI V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.308	0.835	0.783	0.954

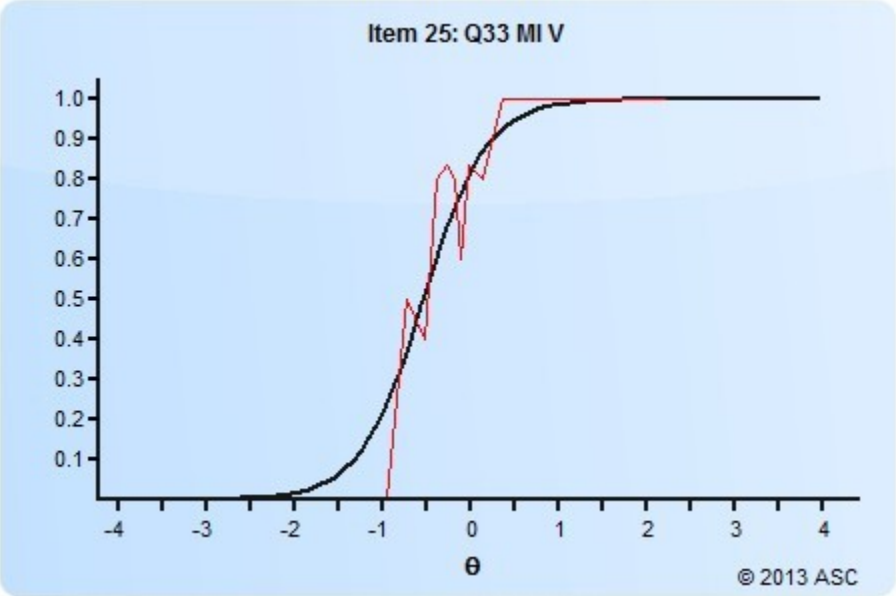
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.548	-0.474	0.194	0.106	13.961	18	0.732	1.316	0.188

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	164	0.692	-0.835	-0.783	-5.556	2.607	
1	73	0.308	0.835	0.783	1.035	1.962	**KEY**
Omit	0						
Not Admin	0						

Item 7: I felt physical pain that still hurt the next day because of a fight with my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
25	Q33 MI V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.308	0.847	0.786	0.954

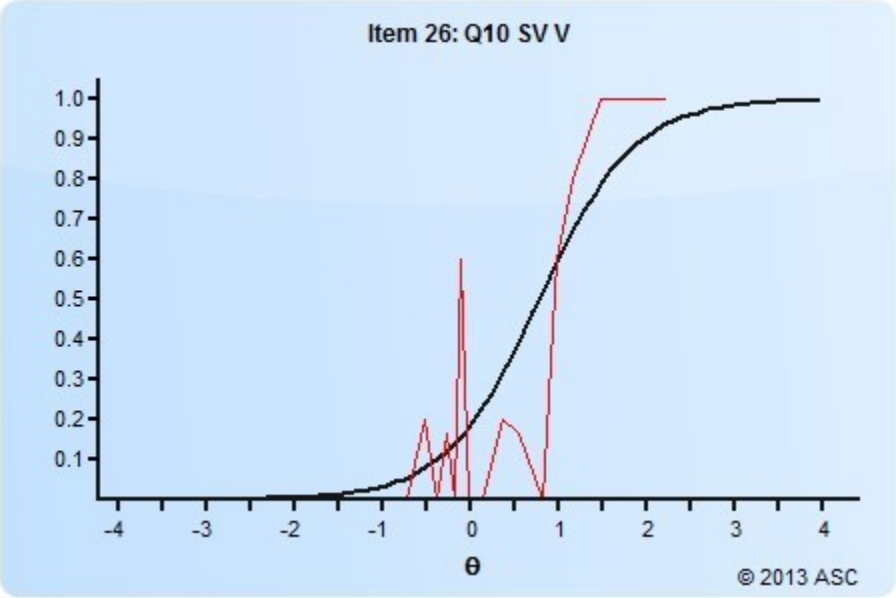
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.612	-0.477	0.194	0.104	8.246	18	0.975	1.362	0.173

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	164	0.692	-0.847	-0.786	-5.563	2.593	
1	73	0.308	0.847	0.786	1.050	1.953	**KEY**
Omit	0						
Not Admin	0						

Item 8: My partner used a knife or gun on me.



Item information

Seq.	ID	Model	Key	Scored	Num Options	Domain	Flags
26	Q10 SV V	2PL	1	Yes	2	1	

Classical statistics

N	P	S-Rpbis	T-Rpbis	Alpha w/o
237	0.131	0.678	0.577	0.957

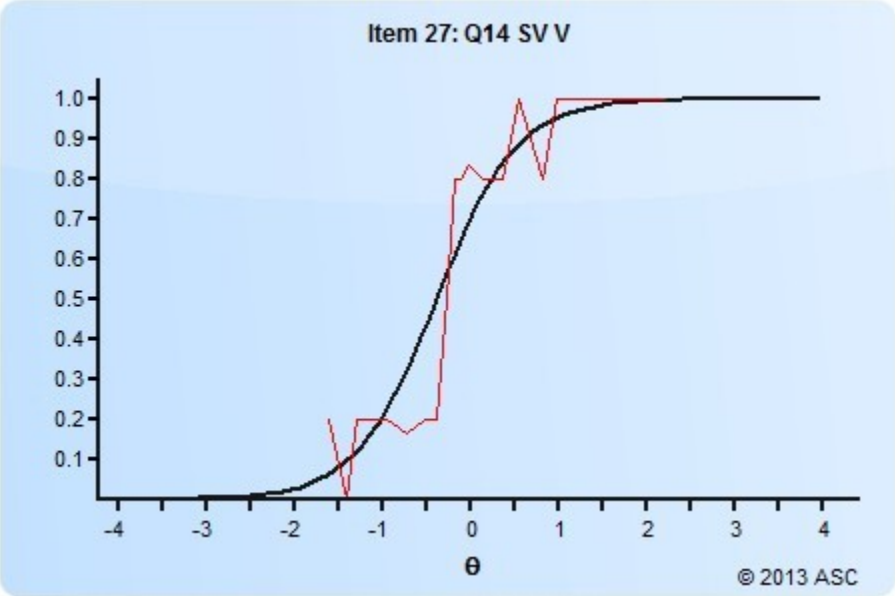
IRT parameters

a	b	a SE	b SE	Chi-sq	df	p	z Resid	p
1.087	0.869	0.183	0.153	23.044	18	0.189	0.675	0.500

Option statistics

Option	N	Prop.	S-Rpbis	T-Rpbis	Mean	SD	
0	206	0.869	-0.678	-0.577	-4.396	3.278	
1	31	0.131	0.678	0.577	2.255	2.486	**KEY**
Omit	0						
Not Admin	0						

Item 9: My partner punched or hit me with something that could hurt.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
27	Q14 SV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.287	0.811	0.754	0.955

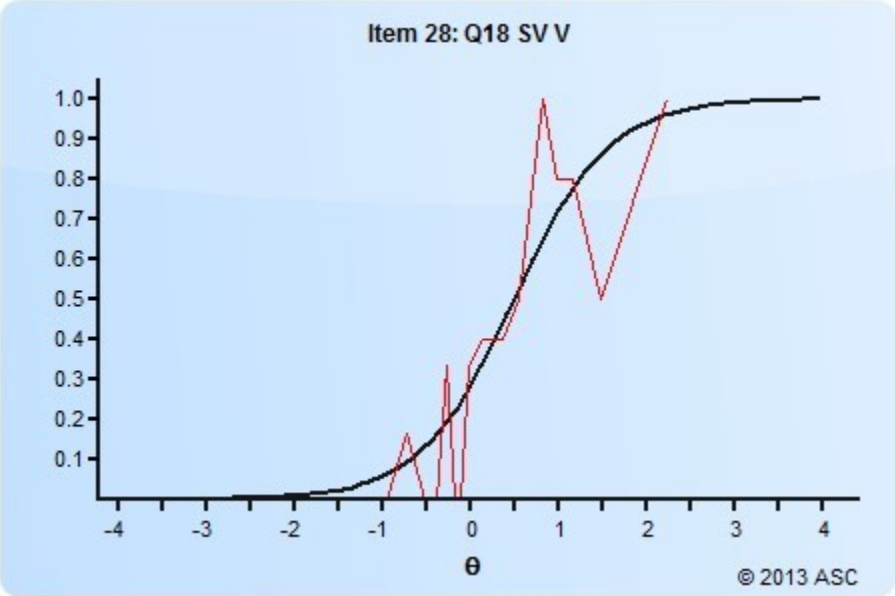
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.272	-0.325	0.195	0.120	10.609	18	0.910	1.119	0.263

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	169	0.713	-0.811	-0.754	-5.386	2.745	
1	68	0.287	0.811	0.754	1.096	2.031	**KEY**
Omit	0						
Not Admin	0						

Item 10: My partner choked me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
28	Q18 SV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.165	0.723	0.615	0.956

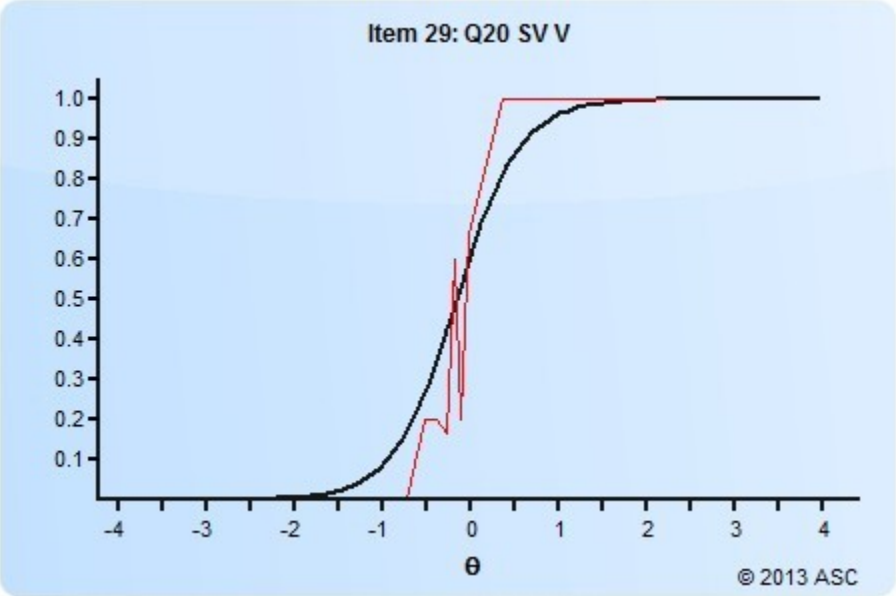
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.071	0.574	0.185	0.145	16.088	18	0.586	0.845	0.398

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	198	0.835	-0.723	-0.615	-4.586	3.202	
1	39	0.165	0.723	0.615	1.856	2.336	**KEY**
Omit	0						
Not Admin	0						

Item 11: My partner slammed me against the wall.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
29	Q20 SV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.245	0.858	0.723	0.954

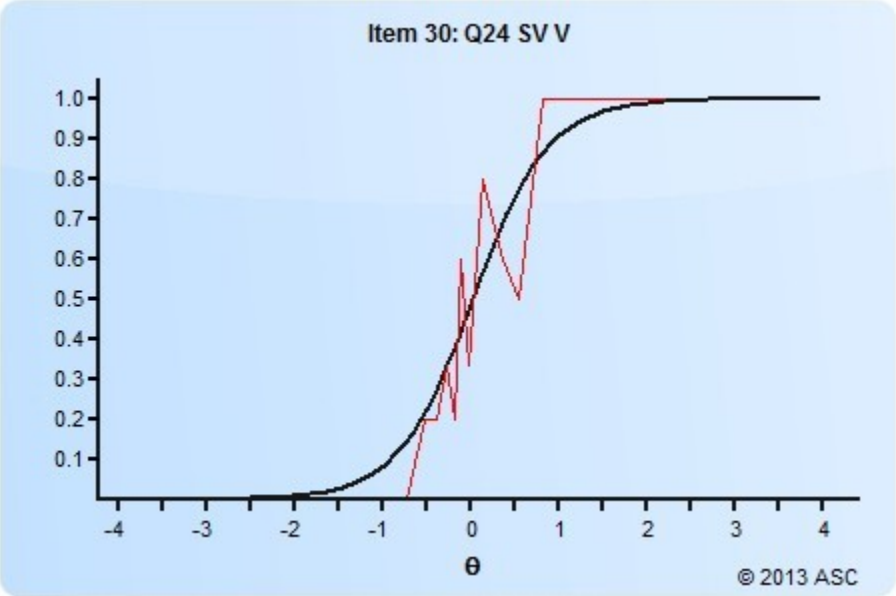
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.604	-0.091	0.195	0.103	9.857	18	0.936	0.910	0.363

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	179	0.755	-0.858	-0.723	-5.126	2.873	
1	58	0.245	0.858	0.723	1.413	2.039	**KEY**
Omit	0						
Not Admin	0						

Item 12: My partner beat me up.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
30	Q24 SV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.219	0.819	0.690	0.954

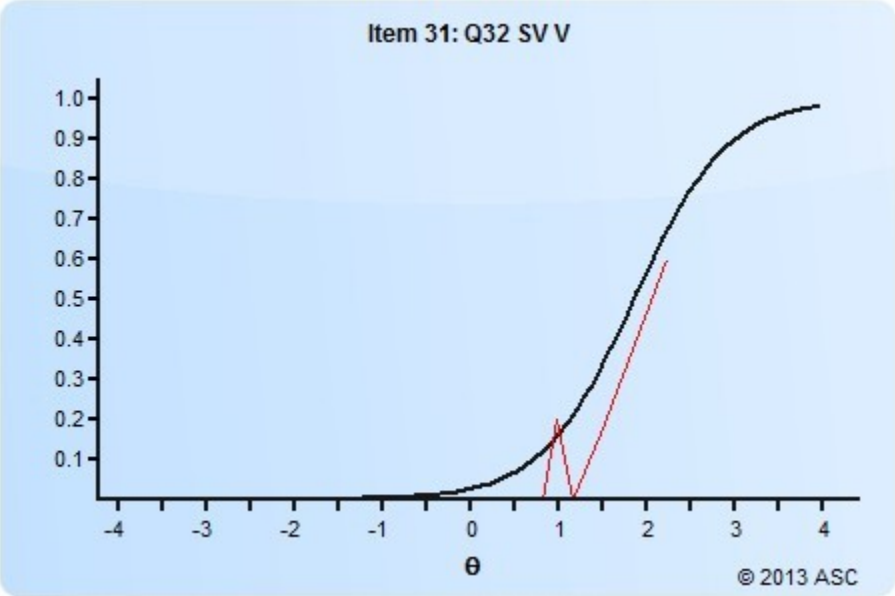
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.350	0.100	0.190	0.117	9.732	18	0.940	0.957	0.338

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	185	0.781	-0.819	-0.690	-4.948	2.991	
1	52	0.219	0.819	0.690	1.532	2.122	**KEY**
Omit	0						
Not Admin	0						

Item 13: My partner burned or scalded me on purpose.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
31	Q32 SV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.046	0.489	0.467	0.959

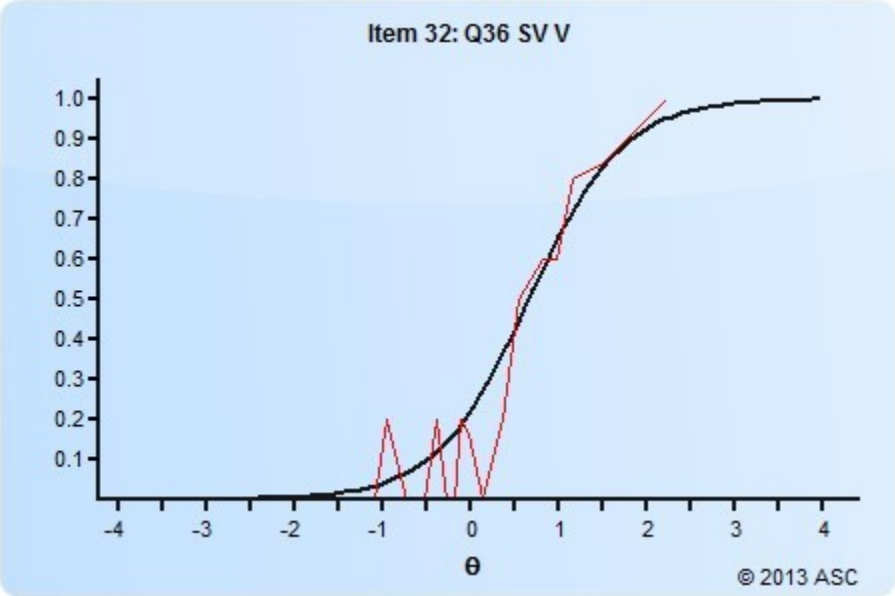
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.105	1.959	0.197	0.225	3.594	18	1.000	0.441	0.659

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	226	0.954	-0.489	-0.467	-3.926	3.481	
1	11	0.046	0.489	0.467	4.704	2.672	**KEY**
Omit	0						
Not Admin	0						

Item 14: My partner kicked me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
32	Q36 SV V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.143	0.700	0.593	0.956

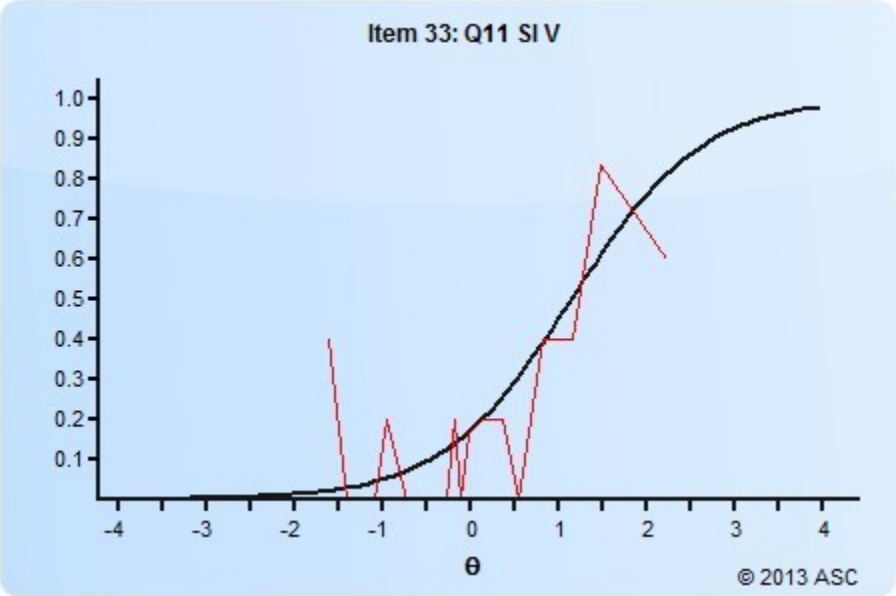
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.097	0.747	0.184	0.148	9.783	18	0.939	0.633	0.527

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	203	0.857	-0.700	-0.593	-4.470	3.244	
1	34	0.143	0.700	0.593	2.109	2.419	**KEY**
Omit	0						
Not Admin	0						

Item 15: I passed out from being hit on the head by my partner in a fight.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
33	Q11 SI V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.114	0.568	0.524	0.958

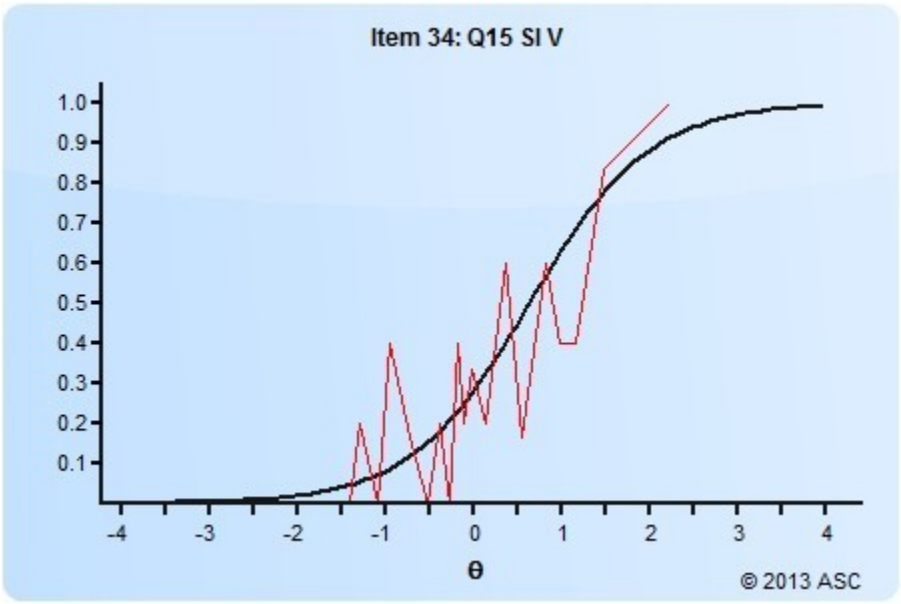
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.788	1.237	0.179	0.203	44.488	18	0.000	0.422	0.673

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	210	0.886	-0.568	-0.524	-4.256	3.381	
1	27	0.114	0.568	0.524	2.153	2.832	**KEY**
Omit	0						
Not Admin	0						

Item 16: I went to the doctor because of a fight with my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
34	Q15 SI V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.160	0.640	0.589	0.957

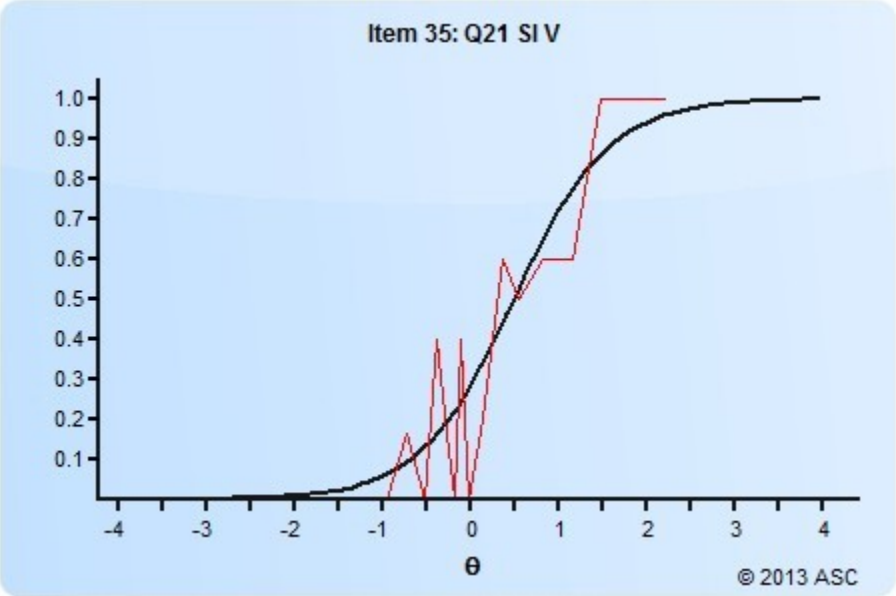
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.855	0.718	0.189	0.172	19.836	18	0.342	0.753	0.451

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	199	0.840	-0.640	-0.589	-4.526	3.261	
1	38	0.160	0.640	0.589	1.714	2.500	**KEY**
Omit	0						
Not Admin	0						

Item 17: I needed to see a doctor because of a fight with my partner, but didn't.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
35	Q21 SI V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.165	0.718	0.616	0.956

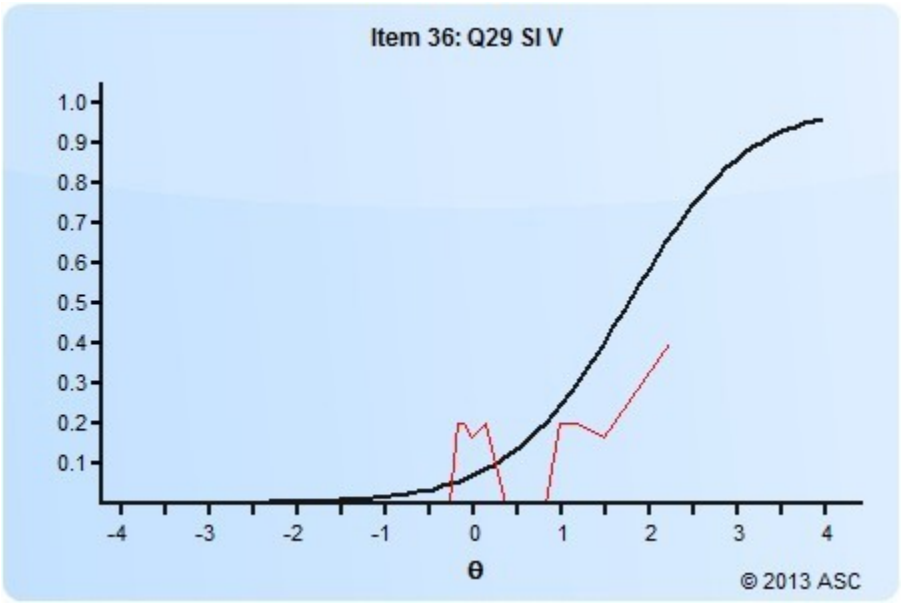
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.073	0.573	0.185	0.145	11.941	18	0.850	0.812	0.417

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	198	0.835	-0.718	-0.616	-4.588	3.197	
1	39	0.165	0.718	0.616	1.864	2.351	**KEY**
Omit	0						
Not Admin	0						

Item 18: I had a broken bone from a fight with my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
36	Q29 SI V	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
236	0.064	0.473	0.462	0.960

IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.848	1.860	0.176	0.237	12.162	18	0.839	0.712	0.476

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	221	0.936	-0.473	-0.462	-3.979	3.481	
1	15	0.064	0.473	0.462	3.387	3.159	**KEY**
Omit	0						
Not Admin	1				-7.000	0.000	

Appendix E

IRT Item Parameter Calibration Report

CTS-2 All Perpetrator Items

Report created on 12/9/2013

Introduction

This report provides the results of the IRT item parameter calibration by the computer program Xcalibre Version 4.2.0.1 (Assessment Systems Corporation, 2012) for CTS Perp all items. The output is divided into four sections:

1. Specifications
2. E-M Algorithm
3. Summary statistics
4. Item-by-item results.

The statistical output is also recorded in a comma-separated value (CSV) file of the same name.

Specifications

This section records the input/output specifications and settings for historical purposes.

The Windows paths for the input files used in this analysis were:

C:\Users\Douglas Klinman\Documents\IRT_CTSc+e.csv
C:\Users\Douglas Klinman\Documents\IRT_CTSc+eCon.csv

The Windows paths for the output files produced by this analysis were:

CTS_P_All.rtf
CTS_P_All.csv
CTS_P_All Scores.csv

Table 1 presents the file specifications. Table 2 presents the IRT specifications used to perform the IRT item parameter calibration. Table 3 presents the flag specifications.

Table 1: File Specifications

Specification	Value	Specification	Value
Number of examinees	237	Total Items	36
Calibrated Items	18	Pretest Items	0
Excluded Items	18	Number of domains	1
Classic Data Header	No	Delimited input	Yes
Delimiter for input	Comma	Number of ID columns	N/A
ID begins in column	N/A	Responses begin in column	N/A
Omit character	O	Not Admin character	-
Save item parameters	No	Item parameter format	N/A
Save data matrix	No	Omit codes are	N/A
Not Admin codes are	N/A	Score Not Admin as omits	No
Plot the IRFs	Yes	Save the IRFs and IIFs	No
Produce the fit line	Yes	# Groups for Plot	20
Type of score groups	Equally sized	# Groups for Chi-square	20
Perform classification	No	Classify using	N/A
Two-group cutpoint	N/A	Low group label	N/A
High group label	N/A	Merge empty poly categories	N/A

Table 2: IRT Calibration Specifications

Specification	Value	Specification	Value
IRT Specification	Dichotomous	Model constant	1.7
Polytomous IRT Model	N/A	Dichotomous IRT Model	2-parameter
Center the boundary locations	No	Centered value	N/A
Floating priors	Yes	a parameter prior mean (sd)	1.000 (0.250)
b parameter prior mean (sd)	0.000 (1.000)	c parameter prior mean (sd)	0.250 (0.025)
Theta estimation method	MLE	Bayesian prior mean (sd)	N/A
Maximum E-M loops	60	Convergence criterion	0.001
Quadrature points	20	Center dich item parameters on	theta
Acceptable P range	0.00 to 1.00	Acceptable item-corr range	-1.00 to 1.00
Acceptable item mean range	0.00 to 15.00	Correct for spuriousness	Yes
Fit statistic critical alpha	0.050	Minimum a	0.05
Maximum a	6.00	Minimum b	-4.00
Maximum b	4.00	Minimum c	0.00
Maximum c	0.70	Minimum theta	-7.00
Maximum theta	7.00	Treat scored items as poly	No
Center poly parameters on theta	No	Test for DIF	No
Group status column	N/A	Ability levels for DIF Test	N/A
Group 1 code	N/A	Group 2 code	N/A
Group 1 label	N/A	Group 2 label	N/A
Exclude items with low N	No	Minimum valid N	N/A
Compute scaled scores	No	Mean (SD) of scaled scores	N/A
Minimum scaled score	N/A	Maximum scaled score	N/A
Save statistics output	Yes	Delimiter	Comma
Save scores output	Yes	Delimiter	Comma
Save test information output	Yes	Delimiter	Comma
Save item information output	Yes	Delimiter	Comma

Table 3: Flag Specifications

Specification	Value	Specification	Value
Low a Flag Bound	0.30	High a Flag Bound	4.00
Low b Flag Bound	-3.00	High b Flag Bound	3.00
Low c Flag Bound	0.00	High c Flag Bound	0.40
Key Flag	K	Fit Flag	F
Low a Flag	La	High a Flag	Ha
Low b Flag	Lb	High b Flag	Hb
Low c Flag	Lc	High c Flag	Hc

E-M Algorithm

Xcalibre uses the expectation-maximization approach to calibrate item parameters. The estimation process is iterative, and repeated in loops until the convergence criterion is satisfied. The following list presents the item with the largest parameter change after each loop, and the value of the change.

The number of loops needed is evidence regarding the fit of the data; if many loops are required, or

convergence is never reached, it means that the data does not fit well with the selected IRT model.

Maximum change after Loop 1 was 2.8899 for Item 13 for the b parameter
Maximum change after Loop 2 was -0.2991 for Item 16 for the b parameter
Maximum change after Loop 3 was 0.2727 for Item 13 for the b parameter
Maximum change after Loop 4 was -0.1938 for Item 13 for the b parameter
Maximum change after Loop 5 was 0.1032 for Item 13 for the b parameter
Maximum change after Loop 6 was -0.0438 for Item 13 for the b parameter
Maximum change after Loop 7 was -0.0074 for Item 2 for the b parameter
Maximum change after Loop 8 was -0.0062 for Item 2 for the b parameter
Maximum change after Loop 9 was -0.0053 for Item 2 for the b parameter
Maximum change after Loop 10 was -0.0046 for Item 2 for the b parameter
Maximum change after Loop 11 was -0.0039 for Item 2 for the b parameter
Maximum change after Loop 12 was -0.0034 for Item 2 for the b parameter
Maximum change after Loop 13 was -0.0030 for Item 2 for the b parameter
Maximum change after Loop 14 was -0.0026 for Item 2 for the b parameter
Maximum change after Loop 15 was -0.0023 for Item 2 for the b parameter
Maximum change after Loop 16 was -0.0020 for Item 2 for the b parameter
Maximum change after Loop 17 was -0.0017 for Item 2 for the b parameter
Maximum change after Loop 18 was -0.0015 for Item 2 for the b parameter
Maximum change after Loop 19 was -0.0013 for Item 2 for the b parameter
Maximum change after Loop 20 was -0.0012 for Item 2 for the b parameter
Maximum change after Loop 21 was -0.0010 for Item 2 for the b parameter
Maximum change after Loop 22 was -0.0009 for Item 2 for the b parameter

Summary statistics

Table 4 presents the summary statistics for the item parameters for all calibrated items. Table 5 summarizes the total scores for the full test for just the calibrated items. Table 6 summarizes the theta estimates for the full test. Table 7 provides the overall model fit chi-square(s) for the full test. Definitions of these statistics are found in the Xcalibre manual.

Table 4: Summary Statistics for All Calibrated Items

Parameter	Items	Mean	SD	Min	Max
a	18	0.773	0.109	0.614	1.052
b	18	1.723	1.273	-0.707	3.717

Table 5: Summary Statistics for the Total Scores

Test	Items	Alpha	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	18	0.890	1.722	2.993	2.528	0	0.00	0.0	2.00	18	2.00

Table 6: Summary Statistics for the Theta Estimates

Test	Examinees	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	237	0.000	0.644	0.973	-7.000	-7.000	-7.000	-0.446	7.000	6.554

Table 7: Overall Model Fit

Test	Items	Chi-square	df	p	-2LL
Full Test	18	324.115	324	0.488	1051

Table 8 presents the item control information and item status for each item

Table 8: Item Control and Item Status for All Items

Seq.	Item ID	Key	Options	Domain	Inclusion	Item Type	Status
1	Q1 MV P	1	2	1	Y	P	Included
2	Q3 MV P	1	2	1	Y	P	Included
3	Q7 MV P	1	2	1	Y	P	Included
4	Q25 MV P	1	2	1	Y	P	Included
5	Q27 MV P	1	2	1	Y	P	Included
6	Q6 MI P	1	2	1	Y	P	Included
7	Q34 MI P	1	2	1	Y	P	Included
8	Q9 SV P	1	2	1	Y	P	Included
9	Q13 SV P	1	2	1	Y	P	Included
10	Q17 SV P	1	2	1	Y	P	Included
11	Q19 SV P	1	2	1	Y	P	Included
12	Q23 SV P	1	2	1	Y	P	Included
13	Q31 SV P	1	2	1	Y	P	Included
14	Q35 SV P	1	2	1	Y	P	Included
15	Q12 SI P	1	2	1	Y	P	Included
16	Q16 SI P	1	2	1	Y	P	Included
17	Q22 SI P	1	2	1	Y	P	Included
18	Q30 SI P	1	2	1	Y	P	Included
19	Q2 MV V	1	2	1	N	P	Not Included
20	Q4 MV V	1	2	1	N	P	Not Included
21	Q8 MV V	1	2	1	N	P	Not Included
22	Q26 MV V	1	2	1	N	P	Not Included
23	Q28 MV V	1	2	1	N	P	Not Included
24	Q5 MI V	1	2	1	N	P	Not Included
25	Q33 MI V	1	2	1	N	P	Not Included
26	Q10 SV V	1	2	1	N	P	Not Included
27	Q14 SV V	1	2	1	N	P	Not Included
28	Q18 SV V	1	2	1	N	P	Not Included
29	Q20 SV V	1	2	1	N	P	Not Included
30	Q24 SV V	1	2	1	N	P	Not Included
31	Q32 SV V	1	2	1	N	P	Not Included
32	Q36 SV V	1	2	1	N	P	Not Included
33	Q11 SI V	1	2	1	N	P	Not Included
34	Q15 SI V	1	2	1	N	P	Not Included
35	Q21 SI V	1	2	1	N	P	Not Included
36	Q29 SI V	1	2	1	N	P	Not Included

Table 9 presents the classical statistics, the item parameters, and any flags for each calibrated item. The K flag indicates that the keyed alternative did not have the highest correlation with total score. The F flag indicates that the item fit statistic (z Resid for dichotomous / chi-square for polytomous) was significant, and the item did not fit the IRT model.The La, Lb, and Lc flags indicate that the a/b/c parameters were lower than the minimum acceptable value.The Ha, Hb, and Hc flags indicate that the a/b/c parameters were higher than the maximum acceptable value

Table 9: Item Parameters for All Calibrated Items

Seq.	Item ID	P	R	a	b	Flag(s)
1	Q1 MV P	0.194	0.600	0.813	0.248	
2	Q3 MV P	0.080	0.514	0.724	1.615	
3	Q7 MV P	0.316	0.612	1.052	-0.707	F
4	Q25 MV P	0.232	0.637	0.935	-0.119	
5	Q27 MV P	0.169	0.665	0.927	0.383	
6	Q6 MI P	0.114	0.584	0.743	1.095	
7	Q34 MI P	0.093	0.630	0.802	1.293	
8	Q9 SV P	0.042	0.497	0.673	2.456	
9	Q13 SV P	0.156	0.609	0.729	0.651	
10	Q17 SV P	0.038	0.472	0.682	2.552	
11	Q19 SV P	0.063	0.555	0.709	1.899	
12	Q23 SV P	0.046	0.608	0.840	2.049	
13	Q31 SV P	0.008	0.479	0.790	3.717	Hb
14	Q35 SV P	0.076	0.565	0.712	1.668	
15	Q12 SI P	0.017	0.450	0.717	3.285	Hb
16	Q16 SI P	0.030	0.377	0.614	3.004	Hb
17	Q22 SI P	0.025	0.545	0.771	2.782	
18	Q30 SI P	0.021	0.437	0.685	3.146	Hb

Figure 1 displays the distribution of the theta estimates for all calibrated items.
Table 10 displays the frequency distribution for the theta estimates.

Figure 1: Theta Estimates for All Calibrated Items

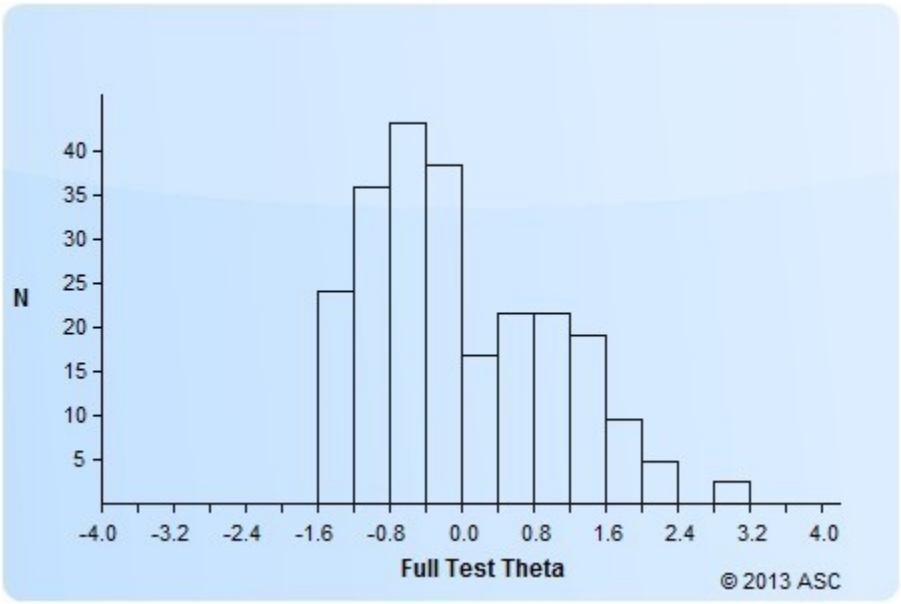


Table 10: Frequency Distribution for Full Test Theta

Range	Frequency
Below -4	136
-4.0 to -3.6	0
-3.6 to -3.2	0
-3.2 to -2.8	0
-2.8 to -2.4	0
-2.4 to -2.0	0
-2.0 to -1.6	0
-1.6 to -1.2	10
-1.2 to -0.8	15
-0.8 to -0.4	18
-0.4 to 0.0	16
0.0 to 0.4	7
0.4 to 0.8	9
0.8 to 1.2	9
1.2 to 1.6	8
1.6 to 2.0	4
2.0 to 2.4	2
2.4 to 2.8	0
2.8 to 3.2	1
3.2 to 3.6	0
3.6 to 4.0	0
Above +4	2

Figure 2 displays the distribution of the a parameters.
Table 11 displays the frequency distribution of the a parameters shown in Figure 2.

Figure 2: Histogram of the a Parameters

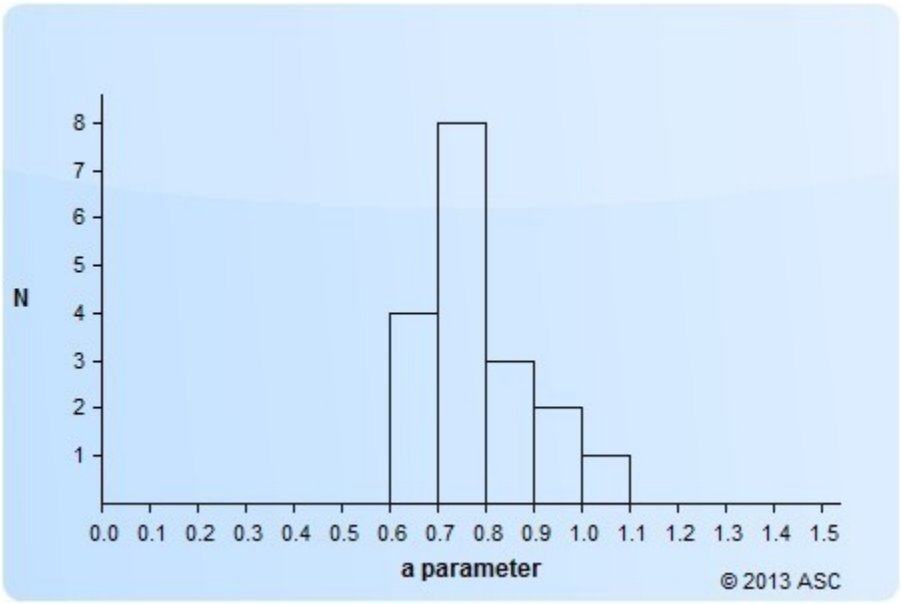


Table 11: Frequency Distribution for the a Parameters

Range	Frequency
0.00 to 0.10	0
0.10 to 0.20	0
0.20 to 0.30	0
0.30 to 0.40	0
0.40 to 0.50	0
0.50 to 0.60	0
0.60 to 0.70	4
0.70 to 0.80	8
0.80 to 0.90	3
0.90 to 1.00	2
1.00 to 1.10	1
1.10 to 1.20	0
1.20 to 1.30	0
1.30 to 1.40	0
1.40 to 1.50	0

Figure 3 displays the distribution of the b parameters.
Table 12 displays the frequency distribution of the b parameters shown in Figure 3.

Figure 3: Histogram of the b Parameters

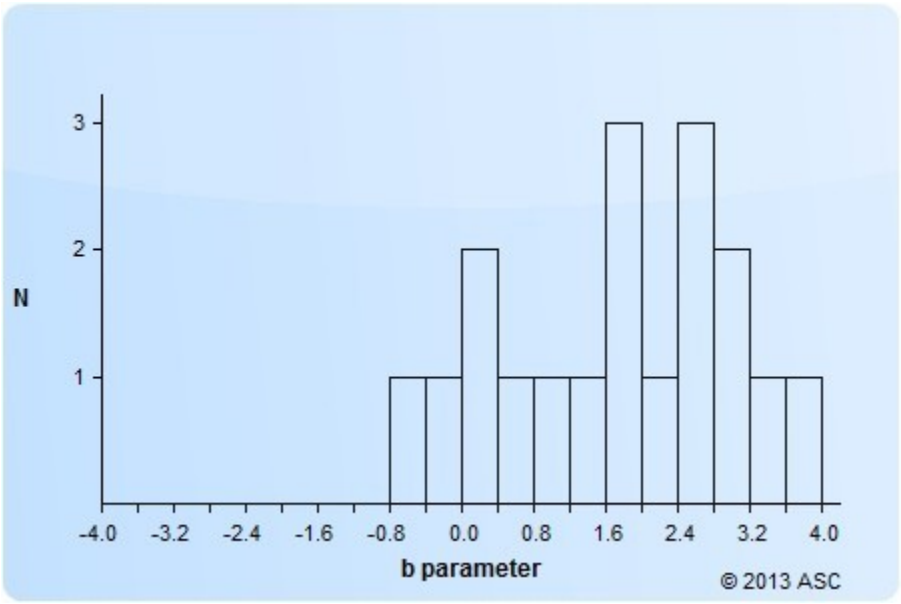


Table 12: Frequency Distribution for the b Parameters

Range	Frequency
-4.0 to -3.6	0
-3.6 to -3.2	0
-3.2 to -2.8	0
-2.8 to -2.4	0
-2.4 to -2.0	0
-2.0 to -1.6	0
-1.6 to -1.2	0
-1.2 to -0.8	0
-0.8 to -0.4	1
-0.4 to 0.0	1
0.0 to 0.4	2
0.4 to 0.8	1
0.8 to 1.2	1
1.2 to 1.6	1
1.6 to 2.0	3
2.0 to 2.4	1
2.4 to 2.8	3
2.8 to 3.2	2
3.2 to 3.6	1
3.6 to 4.0	1

Figure 4 displays the scatterplot of the b parameter (difficulty) by the a parameter (discrimination) for all calibrated items.

Figure 4: b Parameter by a Parameter

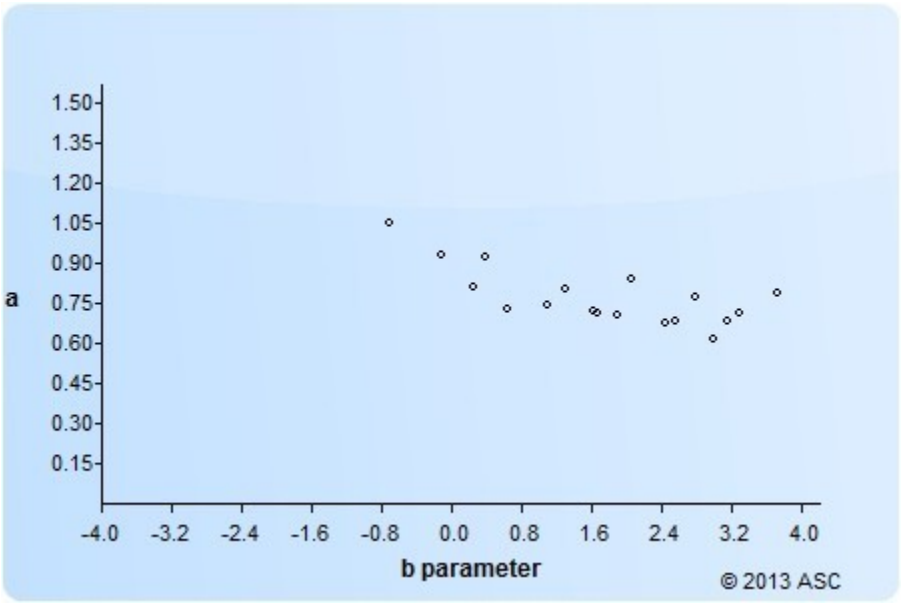


Figure 5 displays the joint distribution of the b parameter by Theta.

Figure 5: b parameter by Theta

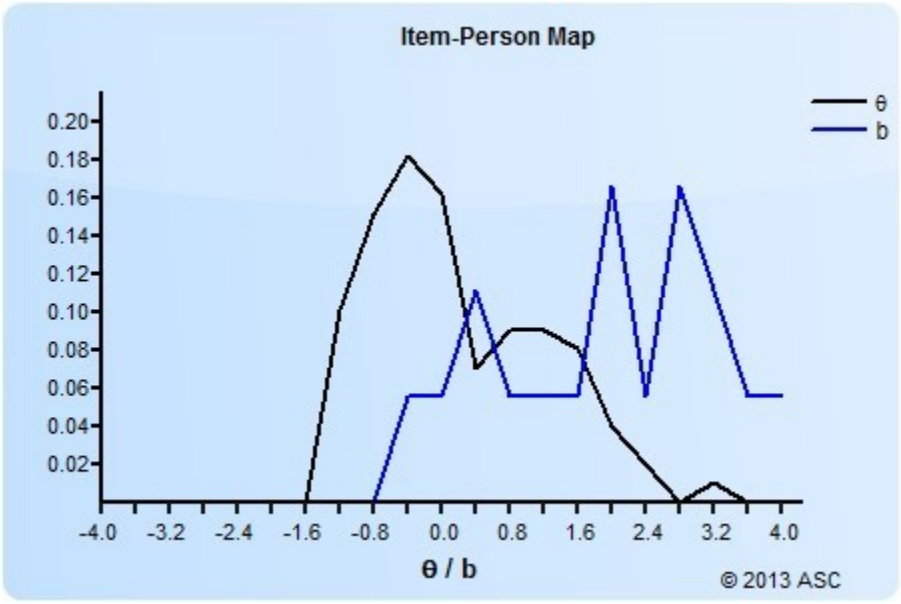


Figure 6 displays a graph of the Test Response Function (TRF) for all calibrated items. The TRF predicts the proportion or number of items that an examinee would answer correctly as a function of theta. The left Y-axis is in proportion correct units while the right Y-axis is in number-correct units.

Figure 6: Test Response Function

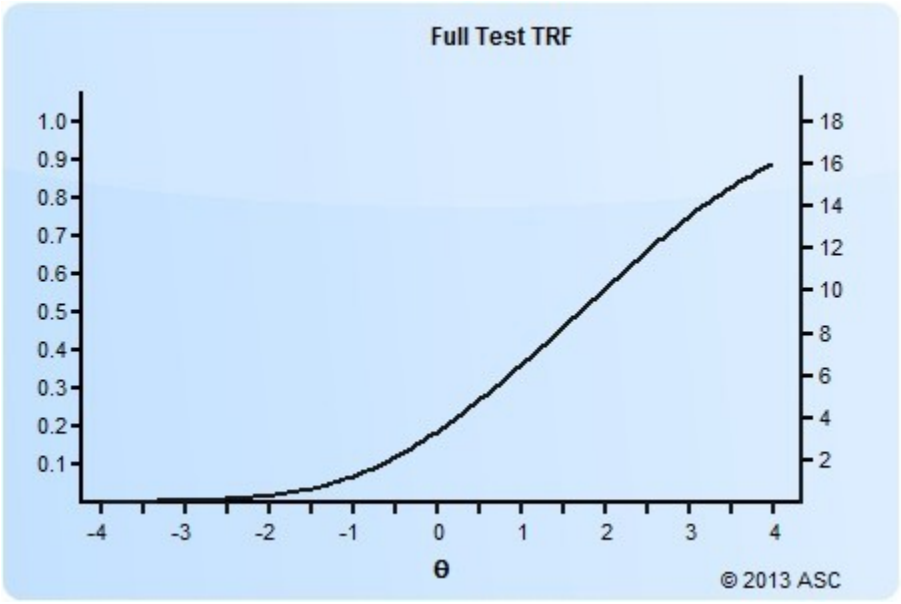


Figure 7 displays a graph of the Test Information Function for all calibrated items. The TIF is a graphical representation of how much information the test is providing at each level of theta. Maximum information was 4.668 at theta = 1.600.

Figure 7: Test Information Function

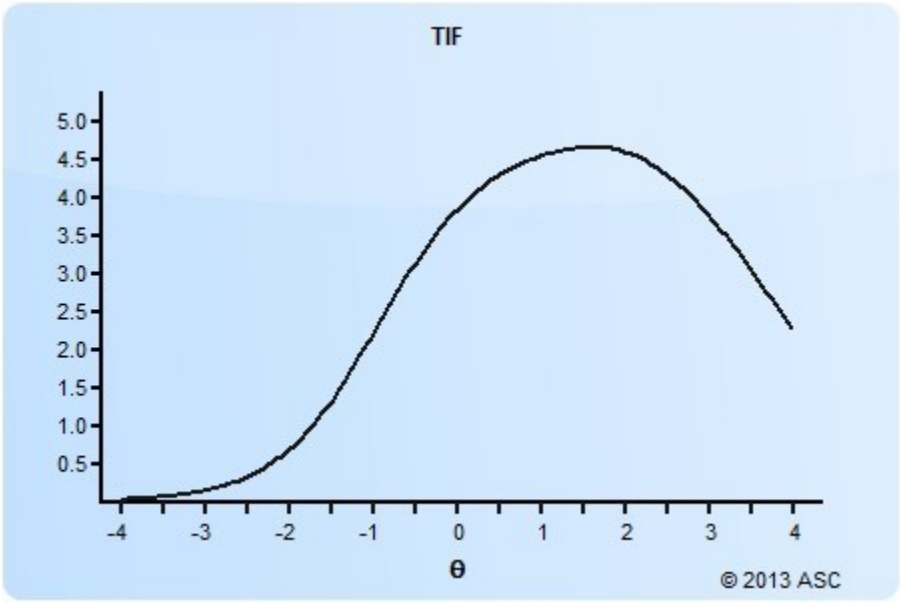
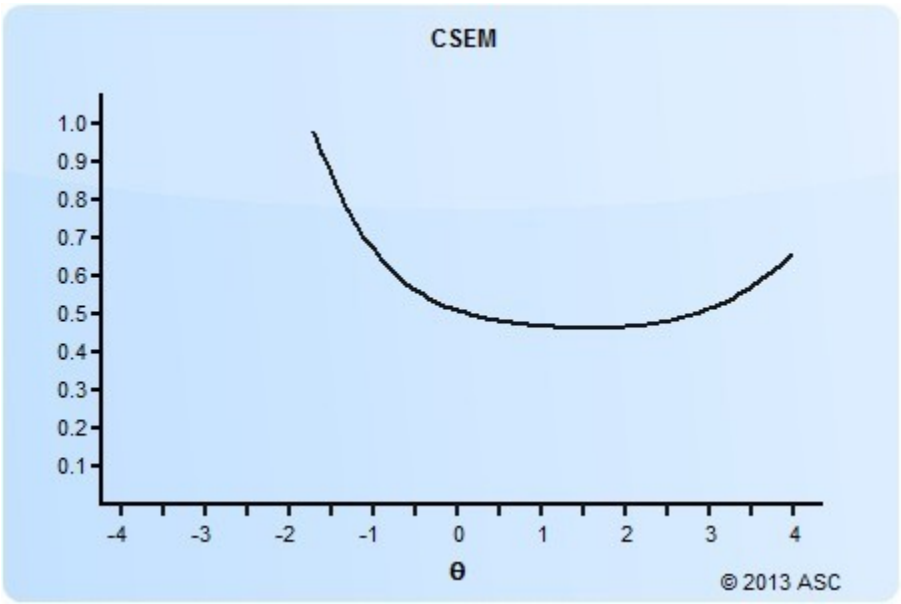


Figure 8 displays a graph of the Conditional Standard Error of Measurement (CSEM) Function. The CSEM is an inverted function of the TIF, and estimates the amount of error in theta estimation for each level of theta. The minimum CSEM was 0.463 at theta = 1.600.

Figure 8: CSEM Function



Item-by-item results

The following section presents the item-by-item results of the analysis. Each scored item has four tables and a plot of the item response function (IRF). The red line (fit line) represents the observed proportion correct conditional on theta. Large deviations of the red line from the IRF are suggestive of poor item fit. Thus, the fit line could be used to identify why items are not fitting the chosen IRT model.

There are four tables presented for each item.

1. Item information table: records the information supplied by the control file (or Classic Data Header) for this item.
2. Classical statistics table: classical statistics for the item.
3. IRT parameters table: item parameter estimates for the item.
4. Option/Category statistics: detailed statistics for each item, which helps diagnose issues in items with poor statistics.

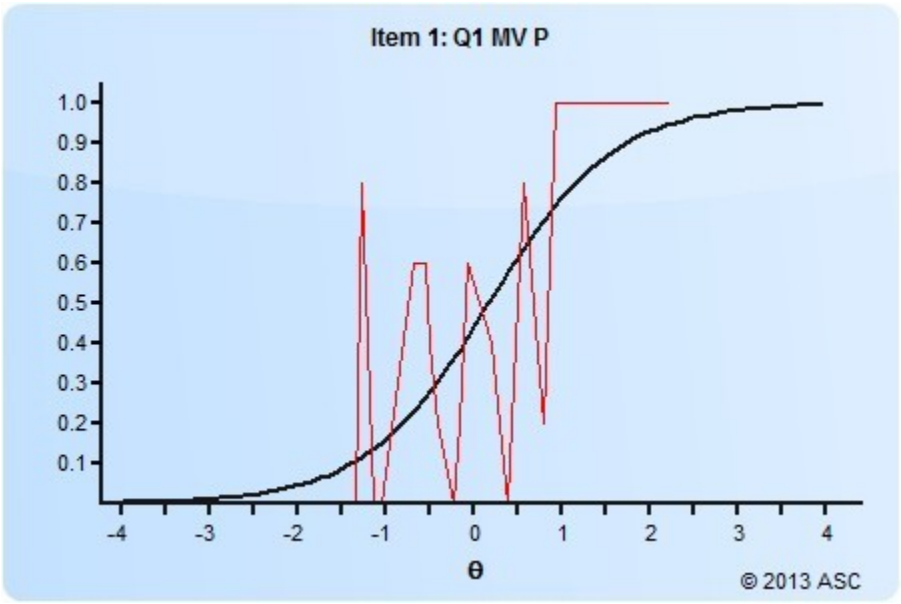
The classical statistics presents classical summary statistics for the item. For multiple choice items the P value and the point-biserial correlations are presented in the first three columns of the table. The P value is the proportion of examinees that answered an item in the keyed direction and ranges from 0 to 1. The S-Rpbis and T-Rpbis are the point-biserial correlations of an item with total score and theta, respectively. The Alpha w/o is Cronbach's alpha computed with the current item excluded. The item-total correlation is a measure of the discriminating power of the item and is related to the IRT discrimination parameter.

The IRT parameters table presents the IRT item parameters and the fit statistics. The latent trait theta is expressed on a standardized scale, so a one unit change equals a one standard deviation change. The "a" parameter indexes the discrimination of the item, as larger values for "a" will result in a greater slope of the IRF and indicate the item differentiates examinees well. The "b" parameter is the item difficulty parameter and equals the location on the theta continuum where the probability of a correct response equals .50. It follows that multiple choice items with more positive "b" parameters are more difficult for examinees, as a higher trait level is required to endorse the keyed response 50% of the time.

The standard errors (SE) for each item parameter estimate are also presented in the item parameter table. A large SE for an item parameter (compared to the other items) indicates that the item parameter was poorly estimated. The IRT standardized (z) residual is the last entry in the item parameter table. It indexes the fit of the data to the item response function. For dichotomous items, the p-value for rejecting the item as poor fit was computed using the z residual with the standard normal distribution as its sampling distribution. The chi-square fit statistic and its degrees of freedom are reported for each item. The chi-square fit statistic and its degrees of freedom are reported for each item.

The option statistics table presents statistics for each individual option (alternative). The key thing to examine in this portion of the table is that no distractors have a higher S-Rpbis or T-Rpbis than the correct answer. That indicates that higher scoring examinees are selecting the incorrect answer, which therefore might be arguably correct.

Item 1: I threw something at my partner that could hurt.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
1	Q1 MV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.194	0.600	0.645	0.882

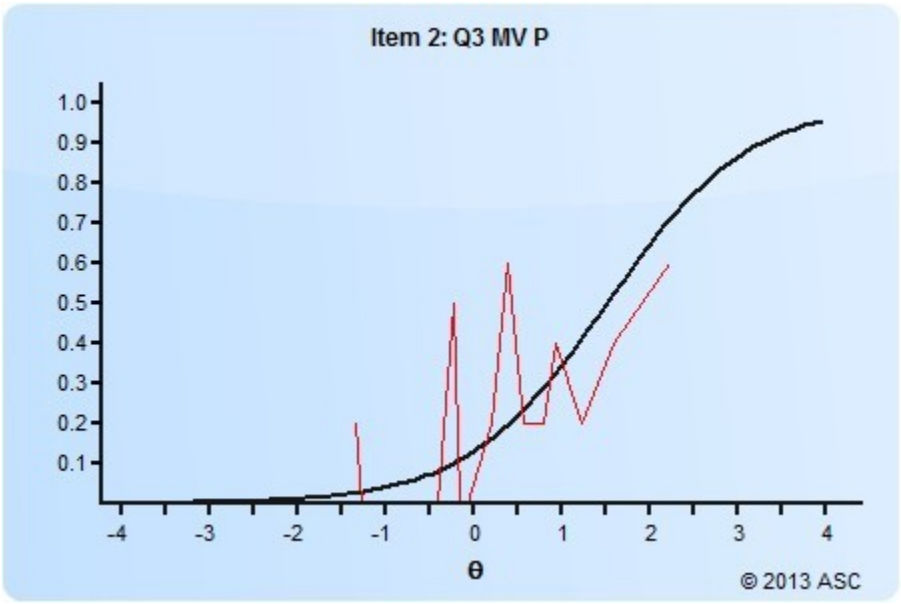
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.813	0.248	0.207	0.174	55.487	18	0.000	1.318	0.188

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	191	0.806	-0.600	-0.645	-5.111	3.000	
1	46	0.194	0.600	0.645	0.828	1.697	**KEY**
Omit	0						
Not Admin	0						

Item 2: I twisted my partner's arm or hair.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
2	Q3 MV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.080	0.514	0.443	0.884

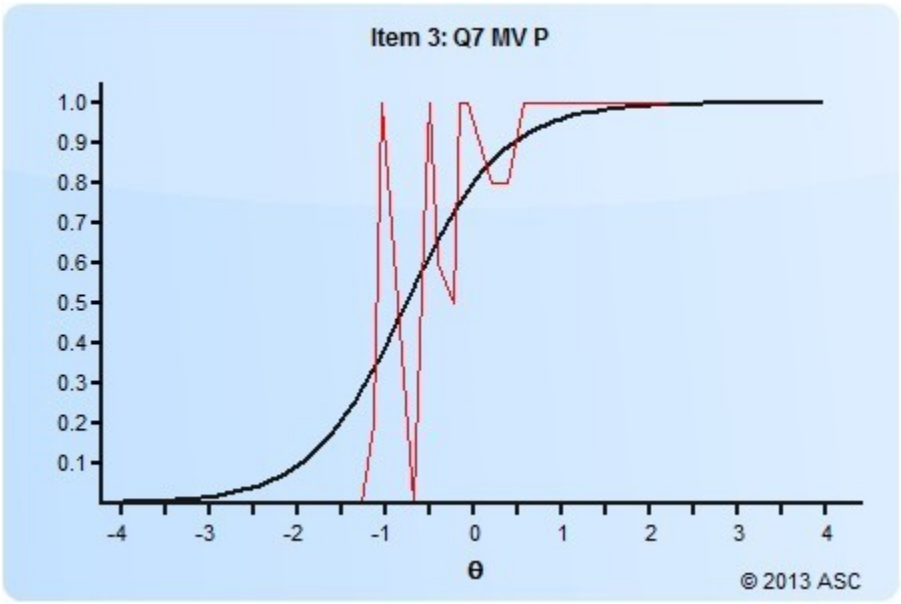
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.724	1.615	0.178	0.243	24.242	18	0.147	0.747	0.455

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	218	0.920	-0.514	-0.443	-4.435	3.357	
1	19	0.080	0.514	0.443	1.513	2.158	**KEY**
Omit	0						
Not Admin	0						

Item 3: I pushed or shoved my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
3	Q7 MV P	2PL	1	Yes	2	1	F

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.316	0.612	0.829	0.883

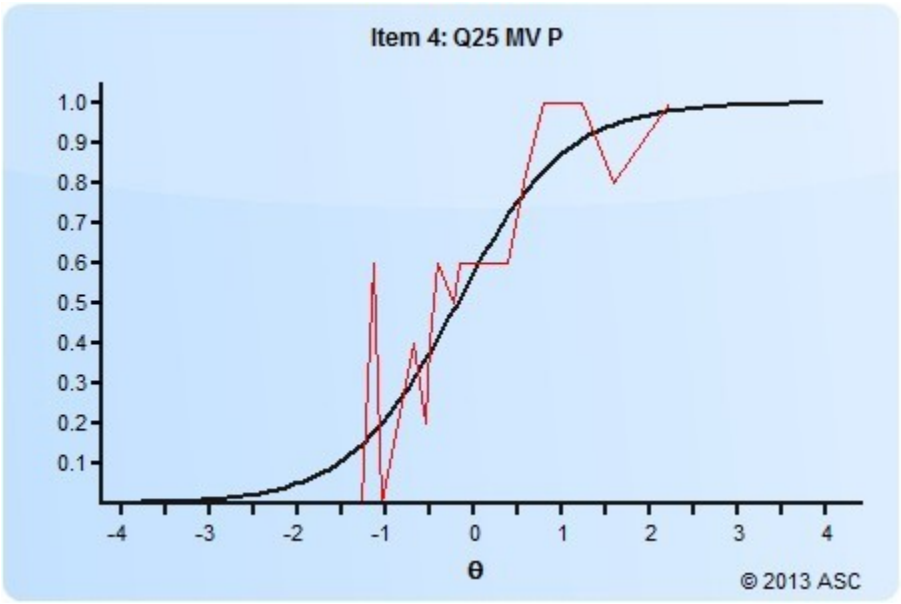
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
1.052	-0.707	0.217	0.143	37.307	18	0.005	2.604	0.009

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	162	0.684	-0.612	-0.829	-6.012	2.275	
1	75	0.316	0.612	0.829	0.480	1.442	**KEY**
Omit	0						
Not Admin	0						

Item 4: I grabbed my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
4	Q25 MV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.232	0.637	0.715	0.881

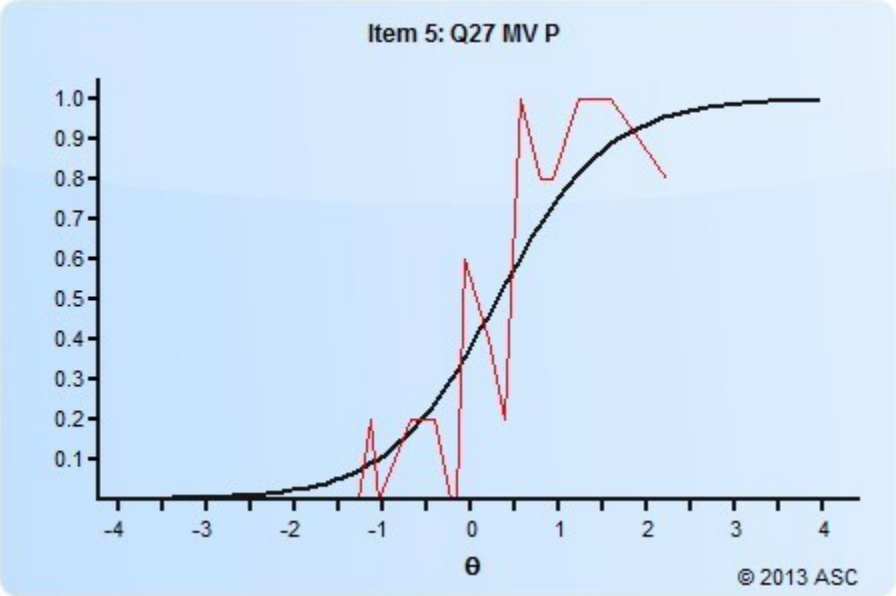
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.935	-0.119	0.210	0.153	17.106	18	0.516	1.812	0.070

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	182	0.768	-0.637	-0.715	-5.391	2.794	
1	55	0.232	0.637	0.715	0.784	1.519	**KEY**
Omit	0						
Not Admin	0						

Item 5: I slapped my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
5	Q27 MV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.169	0.665	0.630	0.878

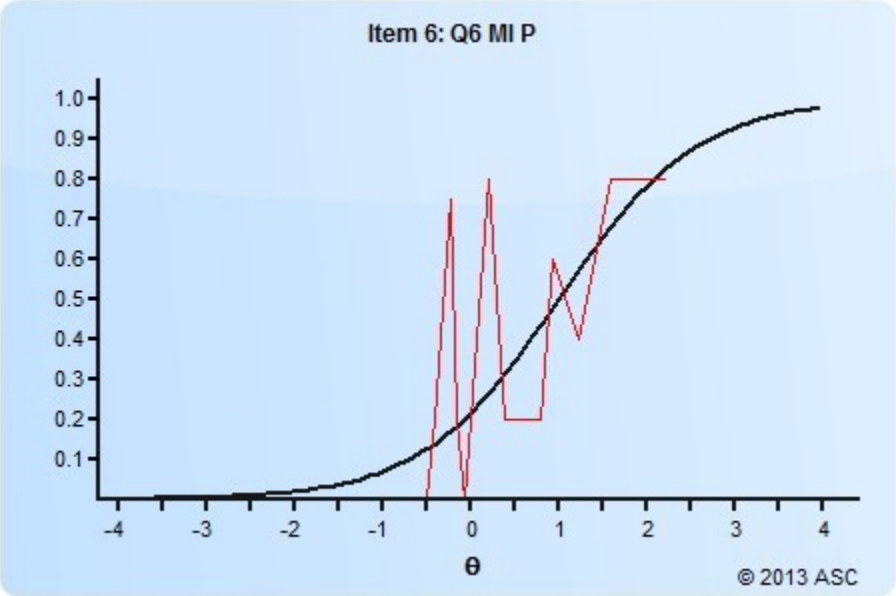
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.927	0.383	0.195	0.161	18.636	18	0.415	1.395	0.163

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	197	0.831	-0.665	-0.630	-4.993	3.028	
1	40	0.169	0.665	0.630	1.139	1.604	**KEY**
Omit	0						
Not Admin	0						

Item 6: My partner had a sprain, bruise, or small cut because of a fight with me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
6	Q6 MI P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.114	0.584	0.521	0.882

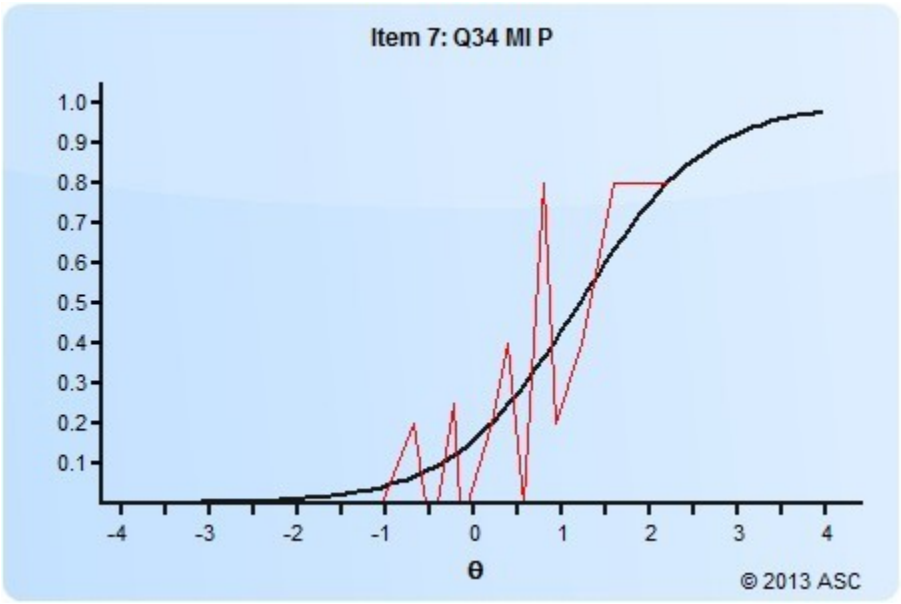
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.743	1.095	0.187	0.211	26.034	18	0.099	1.121	0.262

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	210	0.886	-0.584	-0.521	-4.639	3.248	
1	27	0.114	0.584	0.521	1.339	1.842	**KEY**
Omit	0						
Not Admin	0						

Item 7: My partner still felt physical pain the next day because of a fight we had.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
7	Q34 MI P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.093	0.630	0.494	0.880

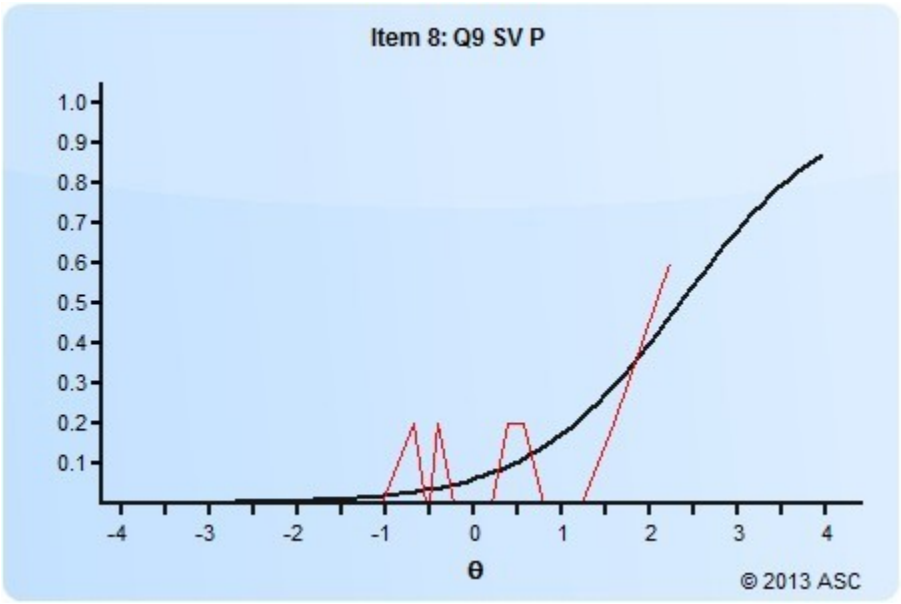
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.802	1.293	0.183	0.211	14.929	18	0.667	0.945	0.345

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	215	0.907	-0.630	-0.494	-4.534	3.279	
1	22	0.093	0.630	0.494	1.673	1.918	**KEY**
Omit	0						
Not Admin	0						

Item 8: I used a knife or gun on my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
8	Q9 SV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.042	0.497	0.360	0.885

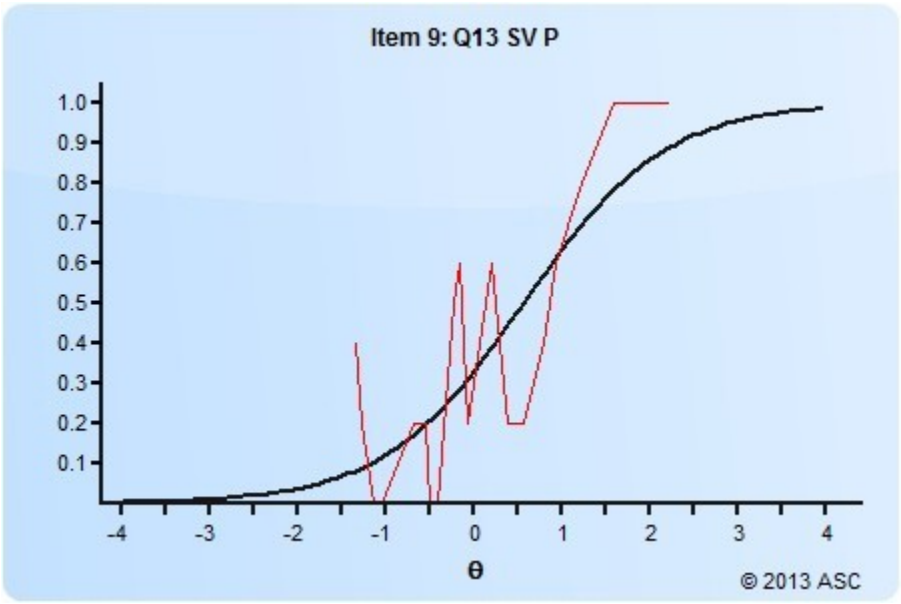
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.673	2.456	0.174	0.327	15.563	18	0.623	0.487	0.626

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	227	0.958	-0.497	-0.360	-4.233	3.439	
1	10	0.042	0.497	0.360	2.288	2.747	**KEY**
Omit	0						
Not Admin	0						

Item 9: I punched or hit my partner with something that could hurt.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
9	Q13 SV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.156	0.609	0.589	0.881

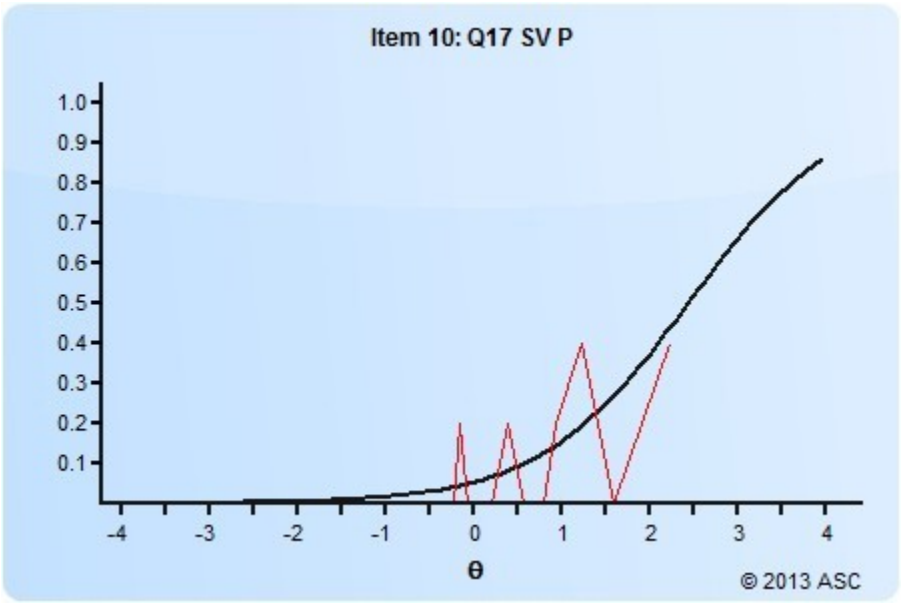
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.729	0.651	0.200	0.198	23.274	18	0.180	1.238	0.216

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	200	0.844	-0.609	-0.589	-4.881	3.123	
1	37	0.156	0.609	0.589	1.030	1.790	**KEY**
Omit	0						
Not Admin	0						

Item 10: I choked my partner



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
10	Q17 SV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.038	0.472	0.349	0.886

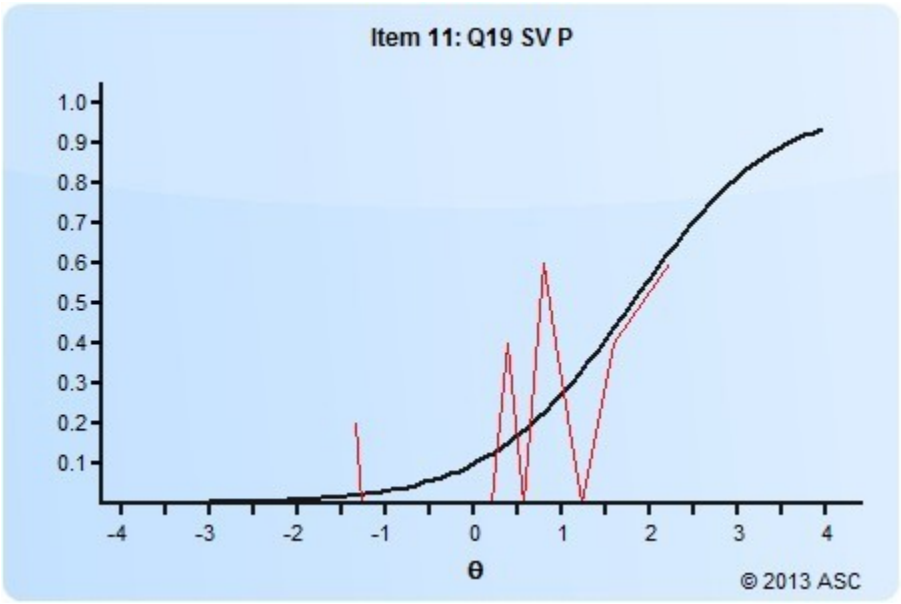
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.682	2.552	0.175	0.338	10.047	18	0.930	0.619	0.536

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	228	0.962	-0.472	-0.349	-4.211	3.453	
1	9	0.038	0.472	0.349	2.452	2.676	**KEY**
Omit	0						
Not Admin	0						

Item 11: I slammed my partner against the wall.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
11	Q19 SV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.063	0.555	0.421	0.883

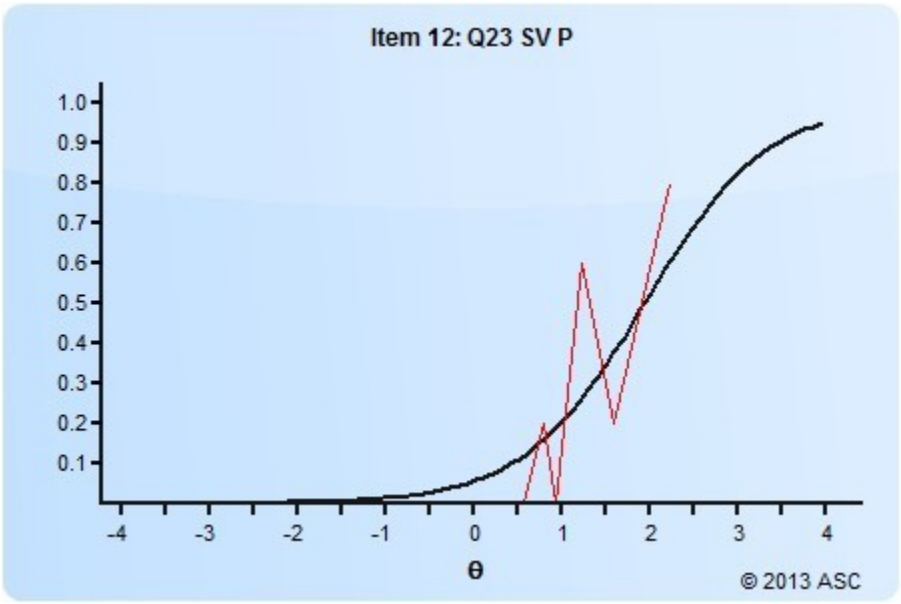
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.709	1.899	0.176	0.267	14.811	18	0.675	0.783	0.434

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	222	0.937	-0.555	-0.421	-4.356	3.375	
1	15	0.063	0.555	0.421	1.939	2.286	**KEY**
Omit	0						
Not Admin	0						

Item 12: I beat up my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
12	Q23 SV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.046	0.608	0.402	0.882

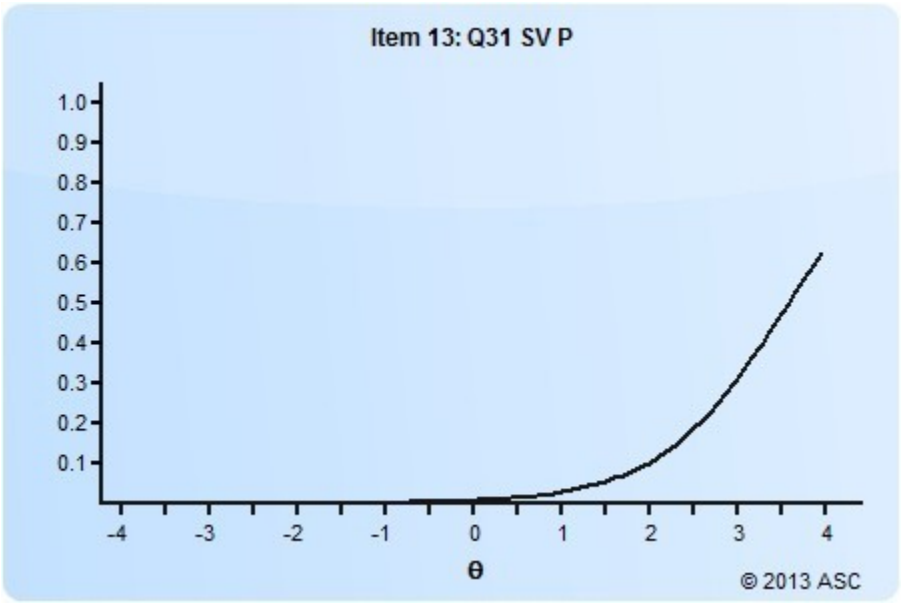
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.840	2.049	0.184	0.263	8.691	18	0.966	0.706	0.480

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	226	0.954	-0.608	-0.402	-4.281	3.393	
1	11	0.046	0.608	0.402	2.677	2.228	**KEY**
Omit	0						
Not Admin	0						

Item 13: I burned or scalded my partner on purpose.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
13	Q31 SV P	2PL	1	Yes	2	1	Hb

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.008	0.479	0.277	0.888

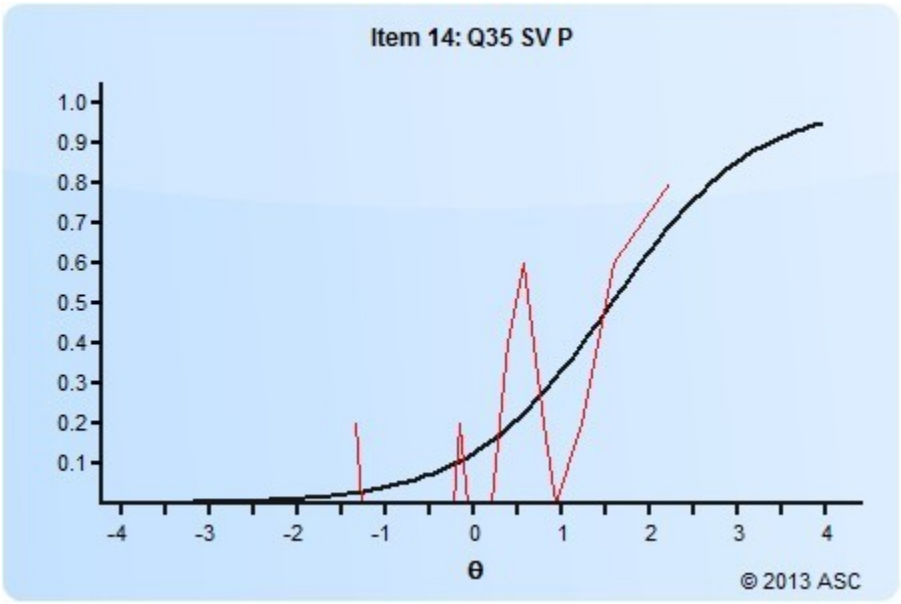
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.790	3.717	0.232	0.588	1.430	18	1.000	0.357	0.721

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	235	0.992	-0.479	-0.277	-4.051	3.524	
1	2	0.008	0.479	0.277	7.000	0.000	**KEY**
Omit	0						
Not Admin	0						

Item 14: I kicked my partner.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
14	Q35 SV P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.076	0.565	0.448	0.882

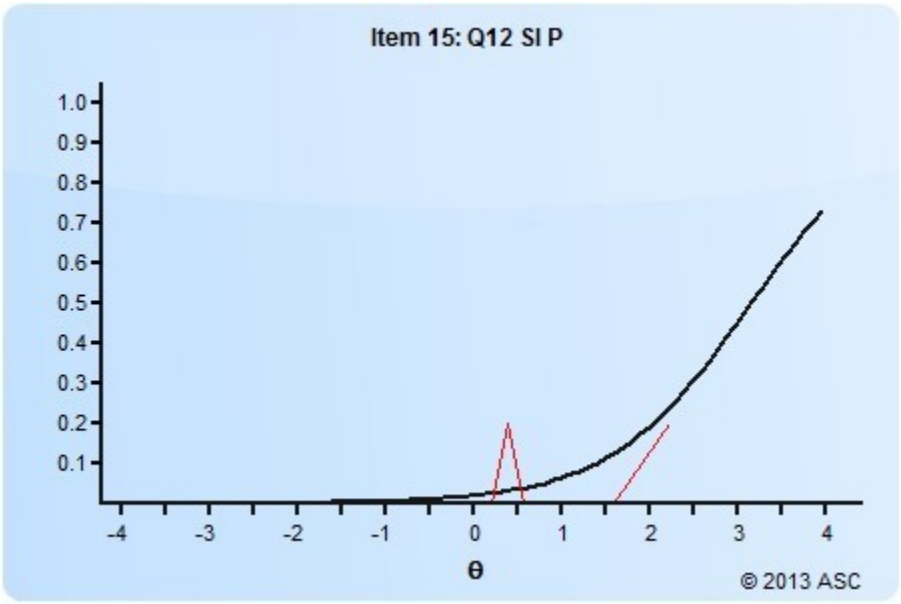
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.712	1.668	0.178	0.249	20.161	18	0.324	0.841	0.400

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	219	0.924	-0.565	-0.448	-4.426	3.344	
1	18	0.076	0.565	0.448	1.736	2.148	**KEY**
Omit	0						
Not Admin	0						

Item 15: My partner passed out from being hit on the head in a fight with me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
15	Q12 SI P	2PL	1	Yes	2	1	Hb

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.017	0.450	0.289	0.888

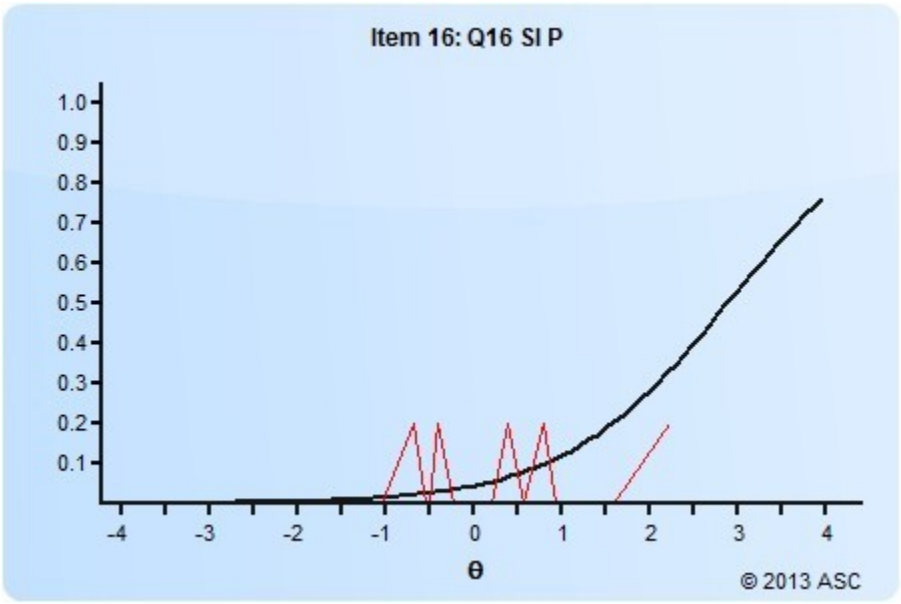
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.717	3.285	0.195	0.459	7.084	18	0.989	0.404	0.686

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	233	0.983	-0.450	-0.289	-4.096	3.504	
1	4	0.017	0.450	0.289	4.089	3.420	**KEY**
Omit	0						
Not Admin	0						

Item 16: My partner went to a doctor because of a fight with me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
16	Q16 SI P	2PL	1	Yes	2	1	Hb

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.030	0.377	0.303	0.888

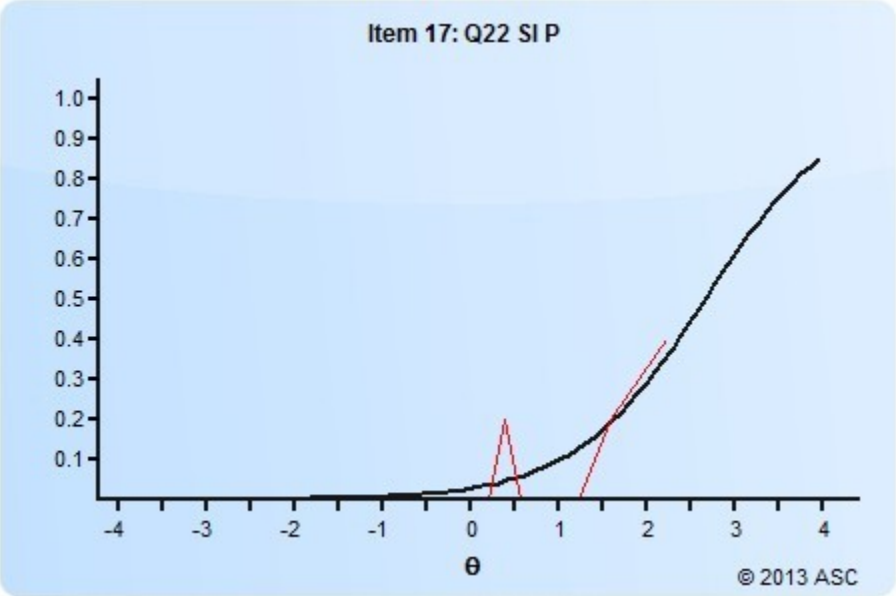
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.614	3.004	0.173	0.407	19.714	18	0.349	0.443	0.658

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	230	0.970	-0.377	-0.303	-4.151	3.492	
1	7	0.030	0.377	0.303	2.374	3.297	**KEY**
Omit	0						
Not Admin	0						

Item 17: My partner needed to see a doctor because of a fight with me, but didn't.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
17	Q22 SI P	2PL	1	Yes	2	1	

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.025	0.545	0.333	0.885

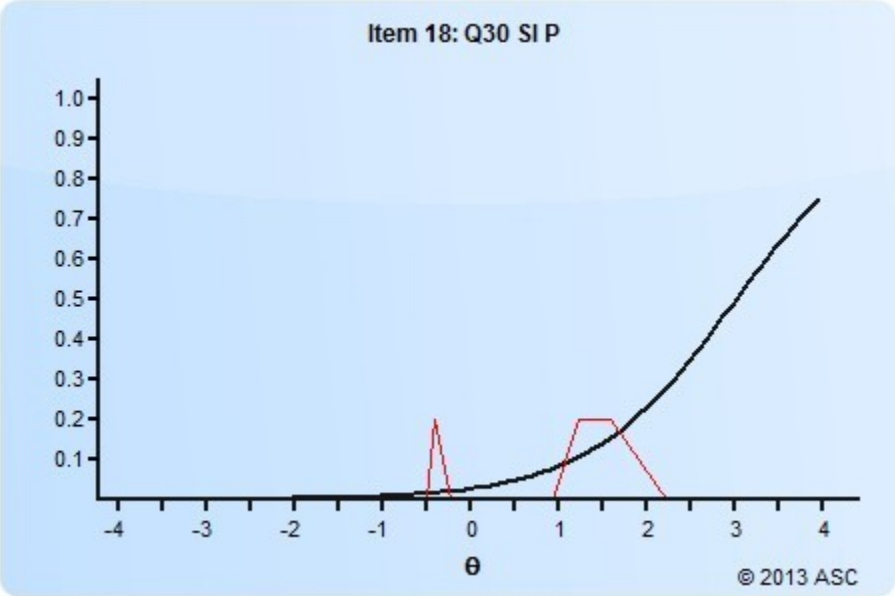
IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.771	2.782	0.190	0.364	5.343	18	0.998	0.482	0.630

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	231	0.975	-0.545	-0.333	-4.153	3.464	
1	6	0.025	0.545	0.333	3.570	2.790	**KEY**
Omit	0						
Not Admin	0						

Item 18: My partner had a broken bone from a fight with me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
18	Q30 SI P	2PL	1	Yes	2	1	Hb

Classical statistics

N	*P*	*S-Rpbis*	*T-Rpbis*	*Alpha w/o*
237	0.021	0.437	0.296	0.888

IRT parameters

a	*b*	*a SE*	*b SE*	*Chi-sq*	*df*	*p*	*z Resid*	*p*
0.685	3.146	0.186	0.429	4.255	18	1.000	0.376	0.707

Option statistics

Option	*N*	*Prop.*	*S-Rpbis*	*T-Rpbis*	*Mean*	*SD*	
0	232	0.979	-0.437	-0.296	-4.116	3.497	
1	5	0.021	0.437	0.296	3.401	3.388	**KEY**
Omit	0						
Not Admin	0						

Appendix F

IRT Item Parameter Calibration Report

PSI-PCDI all Items

Report created on 12/14/2013

Introduction

This report provides the results of the IRT item parameter calibration by the computer program Xcalibre Version 4.2.0.1 (Assessment Systems Corporation, 2012) for PSI-PCDI all Items. The output is divided into four sections:

1. Specifications
2. E-M Algorithm
3. Summary statistics
4. Item-by-item results.

The statistical output is also recorded in a comma-separated value (CSV) file of the same name.

Specifications

This section records the input/output specifications and settings for historical purposes.

The Windows paths for the input files used in this analysis were:

C:\Users\Douglas Klinman\Documents\IRT Models\PSI_PCDIex.csv
C:\Users\Douglas Klinman\Documents\IRT Models\PSIDC_PCDcon.csv

The Windows paths for the output files produced by this analysis were:

PSIPCDall.rtf
PSIPCDall.csv
PSIPCDall Scores.csv

Table 1 presents the file specifications. Table 2 presents the IRT specifications used to perform the IRT item parameter calibration. Table 3 presents the flag specifications.

Table 1: File Specifications

Specification	Value	Specification	Value
Number of examinees	240	Total Items	12
Calibrated Items	12	Pretest Items	0
Excluded Items	0	Number of domains	1
Classic Data Header	No	Delimited input	Yes
Delimiter for input	Comma	Number of ID columns	N/A
ID begins in column	N/A	Responses begin in column	N/A
Omit character	O	Not Admin character	-
Save item parameters	No	Item parameter format	N/A
Save data matrix	No	Omit codes are	N/A
Not Admin codes are	N/A	Score Not Admin as omits	No
Plot the IRFs	Yes	Save the IRFs and IIFs	No
Produce the fit line	No	# Groups for Plot	N/A
Type of score groups	N/A	# Groups for Chi-square	20
Perform classification	No	Classify using	N/A
Two-group cutpoint	N/A	Low group label	N/A
High group label	N/A	Merge empty poly categories	No

Table 2: IRT Calibration Specifications

Specification	Value	Specification	Value
IRT Specification	Polytomous	Model constant	1.7
Polytomous IRT Model	SGRM	Dichotomous IRT Model	N/A
Center the boundary locations	No	Centered value	N/A
Floating priors	Yes	a parameter prior mean (sd)	1.000 (0.250)
b parameter prior mean (sd)	0.000 (1.000)	c parameter prior mean (sd)	0.250 (0.025)
Theta estimation method	MLE	Bayesian prior mean (sd)	N/A
Maximum E-M loops	60	Convergence criterion	0.001
Quadrature points	20	Center dich item parameters on	N/A
Acceptable P range	0.00 to 1.00	Acceptable item-corr range	-1.00 to 1.00
Acceptable item mean range	0.00 to 15.00	Correct for spuriousness	Yes
Fit statistic critical alpha	0.050	Minimum a	0.05
Maximum a	6.00	Minimum b	-4.00
Maximum b	4.00	Minimum c	0.00
Maximum c	0.70	Minimum theta	-7.00
Maximum theta	7.00	Treat scored items as poly	Yes
Center poly parameters on theta	Yes	Test for DIF	No
Group status column	N/A	Ability levels for DIF Test	N/A
Group 1 code	N/A	Group 2 code	N/A
Group 1 label	N/A	Group 2 label	N/A
Exclude items with low N	No	Minimum valid N	N/A
Compute scaled scores	No	Mean (SD) of scaled scores	N/A
Minimum scaled score	N/A	Maximum scaled score	N/A
Save statistics output	Yes	Delimiter	Comma
Save scores output	Yes	Delimiter	Comma
Save test information output	Yes	Delimiter	Comma
Save item information output	Yes	Delimiter	Comma

Table 3: Flag Specifications

Specification	Value	Specification	Value
Low a Flag Bound	0.30	High a Flag Bound	4.00
Low b Flag Bound	-3.00	High b Flag Bound	3.00
Low c Flag Bound	0.00	High c Flag Bound	0.40
Key Flag	K	Fit Flag	F
Low a Flag	La	High a Flag	Ha
Low b Flag	Lb	High b Flag	Hb
Low c Flag	Lc	High c Flag	Hc

E-M Algorithm

Xcalibre uses the expectation-maximization approach to calibrate item parameters. The estimation process is iterative, and repeated in loops until the convergence criterion is satisfied. The following list presents the item with the largest parameter change after each loop, and the value of the change.

The number of loops needed is evidence regarding the fit of the data; if many loops are required, or

convergence is never reached, it means that the data does not fit well with the selected IRT model.

Maximum change after Loop 1 was -4.6048 for Item 5 for Category parameter 4
Maximum change after Loop 2 was -0.7055 for Item 7 for Category parameter 4
Maximum change after Loop 3 was -0.3103 for Item 7 for Category parameter 4
Maximum change after Loop 4 was -0.1934 for Item 5 for Category parameter 4
Maximum change after Loop 5 was -0.1415 for Item 5 for Category parameter 4
Maximum change after Loop 6 was -0.1123 for Item 5 for Category parameter 4
Maximum change after Loop 7 was -0.0930 for Item 5 for Category parameter 4
Maximum change after Loop 8 was -0.0793 for Item 3 for Category parameter 4
Maximum change after Loop 9 was -0.0700 for Item 3 for Category parameter 4
Maximum change after Loop 10 was -0.0623 for Item 3 for Category parameter 4
Maximum change after Loop 11 was -0.0558 for Item 3 for Category parameter 4
Maximum change after Loop 12 was -0.0502 for Item 3 for Category parameter 4
Maximum change after Loop 13 was -0.0454 for Item 3 for Category parameter 4
Maximum change after Loop 14 was -0.0411 for Item 3 for Category parameter 4
Maximum change after Loop 15 was -0.0373 for Item 3 for Category parameter 4
Maximum change after Loop 16 was -0.0340 for Item 3 for Category parameter 4
Maximum change after Loop 17 was -0.0310 for Item 3 for Category parameter 4
Maximum change after Loop 18 was -0.0283 for Item 3 for Category parameter 4
Maximum change after Loop 19 was -0.0258 for Item 3 for Category parameter 4
Maximum change after Loop 20 was -0.0236 for Item 3 for Category parameter 4
Maximum change after Loop 21 was -0.0216 for Item 3 for Category parameter 4
Maximum change after Loop 22 was -0.0197 for Item 3 for Category parameter 4
Maximum change after Loop 23 was -0.0177 for Item 3 for Category parameter 4
Maximum change after Loop 24 was -0.0159 for Item 3 for Category parameter 4
Maximum change after Loop 25 was -0.0142 for Item 3 for Category parameter 4
Maximum change after Loop 26 was -0.0128 for Item 3 for Category parameter 4
Maximum change after Loop 27 was -0.0115 for Item 3 for Category parameter 4
Maximum change after Loop 28 was -0.0104 for Item 3 for Category parameter 4
Maximum change after Loop 29 was -0.0094 for Item 3 for Category parameter 4
Maximum change after Loop 30 was -0.0086 for Item 3 for Category parameter 4
Maximum change after Loop 31 was -0.0077 for Item 3 for Category parameter 4
Maximum change after Loop 32 was -0.0070 for Item 3 for Category parameter 4
Maximum change after Loop 33 was -0.0063 for Item 3 for Category parameter 4
Maximum change after Loop 34 was -0.0057 for Item 3 for Category parameter 4
Maximum change after Loop 35 was -0.0051 for Item 3 for Category parameter 4
Maximum change after Loop 36 was -0.0046 for Item 3 for Category parameter 4
Maximum change after Loop 37 was -0.0042 for Item 3 for Category parameter 4
Maximum change after Loop 38 was -0.0038 for Item 3 for Category parameter 4
Maximum change after Loop 39 was -0.0034 for Item 3 for Category parameter 4
Maximum change after Loop 40 was -0.0031 for Item 3 for Category parameter 4
Maximum change after Loop 41 was -0.0028 for Item 3 for Category parameter 4
Maximum change after Loop 42 was -0.0025 for Item 3 for Category parameter 4
Maximum change after Loop 43 was -0.0023 for Item 3 for Category parameter 4
Maximum change after Loop 44 was -0.0021 for Item 3 for Category parameter 4
Maximum change after Loop 45 was -0.0019 for Item 3 for Category parameter 4
Maximum change after Loop 46 was -0.0017 for Item 3 for Category parameter 4
Maximum change after Loop 47 was -0.0016 for Item 3 for Category parameter 4
Maximum change after Loop 48 was -0.0014 for Item 3 for Category parameter 4
Maximum change after Loop 49 was -0.0013 for Item 3 for Category parameter 4
Maximum change after Loop 50 was -0.0012 for Item 3 for Category parameter 4
Maximum change after Loop 51 was -0.0010 for Item 3 for Category parameter 4
Maximum change after Loop 52 was -0.0010 for Item 7 for Category parameter 4

Summary statistics

Table 4 presents the summary statistics for the item parameters for all calibrated items. Table 5 summarizes the total scores for the full test for just the calibrated items. Table 6 summarizes the theta estimates for the full test. Table 7 provides the overall model fit chi-square(s) for the full test. Definitions of these statistics are found in the Xcalibre manual.

Table 4: Summary Statistics for All Calibrated Items

Parameter	Items	Mean	SD	Min	Max
a	12	1.335	0.399	0.387	1.728

Table 5: Summary Statistics for the Total Scores

Test	Items	Alpha	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	12	0.876	23.321	8.670	1.002	12	16.25	22.0	28.00	54	11.75

Table 6: Summary Statistics for the Theta Estimates

Test	Examinees	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	240	0.000	0.985	-0.252	-7.000	-0.622	0.111	0.644	3.223	1.266

Table 7: Overall Model Fit

Test	Items	Chi-square	df	p	-2LL
Full Test	12	957.146	900	0.091	5015

Table 8 presents the item control information and item status for each item

Table 8: Item Control and Item Status for All Items

Seq.	Item ID	Key	Options	Domain	Inclusion	Item Type	Status
1	PSI Item 13	+	5	1	Y	R	Included
2	PSI Item 14	+	5	1	Y	R	Included
3	PSI Item 15	+	5	1	Y	R	Included
4	PSI Item 16	+	5	1	Y	R	Included
5	PSI Item 17	+	5	1	Y	R	Included
6	PSI Item 18	+	5	1	Y	R	Included
7	PSI Item 19	+	5	1	Y	R	Included
8	PSI Item 20	+	5	1	Y	R	Included
9	PSI Item 21	+	5	1	Y	R	Included
10	PSI Item 22	+	5	1	Y	R	Included
11	PSI Item 23	+	5	1	Y	R	Included
12	PSI Item 24	+	5	1	Y	R	Included

Table 9 presents the classical statistics, the item parameters, and any flags for each calibrated item. The K flag indicates that the keyed alternative did not have the highest correlation with total score. The F flag indicates that the item fit statistic (z Resid for dichotomous / chi-square for polytomous) was significant, and the item did not fit the IRT model.The La, Lb, and Lc flags indicate that the a/b/c parameters were lower than the minimum acceptable value.The Ha, Hb, and Hc flags indicate that the a/b/c parameters were higher than the maximum acceptable value

Table 9: Item Parameters for All Calibrated Items

Seq.	Item ID	Item Mean	R	a	b	Flag(s)
1	PSI Item 13	1.650	0.502	1.527	0.18, 1.39, 1.56, 2.43	
2	PSI Item 14	1.975	0.541	1.595	0.03, 0.78, 1.04, 1.96	
3	PSI Item 15	1.700	0.508	1.728	0.26, 1.18, 1.32, 2.11	
4	PSI Item 16	2.154	0.580	1.612	-0.23, 0.71, 0.91, 1.61	
5	PSI Item 17	1.442	0.470	1.572	0.51, 1.69, 1.93, 2.53	
6	PSI Item 18	2.142	0.441	1.003	-0.11, 0.77, 1.11, 2.01	F
7	PSI Item 19	1.575	0.527	1.700	0.30, 1.35, 1.70, 2.35	
8	PSI Item 20	1.846	0.541	1.532	-0.03, 1.06, 1.30, 2.37	
9	PSI Item 21	2.258	0.500	1.017	-0.48, 0.69, 1.06, 2.10	
10	PSI Item 22	2.421	0.274	0.387	-2.05, -0.12, 3.29, 6.40	F
11	PSI Item 23	1.704	0.467	1.284	0.29, 1.18, 1.57, 2.34	
12	PSI Item 24	2.454	0.518	1.062	-0.44, 0.44, 0.66, 1.72	F

Figure 1 displays the distribution of the theta estimates for all calibrated items.
Table 10 displays the frequency distribution for the theta estimates.

Figure 1: Theta Estimates for All Calibrated Items

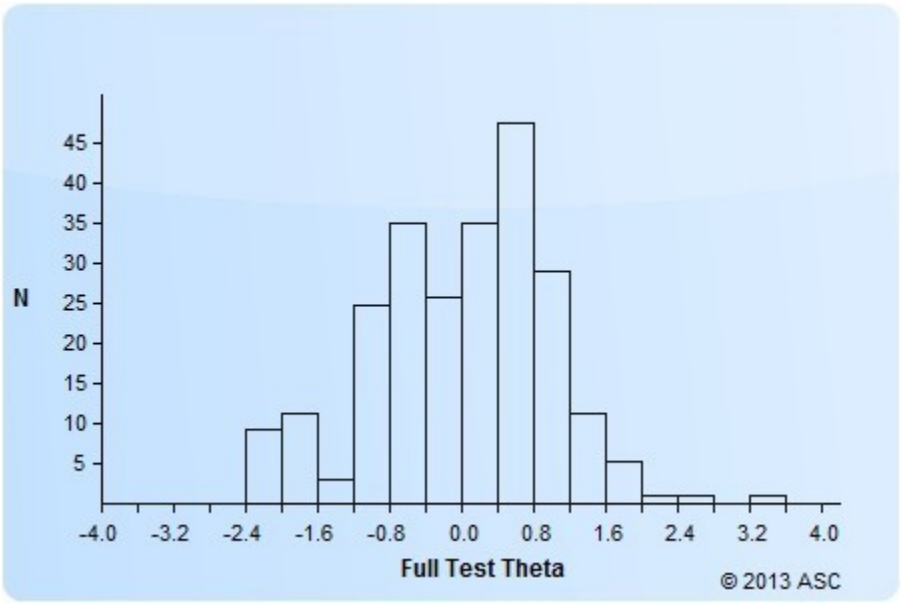


Table 10: Frequency Distribution for Full Test Theta

Range	Frequency
Below -4	7
-4.0 to -3.6	0
-3.6 to -3.2	0
-3.2 to -2.8	0
-2.8 to -2.4	0
-2.4 to -2.0	9
-2.0 to -1.6	11
-1.6 to -1.2	3
-1.2 to -0.8	24
-0.8 to -0.4	34
-0.4 to 0.0	25
0.0 to 0.4	34
0.4 to 0.8	46
0.8 to 1.2	28
1.2 to 1.6	11
1.6 to 2.0	5
2.0 to 2.4	1
2.4 to 2.8	1
2.8 to 3.2	0
3.2 to 3.6	1
3.6 to 4.0	0
Above +4	0

Figure 2 displays a graph of the Test Information Function for all calibrated items. The TIF is a graphical representation of how much information the test is providing at each level of theta. Maximum information was 19.321 at theta = 1.450.

Figure 2: Test Information Function

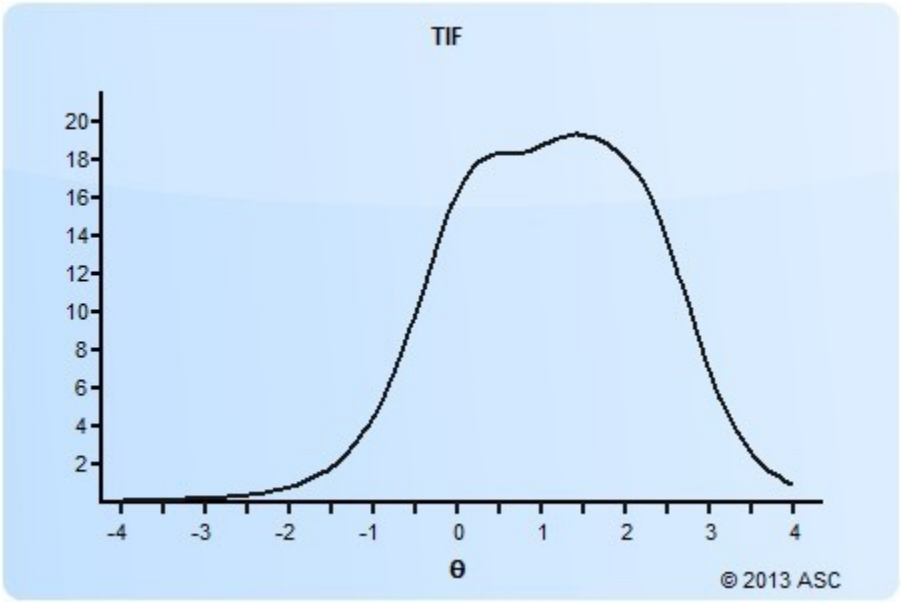
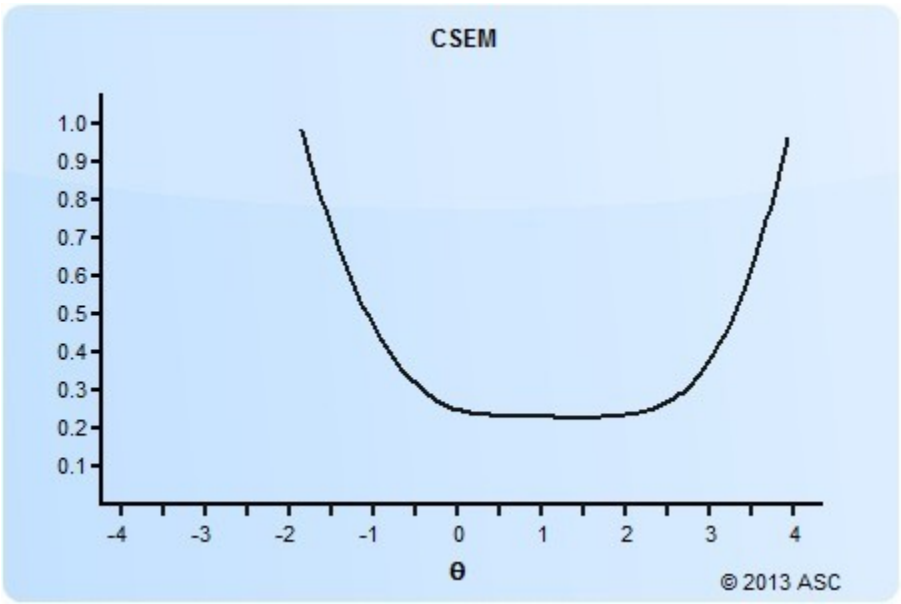


Figure 3 displays a graph of the Conditional Standard Error of Measurement (CSEM) Function. The CSEM is an inverted function of the TIF, and estimates the amount of error in theta estimation for each level of theta. The minimum CSEM was 0.227 at theta = 1.450.

Figure 3: CSEM Function



Item-by-item results

The following section presents the item-by-item results of the analysis. Each scored item has four tables and a plot of the option/category response functions (CRFs).

There are four tables presented for each item.

1. Item information table: records the information supplied by the control file (or Classic Data Header) for this item.
2. Classical statistics table: classical statistics for the item.
3. IRT parameters table: item parameter estimates for the item.
4. Option/Category statistics: detailed statistics for each item, which helps diagnose issues in items with poor statistics.

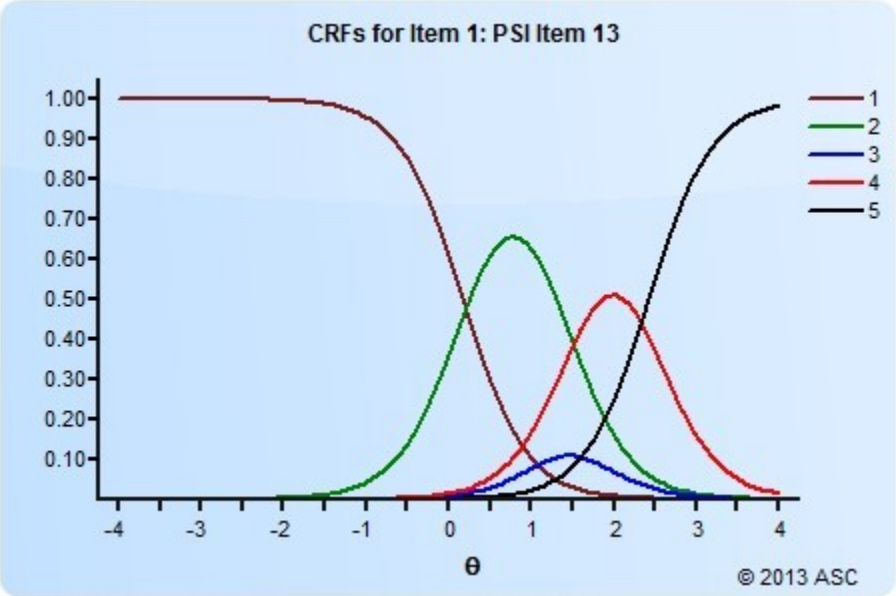
The classical statistics presents classical summary statistics for the item. For polytomous items the item mean and the R and eta correlations are presented in the first three columns of the table. The R correlation is the Pearson correlation between the item and the theta estimates. The Eta coefficient is the square root of the squared eta coefficient from an Analysis of Variance performed on the item and the theta estimates. R ranges from -1 to 1, with a higher value indicating more discrimination. The Alpha w/o is Cronbach's alpha computed with the current item excluded. The item-total correlation is a measure of the discriminating power of the item and is related to the IRT discrimination parameter.

The IRT parameters table presents the IRT item parameters and the fit statistics. The latent trait theta is expressed on a standardized scale, so a one unit change equals a one standard deviation change. The "a" parameter indexes the discrimination of the item, as larger values for "a" will result in a greater slope of the IRF and indicate the item differentiates examinees well. The "b" parameter is the item difficulty parameter and equals the location on the theta continuum where the probability of a correct response equals .50.

The standard errors (SE) for each item parameter estimate are also presented in the item parameter table. A large SE for an item parameter (compared to the other items) indicates that the item parameter was poorly estimated. The chi-square fit statistic and its degrees of freedom are reported for each item. The chi-square fit statistic and its degrees of freedom are reported for each item.

The option/category statistics table presents statistics for each individual option/category. For polytomous items, the boundary location parameters and their SEs are provided in the final two columns of this table. For graded response items, the boundary location represents the location on theta where the boundary response function equals 0.5 (P*). For items with low discrimination, the boundary location will not necessarily be where two category response functions cross.

My child rarely does things for me that make me feel good.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
1	PSI Item 13	SGRM	+	Yes	5	1	

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	1.650	0.502	0.550	0.864

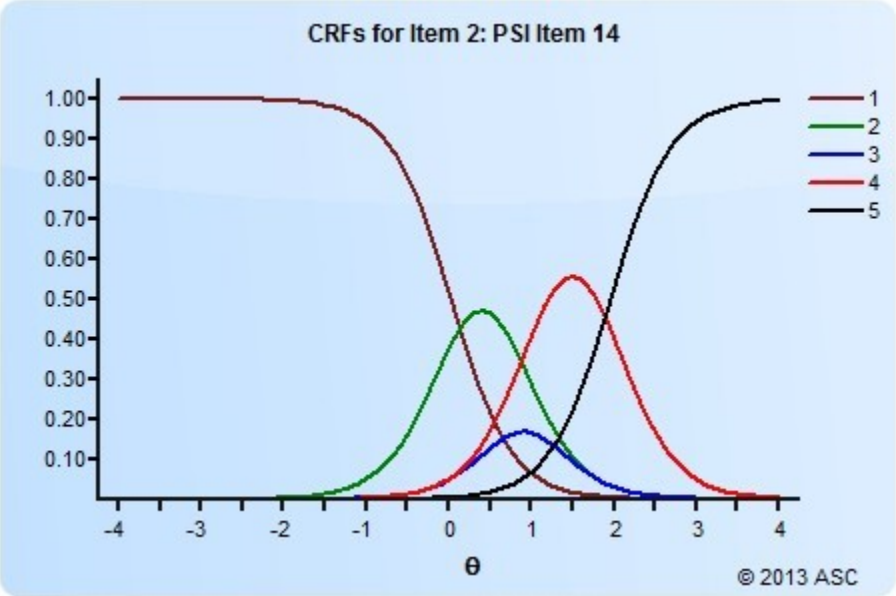
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.527	0.217	38.153	75	1.000

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	132	0.550	-0.931	1.665	0.175	0.068
2	83	0.346	0.509	0.479	1.388	0.057
3	6	0.025	1.224	0.389	1.557	0.063
4	15	0.063	1.008	0.821	2.426	0.206
5	4	0.017	2.279	0.778		
Omit	0					
Not Admin	0					

Sometimes I feel my child doesn't like me and doesn't want to be close to me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
2	PSI Item 14	SGRM	+	Yes	5	1	

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	1.975	0.541	0.596	0.857

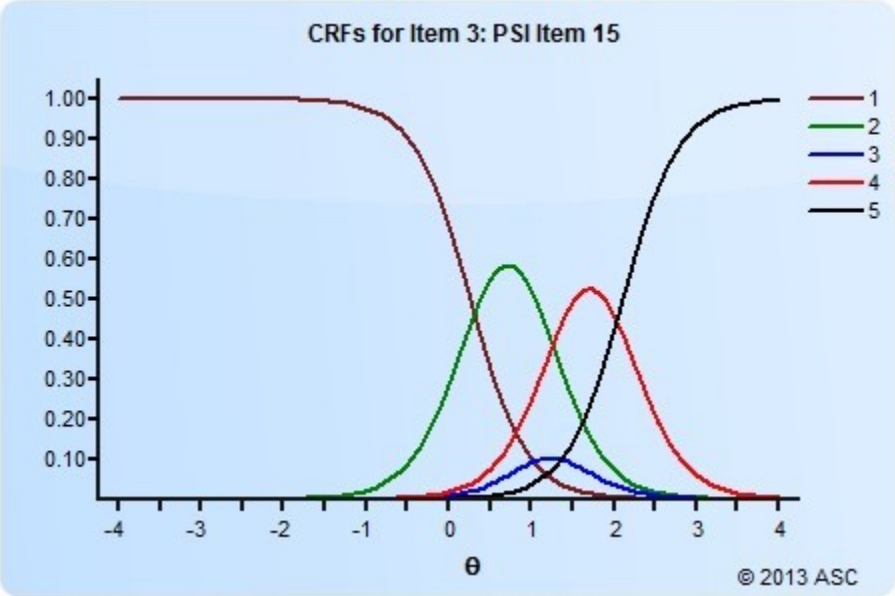
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.595	0.214	89.470	75	0.122

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	117	0.487	-1.101	1.689	0.027	0.064
2	63	0.263	0.407	0.435	0.784	0.049
3	18	0.075	0.696	0.400	1.036	0.054
4	33	0.138	0.755	0.653	1.962	0.140
5	9	0.037	1.853	0.833		
Omit	0					
Not Admin	0					

My child smiles at me much less than I expected.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
3	PSI Item 15	SGRM	+	Yes	5	1	

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	1.700	0.508	0.566	0.862

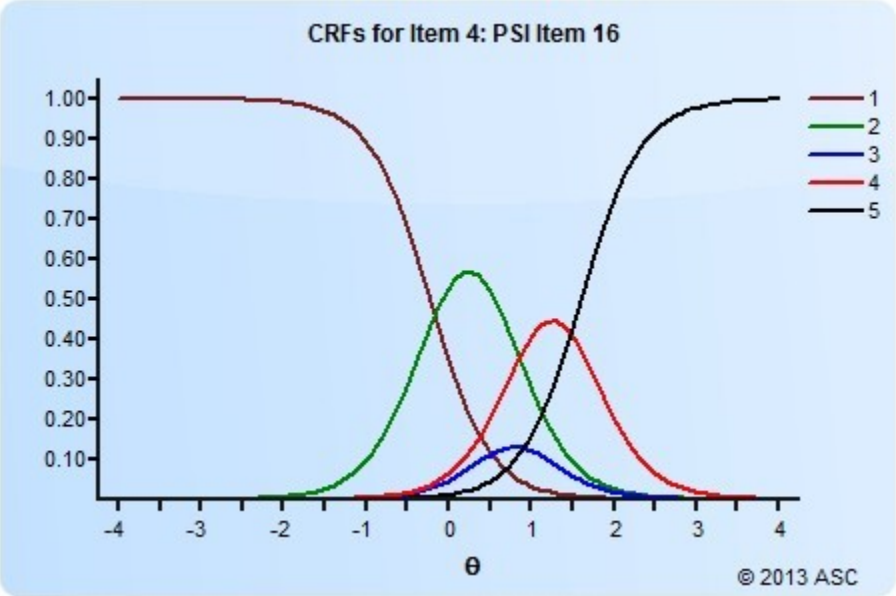
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.728	0.248	52.563	75	0.977

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	139	0.579	-0.925	1.608	0.263	0.062
2	67	0.279	0.590	0.390	1.175	0.046
3	7	0.029	1.174	0.382	1.316	0.050
4	21	0.087	1.047	0.534	2.111	0.158
5	6	0.025	1.636	1.461		
Omit	0					
Not Admin	0					

When I do things for my child, I get the feeling that my efforts are not appreciated very much.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
4	PSI Item 16	SGRM	+	Yes	5	1	

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	2.154	0.580	0.655	0.855

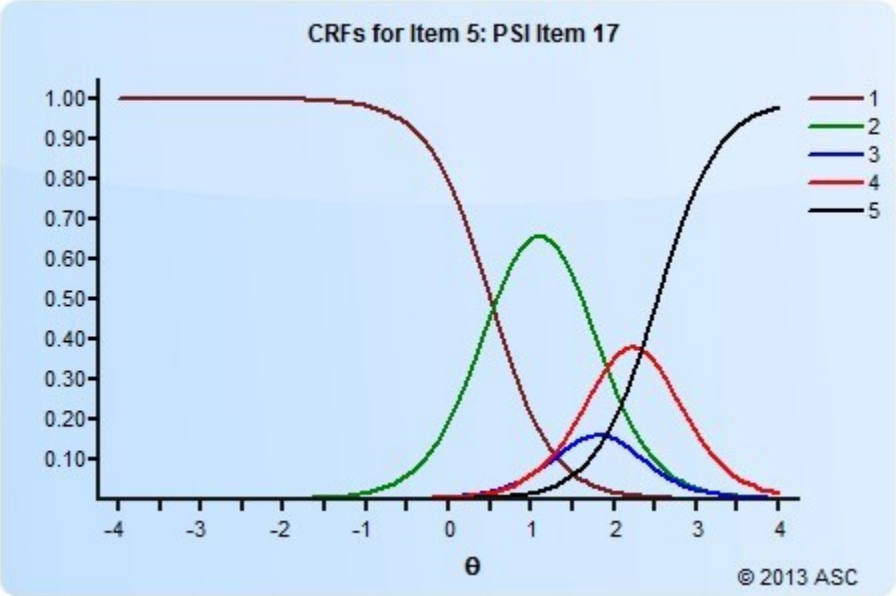
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.612	0.212	91.999	75	0.089

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	95	0.396	-1.384	1.741	-0.228	0.067
2	80	0.333	0.326	0.424	0.713	0.044
3	15	0.063	0.474	0.610	0.907	0.046
4	33	0.138	0.663	0.605	1.607	0.102
5	17	0.071	1.613	0.644		
Omit	0					
Not Admin	0					

When playing, my child doesn't often giggle or laugh.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
5	PSI Item 17	SGRM	+	Yes	5	1	

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	1.442	0.470	0.508	0.868

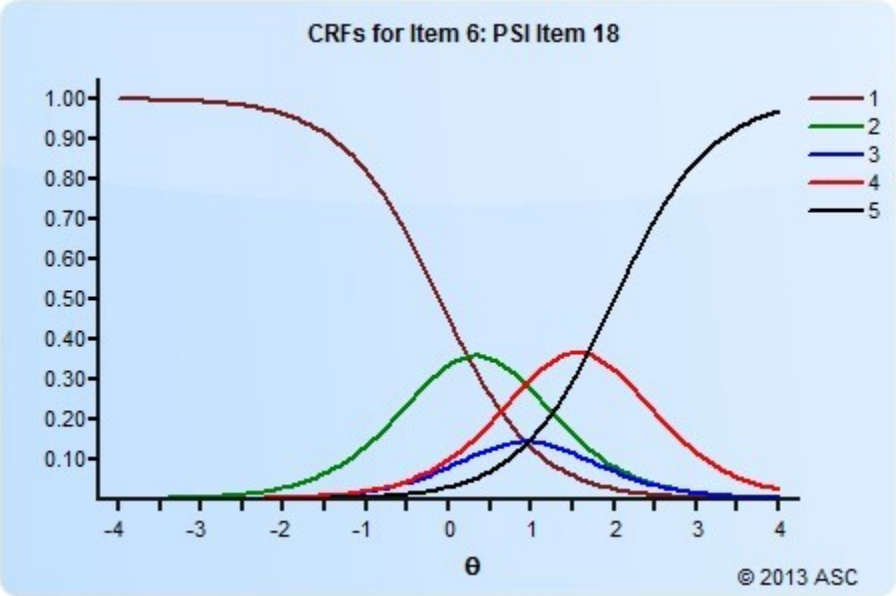
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.572	0.245	64.054	75	0.812

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	160	0.667	-0.738	1.591	0.509	0.068
2	66	0.275	0.747	0.380	1.689	0.081
3	5	0.021	0.901	0.850	1.932	0.093
4	6	0.025	1.391	0.662	2.531	0.215
5	3	0.013	2.323	1.070		
Omit	0					
Not Admin	0					

My child doesn't seem to learn as quickly as most children.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
6	PSI Item 18	SGRM	+	Yes	5	1	F

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	2.142	0.441	0.538	0.876

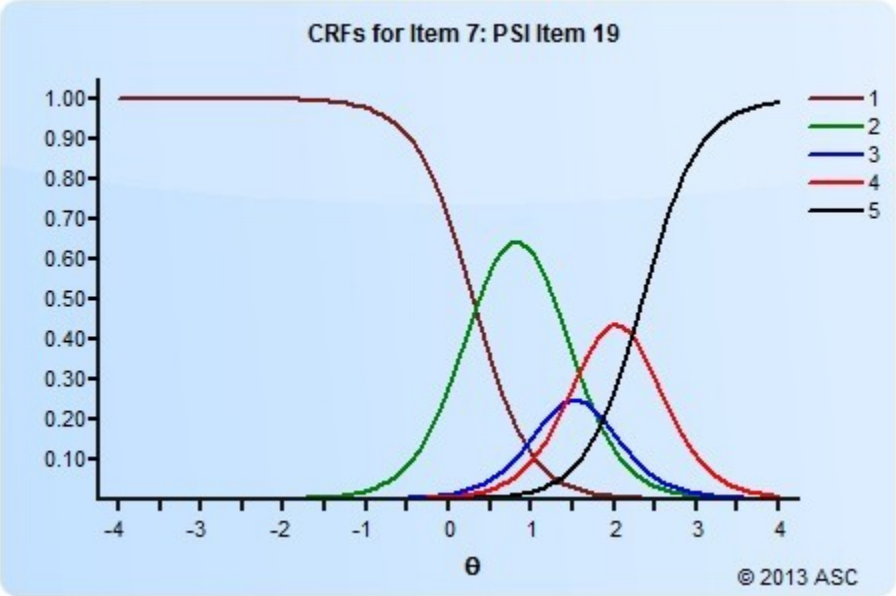
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.003	0.132	154.603	75	0.000

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	106	0.442	-1.128	1.816	-0.111	0.083
2	61	0.254	0.465	0.574	0.770	0.062
3	22	0.092	0.538	0.565	1.111	0.068
4	35	0.146	0.511	0.602	2.010	0.142
5	16	0.067	0.781	1.044		
Omit	0					
Not Admin	0					

My child doesn't seem to smile as much as most children.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
7	PSI Item 19	SGRM	+	Yes	5	1	

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	1.575	0.527	0.565	0.862

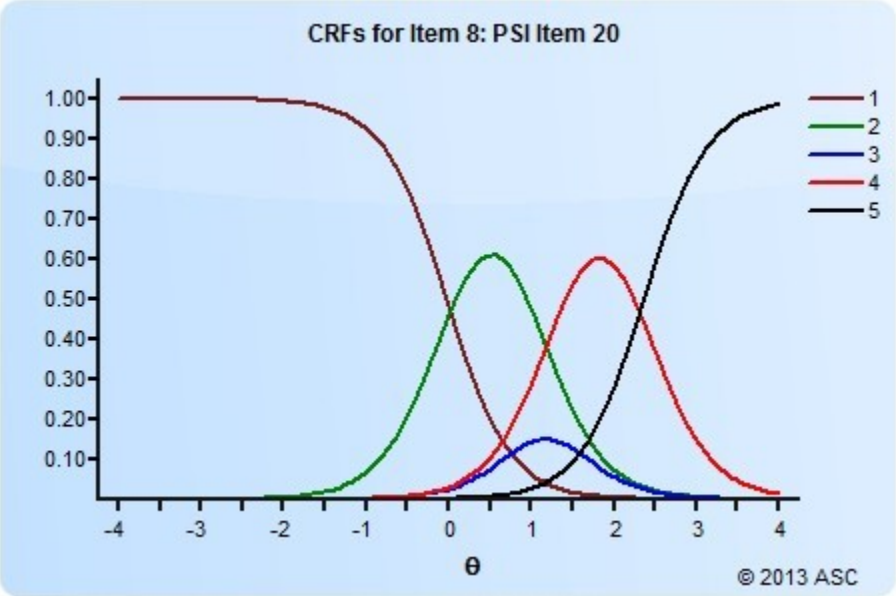
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.700	0.246	41.449	75	0.999

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	143	0.596	-0.889	1.606	0.296	0.063
2	73	0.304	0.627	0.403	1.353	0.071
3	11	0.046	0.970	0.437	1.705	0.089
4	9	0.037	1.418	0.542	2.351	0.191
5	4	0.017	2.239	0.889		
Omit	0					
Not Admin	0					

My child is not able to do as much as I expected.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
8	PSI Item 20	SGRM	+	Yes	5	1	

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	1.846	0.541	0.612	0.860

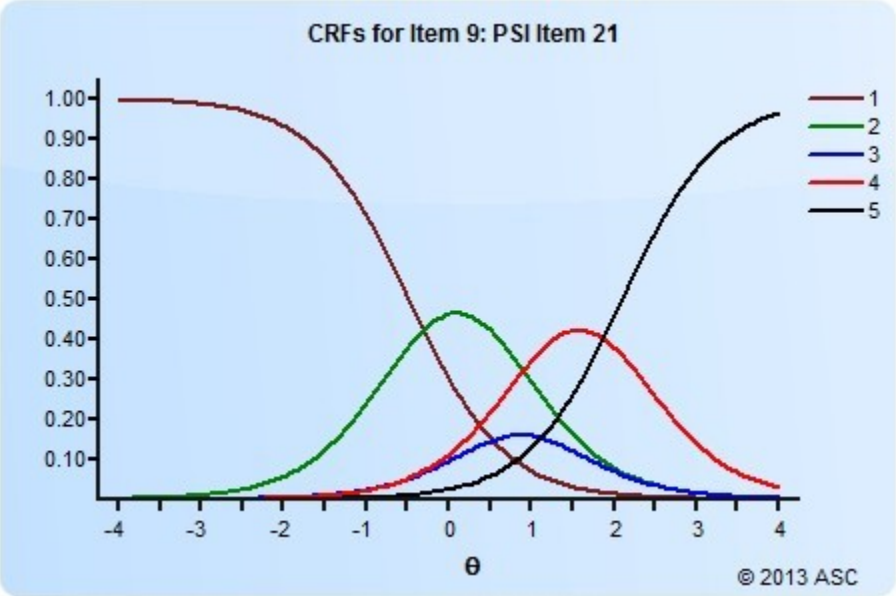
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.532	0.207	59.219	75	0.909

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	114	0.475	-1.168	1.676	-0.025	0.069
2	83	0.346	0.491	0.491	1.063	0.054
3	13	0.054	0.859	0.541	1.297	0.061
4	26	0.108	0.978	0.605	2.368	0.200
5	4	0.017	1.692	1.106		
Omit	0					
Not Admin	0					

It takes a long time and it is very hard for my child to get used to new things.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
9	PSI Item 21	SGRM	+	Yes	5	1	

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	2.258	0.500	0.568	0.868

IRT parameters

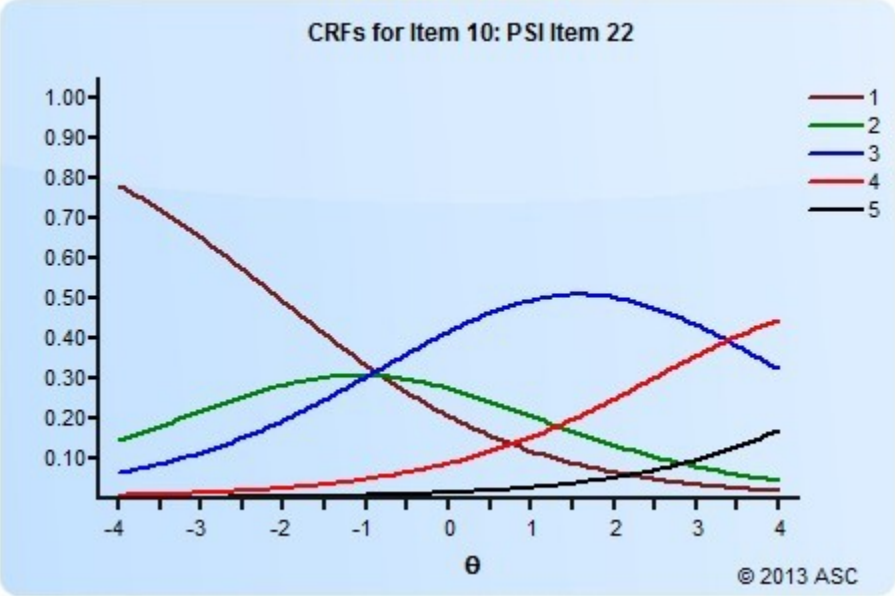
a	*a SE*	*Chi-sq*	*df*	*p*
1.017	0.126	91.184	75	0.098

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	82	0.342	-1.368	1.934	-0.481	0.091
2	82	0.342	0.180	0.609	0.686	0.064
3	23	0.096	0.271	0.736	1.060	0.069
4	38	0.158	0.722	0.823	2.104	0.154
5	15	0.063	0.987	0.948		
Omit	0					
Not Admin	0					

I feel that I am:

1. Not very good at being a parent
2. A person who has some trouble being a parent
3. And average parent
4. A better than average parent
5. A very good parent



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
10	PSI Item 22	SGRM	+	Yes	5	1	F

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	2.421	0.274	0.275	0.888

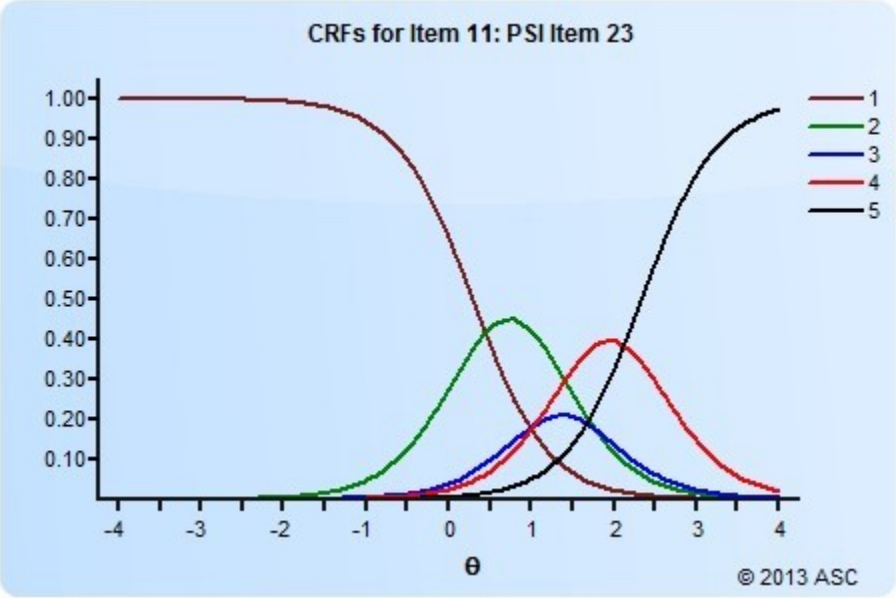
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
0.387	0.044	103.017	75	0.018

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	53	0.221	-0.830	2.540	-2.054	0.210
2	64	0.267	-0.380	1.101	-0.122	0.170
3	96	0.400	0.067	0.997	3.290	0.280
4	23	0.096	0.446	0.835	6.404	0.707
5	4	0.017	0.670	0.457		
Omit	0					
Not Admin	0					

I expected to have a closer and warmer feelings for my child than I do and this bothers me.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
11	PSI Item 23	SGRM	+	Yes	5	1	

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	1.704	0.467	0.508	0.865

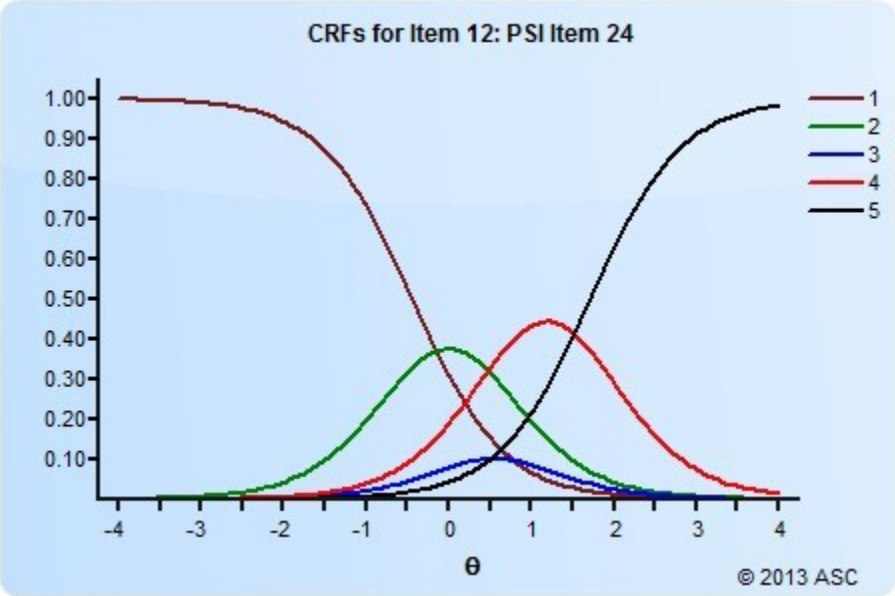
IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.284	0.177	63.473	75	0.826

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	140	0.583	-0.841	1.658	0.290	0.073
2	61	0.254	0.494	0.458	1.177	0.069
3	16	0.067	0.647	0.623	1.568	0.083
4	16	0.067	1.174	0.752	2.336	0.177
5	7	0.029	1.360	1.303		
Omit	0					
Not Admin	0					

Sometimes my child does things that bother me just to be mean.



Item information

Seq.	*ID*	*Model*	*Key*	*Scored*	*Num Options*	*Domain*	*Flags*
12	PSI Item 24	SGRM	+	Yes	5	1	F

Classical statistics

N	*Mean*	*R*	*Eta*	*Alpha w/o*
240	2.454	0.518	0.593	0.868

IRT parameters

a	*a SE*	*Chi-sq*	*df*	*p*
1.062	0.135	107.961	75	0.008

Category statistics

Category	*N*	*Prop.*	*Mean*	*SD*	*Boundary Location (b)*	*b SE*
1	83	0.346	-1.412	1.890	-0.440	0.084
2	64	0.267	0.244	0.557	0.436	0.050
3	17	0.071	0.167	0.729	0.661	0.052
4	53	0.221	0.463	0.735	1.716	0.124
5	23	0.096	1.095	0.888		
Omit	0					
Not Admin	0					