USING ITEM RESPONSE THEORY TO EVALUATE AND REVISE THE CHILD BEHAVIOR QUESTIONNAIRE

A Dissertation

by

DAVID ANGUS CLARK III

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	M. Brent Donnellan
Committee Members,	Vani A. Mathur
	Rebecca J. Schlegel
	Jeffrey Liew
Head of Department,	Heather C. Lench

May 2018

Major Subject: Psychology

ABSTRACT

Early emerging individual differences in reactivity and self-regulation, or temperament, are crucial for understanding development in childhood and beyond. The Child Behavior Questionnaire (CBQ) is currently the most popular measure for assessing temperament in childhood. However, its current length (195 items) may overburden informants. Short forms of the CBQ exist, but these versions may suffer from a lack of measurement precision and content coverage given the procedures used for their development. Modern psychometric techniques based on Item Response Theory (IRT) are well suited to the task of reducing assessment length without compromising measurement quality. Accordingly, the current study used IRT and related techniques to revise the CBQ with the goal of making it more efficient. Result indicated that CBQ could be reduced in length by 44% while still functioning similarly to the original form in terms of measurement precision, inter-parent agreement, and the ability to predict adjustment outcomes. This revised 110-item CBQ, which is substantially shorter and maintains the favorable measurement properties of the original, should prove useful for researchers and clinicians who desire a comprehensive assessment of temperament.

DEDICATION

In memory of Rusty Arthur Clark, whose persistence and self-assurance in life served as a constant source of inspiration on this academic odyssey.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. M. Brent Donnellan, and my committee members, Drs. Vani A. Mathur, Rebecca J. Schlegel, and Jeffrey Liew, for their guidance and support throughout the course of this process. I would also like to thank the wonderful network of collaborators and mentors that I have had throughout my time in graduate school, including Drs. C. Emily Durbin, Brian M. Hicks, Ryan P. Bowles, Amy K. Nuttall, Richard W. Robins, S. Alexandra Burt, and Rebecca J. Brooker. If I have been successful in my scholarly pursuits up to this point, it is largely because of the patience and effort of all those mentioned here. Truly, it takes a village. Thanks must of course also go to all my friends and family for providing all the friend and family stuff that evidence suggests is fairly important for maintaining one's wellbeing. Finally, thanks to the myriad doctors and nurses of Kent General Hospital and the University of Pennsylvania who were a part of my care-network. If not for their efforts I would have been deep in the cold, cold ground long before I ever made it to graduate school.

CONTRIBUTORS AND FUNDING SOURCES

This work was supported by a dissertation committee consisting of Professors M. Brent Donnellan [advisor], Vani A. Mathur, and Rebecca J. Schlegel of the Department of Psychological and Brain Sciences, and Professor Jeffrey Liew of the Department of Educational Psychology. The data analyzed in this dissertation was provided by Professors M. Brent Donnellan and C. Emily Durbin (of Michigan State University). All other work conducted for the dissertation was completed by the student independently.

Funding for the datasets used in this project came from multiple sources. Sample 2 data collection was supported by the Kovler Research Scholar Fund of The Family Institute at Northwestern University. Sample 3 data collection was supported by a grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (HD064687).

TABLE OF CONTENTS

Page

LIST OF TABLES

1. INTRODUCTION

Children vary from one another on a wide variety of observable dimensions, such as activity level, self-control, reaction to novelty, and tolerance of frustration. These early emerging individual differences in emotional reactivity and self-regulation reflect differences in children's *temperament* (Rothbart, Ahadi, & Hershey, 1994; Shiner & DeYoung, 2013; Shiner & Caspi, 2012). Individual differences in child temperament form the foundation of adolescent and adult personality (Shiner & DeYoung, 2013), and predict consequential short- and long-term outcomes such as psychopathology (Caspi, Moffitt, Newman, & Silva, 1996; Klein, Dyson, Kujawa, & Kotov, 2012; Tackett, Martel, & Kushner, 2012), academic performance (Duckworth & Allred, 2012), and substance use (e.g., Clark, Donnellan, Robins, & Conger, 2015; Creemers et al., 2010; Stautz & Cooper, 2013). Research on child temperament thus has the potential to contribute to a better understanding of human development across the lifespan in several domains (e.g. social, emotional, health). Likewise, research on temperament may provide clues about how to promote positive youth development (Moffitt et al., 2011).

Accordingly, temperamental characteristics are increasingly recognized as critical features of the developmental milieu in early life and beyond (Clark, Durbin, Hicks, Iacono, & McGue, 2017; Rothbart, 2011; Zentner & Shiner, 2012), with a growing number of researchers interested in incorporating information on child temperament into their work (Zentner & Shiner, 2012). For example, a recent (May, 2017) search on the "PsychInfo" database for child (ages 0 to 12 specified) temperament related keywords returned 21,627 hits between the years 1963 (the year of the initial New York Longitudinal Study report that laid the groundwork for modern temperament research; Thomas et al., 1963) and 2000 (an average hit rate of roughly 585 per year), and 30,635 hits between the years of 2001 and 2017 (an average hit rate of roughly 1,915

per year). With the rapid growth of this area, more investigators than ever before are being faced with the issue of how to measure child temperament, and most will likely opt to rely on parent reports (Goldsmith & Gagne, 2012).

To be sure, researchers have many parent report questionnaires to choose from (Gartstein, Bridgett, & Low, 2012). Perhaps the most prominent of these is Rothbart and colleagues' Child Behavior Questionnaire (CBQ; Rothbart, Ahadi, Hershey, & Fisher, 2001; Kotelnikova, Olino, Klein, Kryski, & Hayden, 2015). The CBQ's popularity in the field is not unwarranted as it is a truly comprehensive inventory. However, it is also a very long inventory, containing almost 200 items. This may prove restrictive in many circumstances. Short, and very short, forms of the CBQ exist (Putnam & Rothbart, 2006), but researchers and practitioners may be hesitant to sacrifice the precision and content coverage that are associated with short forms in general (e.g., Crede, Harms, Niehorster, & Gaye-Valentine, 2012).

Modern psychometric techniques based Item Response Theory (IRT; Embretson & Reise, 2000; Revicki & Reise, 2014) and categorical structural equation modeling (SEM; Kamata & Bauer, 2008; Wirth & Edwards, 2007), however, provide useful tools for creating particularly efficient short forms. That is, these techniques allow researchers to more simultaneously satisfy competing concerns regarding survey length and quality. In practical terms, this means that the psychometric insights these techniques provide can facilitate the reduction of survey length without substantially impacting measurement precision or content coverage. Accordingly, the goal of the present study was to apply these modern psychometric techniques to the task of optimizing the CBQ by identifying and trimming weak and redundant items. The goal was not to simply create another CBQ short form, rather, it was to improve and revise the standard CBQ.

1.1 Dimensions of Child Temperament

There are many theoretical frameworks for organizing the multitude of temperamental differences that can be observed in children (Goldsmith et al., 1987; Mervielde & De Pauw, 2012). Notably, these various approaches tend to be similar in content and structure to the major personality models used in the adult literature (Rothbart, 2007). Specifically, individual differences between children are organized hierarchically with narrower facets at the bottom (e.g., cheerfulness, feelings of vulnerability) and broader, higher order dimensions (e.g., Extraversion, Neuroticism) at the top. These higher order dimensions function as latent variables that capture patterns of covariation among the lower level facets. Rothbart and colleagues' Psychobiological Model (Rothbart, 2011) of early childhood temperament, which serves as the basis for the CBQ, specifically includes16 lower order facets (see Table 1) that are subsumed by three higher order dimensions: Effortful Control, Negative Affectivity, and Surgency.

The dimension of Effortful Control captures individual differences in the ability to focus and shift attention, inhibit inappropriate responses, and regulate emotions (Rothbart, 2011). Effortful Control is the childhood analog to the traits of Conscientiousness and Constraint from the popular Big Five and Big Three frameworks of adult personality, respectively (Clark & Watson, 2008; Shiner & DeYoung, 2013; Tellegen & Waller, 2008). Negative Affectivity captures individual differences in the tendency to experience and express negatively valenced emotions such as sadness, fear, and anger (Rothbart, 2011). This trait corresponds to the traits of Neuroticism and Negative Emotionality from the Big 5 and Big 3 frameworks (Clark & Watson, 2008; Shiner & DeYoung, 2013; Tellegen & Waller, 2008). Finally, Surgency captures individual differences in the tendency to experience and express positively valenced emotions, and to approach rewarding and novel stimuli (Rothbart, 2011). Surgency corresponds to the Big

5 and Big 3 traits of Extraversion and Positive Emotionality (Clark & Watson, 2008; Shiner & DeYoung, 2013; Tellegen & Waller, 2008).

1.2 The Child Behavior Questionnaire

Research on child temperament has largely relied on parent report measures to assess individual differences between children (Clark et al., 2017; Goldsmith & Gagne, 2012; Lo, Vroman, & Durbin, 2014). That is, instead of asking children directly about their behavior and traits, parents provide the ratings of their child's temperament. Although there are many different parent report forms along these lines, the CBQ (Rothbart et al., 2001) -- based on Rothbart's Psychobiological model of temperament and designed for use with children between 3 and 7 years old -- is currently the most widely used parent report measure for assessing temperament in early childhood (Kotelnikova et al., 2015). Effectively a standard for the field (Kotelnikova et al., 2015), the foundational CBQ paper (Rothbart et al., 2001) has been cited over 1,700 times since its release as of May 2017 according to Google Scholar (with the paper describing the development of the short form -- Putnam & Rothbart, 2006 -- being cited over 500 times since its release). The CBQ has also been translated into over 20 non-English languages (e.g., Arabic, Chinese, Italian), giving it a broad international/cross-cultural presence ("The Children's Behavior Questionnaire", 2017).

The CBQ includes 195 items to measure 16 distinct facets of temperament: Activity Level, Anger/Frustration, Attentional Focusing, Attentional Shifting, Discomfort, Falling Reactivity/Soothability, Fear, High Intensity Pleasure, Impulsivity, Inhibitory Control, Low Intensity Pleasure, Perceptual Sensitivity, Positive Anticipation/Approach, Sadness, Shyness, and Smiling/Laughter (see Table 1 for conceptual definitions). For a given item, respondents rate the degree to which that behavior characterized the target child over the past six months.

Responses are given on a 7-point scale that ranges from “1: extremely untrue of your child” to “7: extremely true of your child”. Item content was derived both rationally, based on the Psychobiological Model of temperament, and from extensive parental interviews. Items tend to emphasize specific behaviors (e.g., “gets mad when only mildly criticized”) instead of more global judgements, which can increase the quality of informant reports, and protect against certain biases (e.g., “contrast effects”; Goldsmith, Lemery, Buss, & Campos, 1999).

The 16 scales of the CBQ load on the broad dimensions of Effortful Control, Negative Affectivity, and Surgency (Rothbart et al., 2001), though alternate higher order factor structures have been identified (e.g. Kotelnikova et al., 2015). The Effortful Control dimension is typically associated with the attentional focusing, attentional shifting, inhibitory control, low intensity pleasure, and perceptual sensitivity scales (Rothbart et al., 2001). Negative Affectivity includes the anger/frustration, discomfort, sadness, fear, and soothability scales (Rothbart et al., 2001). Surgency includes the activity level, smiling and laughter, high intensity pleasure, impulsivity, shyness, and positive anticipation scales (Rothbart et al., 2001). Lower order scale scores are computed by aggregating together a scale’s individual items. Higher order dimensions are typically computed by aggregating the scale scores that are included under a given dimensions.

Overall, the CBQ is a carefully constructed and thoroughly psychometrically evaluated (e.g., Clark, Listro, Lo, Durbin, Donnellan, & Neppl, 2016; Kotelnikova et al., 2015; Rothbart et al., 2001) measure that provides an impressive amount of information about a given child’s temperament. However, this information comes at a rather steep cost of time and effort on the part of respondents given the length of standard form. This issue is compounded by the fact that parental informants are frequently expected to fill out multiple questionnaires for a study, many of which may also be rather lengthy (e.g., the popular “Child Behavior Checklist” which

contains over 200 items; Achenbach & Ruffle, 2000). Additionally, ratings of temperament are being increasingly used in applied settings that favor quick assessment (e.g., pediatrics and interventions) (Carey & McDevitt, 1989; McClowry & Collins, 2012; Schroder, Clark, & Moser, 2017; Smith, McCarthy, & Anderson, 2000).

Given these concerns, a shorter CBQ would be advantageous. To be sure, both short and very short CBQ forms do exist (Putnam & Rothbart, 2006). However, short forms developed using traditional methods often entail sacrificing a substantial degree of measurement precision and content coverage (Smith, McCarthy, & Anderson, 2000). Researchers and practitioners may be hesitant to make this tradeoff. Notably however, though the push and pull between conciseness and assessment quality will always exist to some degree, modern psychometric techniques, specifically those associated with Item Response Theory (IRT; de Ayala, 2009), can help minimize the costs inherent in this tradeoff (Hambleton & Swaminathan, 1985; Schroder, Clark, & Moser, 2017). That is, these more powerful psychometric frameworks incorporate a recognition that more items do necessarily beget better measurement across the underlying latent continuum; some items may be weak and uninformative indicators of the underlying latent variable, whereas others may be redundant past a point of diminishing returns. Accordingly, the application of these techniques to the CBQ could potentially reduce the amount of items on the original form, while also maintaining the original's precision, content coverage, and other desirable measurement properties.

1.3 Item Response Theory

IRT refers to a psychometric meta-framework (not necessarily a "theory" per se) that includes a broad network of latent variable models that provide considerable detail about measurement functioning at both the item and test level (de Ayala, 2009; Embretson & Reise,

2000). Despite their differences, the standard IRT models based on a dominance process (see Stark, Chernyshenko, Drasgow, & Williams, 2006) tend to be based on the same basic underlying principles (Stark, Chernyshenko, & Drasgow, 2006; Reise, Widaman, & Pugh, 1993). Essentially, IRT models are confirmatory factor analyses (CFA) for categorical indicators in which the likelihood of a given item response is modeled as a probabilistic function of the test taker's standing on the latent trait of interest (Kamata & Bauer, 2008). That is, individual items function as indicators of a latent trait, or traits, which undergird responses across the test. This latent trait, often referred to as theta, represents the construct of interest that is being measured (e.g., math ability).

Theta is typically tied to item responses through two major types of parameters: discriminations and location parameters (difficulties or thresholds depending on whether item responses are dichotomous or polytomous). Discrimination values are analogous to factor loadings, and denote how strongly an item is related to theta (i.e., how capable the item is of discriminating between different levels of theta). A discrimination value of .80 has traditionally been considered the threshold for acceptability (a discrimination value of .80 roughly corresponds to a standardized factor loading of around .43; de Ayala, 2009). Location parameters (difficulties/thresholds) are somewhat analogous to indicator intercepts in a CFA model, and capture in one way or another (there are subtle differences across different types of models) the point along theta's continuum at which individuals are more likely to endorse a higher response category (e.g., when children are more likely to be correct versus incorrect on a math question given a particular level of math ability). A more "difficult" item is thus one in which a higher standing on theta is required for there to be non-trivial likelihood of endorsing higher response categories. When measuring psychological attributes (as opposed to abilities), it is generally

advisable to have a wide range of location parameters on a test so that as much of the theta continuum as possible is being adequately assessed (de Ayala, 2009).

Although certain IRT models may include additional parameters (e.g., a lower asymptote; de Ayala, 2009), the discrimination and location parameters are the most critical parameters for most applications of IRT in the psychological sciences. These parameters are also related to the concept of information. "Information" is how IRT models address the issue of reliability, and captures how precisely an item, or test, measures individuals at different levels of theta (i.e., how much information is provided about those test takers; Thissen & Orlando, 2001). The information provided by a given item or test across a range of theta is often presented graphically as an Information Curve (or Information Function). Information at the test level, as opposed to the item level, merely represents an aggregate of all the item information curves (plus a value of 1 if certain estimation techniques are used). Overall, more discriminating items provide more information, and information is generally highest around the location parameters. Importantly, information values are derived in the estimation process, and presented, as "logits", which do not have much inherent meaning and are thus difficult to interpret; however, these logits can easily be converted into interpretable estimates of standard error and reliability. For example, 3 logits of information corresponds to a standard error of .58 ($1/\sqrt{(3)}$), which in turn corresponds to a conditional reliability value of around .67 ($1-.58^2$; alternatively $1 - (1/3)$; Thissen & Orlando, 2001).

IRT is often presented in contrast to Classical Test Theory (CTT), which despite being the predominant measurement framework in psychology, is problematic both practically and conceptually (Borsboom et al., 2009). Indeed, because of this and the fact that the development and evaluation of the CBQ (and other temperament questionnaires) has thus far been based

primarily on CTT, there have been calls to apply IRT techniques to temperament assessment so as to take advantage of its many strengths (Goldsmith & Gagne, 2012). For example, unlike in CTT, IRT parameter estimates are generally sample invariant (though they may sometimes ostensibly appear different given the distribution of theta in a given sample), which means scale development results from IRT analyses more readily generalize than those based on CTT (de Ayala, 2009; Hambleton & Swaminathan, 1985; Markus & Borsboom, 2013).

Furthermore, IRT provides a more nuanced and comprehensive take on the issue of reliability such that test and item precision is conditional on the attribute being assessed (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). Instead of a single number that supposedly captures the reliability of a test, information values are bound to certain levels of theta. Thus, certain tests or items might demonstrate, for instance, considerable precision when assessing individuals at average and above average levels of theta, while simultaneously demonstrating sub-par precision when assessing individuals below the average level of theta. Practically, this means that it is possible to make finer grained distinctions (with confidence) between individuals where information is highest. This also allows for more targeted test construction such that items can be developed that target a specific range of theta, with less effort being put into measuring the less relevant ranges of theta (thus potentially reducing assessment length).

Perhaps one of the most useful conceptual advantages of IRT over CTT is the more explicit recognition and incorporation of the fact that more items per se are not necessarily associated with better, more reliable measurement (a foundational principal behind computer adaptive testing; de Ayala, 2009; Hambleton & Swaminathan, 1985). Long questionnaires may contain several items that help boost coefficient alpha, but beyond that contribute virtually

nothing to the measurement of the target construct (due to weakness and/or redundancy). Alternatively, a small collection of items, or even single items, may provide an adequate degree of information across a reasonable span of theta (e.g., Schroder, Clark, & Moser, 2017). Indeed, IRT models are well suited for helping researchers reach an optimal balance between precision and parsimony. Accordingly, the primary goal of the present study is to use IRT methods to evaluate and streamline the CBQ.

1.4 Present Study

In the present study, parent ratings on the CBQ for 605 children (drawn from three independent samples) were initially submitted to a thorough IRT analysis. This in and of itself provides a large amount of useful information on how the CBQ functions. However, the main goal was to gain insight into which items do not appear to meaningfully contribute to the assessment of the target constructs, and the psychometric consequences of removing these items. This information was used to edit the CBQ by trimming suboptimal items. The functioning of the revised CBQ scales and dimensions were then examined in both the original "calibration sample" (maternal informants), and a semi-independent "validation sample" (paternal informants). That is, the initial evaluation and editing of the CBQ was based on analyses using maternal reports, whereas paternal reports were primarily used to replicate, support, and further refine these results.

The overarching aim was to shorten the CBQ as much as possible while preserving the original form's measurement quality and content coverage. Given the growing interest in child temperament, it is useful to re-visit the popular CBQ with contemporary psychometric techniques in order to improve its efficiency. To be sure, the CBQ is a comprehensive and thoughtfully developed questionnaire based on a rich theory of temperament. It has earned its

place of prominence in the developmental literature. Thus, a revision of this powerful inventory that retains the advantages of the original form promises to offer benefits to both researchers who want to examine temperament in as much detail as possible, as well as informants and practitioners that have a vested interest in parsimonious measurement tools.

2. METHOD[1]

2.1 Participants

The data come from three separate samples of children and their parents. Characteristics of each individual sample are provided below, followed by a brief review of the total sample. This combined sample has been used in a prior study examining measurement invariance across informants (mothers and fathers) and child gender in the CBQ at the scale/dimensional level (Clark et al., 2016).

2.1.1 Sample 1

Participants were recruited from the greater Chicago area for a study of child temperament. Children who did not have any significant medical conditions or developmental disabilities and lived with at least one English-speaking parent were eligible for participation in the study. Participating children visited the laboratory with their mother or father for a 2-hour assessment consisting of tasks designed to elicit discrete emotions and behaviors indicative of temperament traits. At the end of the lab visit, the parent was given a battery of questionnaires to complete and return by mail.

This sample included 206 children between the ages of 3 and 7 years (48.1% girls). The mean age of the children was 56.4 months (*SD* = 12.0; range = 36 - 83). Mothers were between the ages of 23 and 49 years (*M* = 36.9, *SD* = 4.8), and fathers were between the ages of 23 and 57

[1] Parts of this section (participants) have been reprinted from Clark et al., 2016. Copyright © 2016 American Psychological Association. Reproduced with permission. The official citation that should be used in referencing this material is: Clark, D. A., Listro, C. J., Lo, S. L., Durbin, C. E., Donnellan, M. B., & Neppl, T. K. (2016). Measurement invariance and child temperament: An evaluation of sex and informant differences on the Child Behavior Questionnaire. *Psychological Assessment, 28*(12), 1646-1662. This article may not exactly replicate the authoritative document published in the APA journal. It is not the copy of record. No further reproduction or distribution is permitted without written permission from the American Psychological Association.

years (*M* = 38.8, *SD* = 5.8). Data on race and ethnicity and family income were provided by 72.1% of mothers and by 70.2% of fathers. Of those, the ethnic composition was as follows: Caucasian/White (77.4%), Hispanic/Latino (10.1%), African American/Black (8.0%), Asian (5.9%), other (3.1%), and bi- or multiracial (2.8%)[2]. Yearly family income ranged from $21,000 to greater than $100,000; 18.4% reported income less than $41,000. Approximately 74% of the children came from two-parent households, which is slightly higher than the rate of two-parent households in the surrounding area (Cook County, Illinois) from which this sample was drawn (67%; U.S. Census, 2014).

2.1.2 Sample 2

Participants were recruited from the greater Lansing, Michigan area for a study of child temperament. Children who did not have any significant medical conditions or developmental disabilities and lived with at least one English-speaking parent were eligible for participation in the study. Procedures for data collection were identical to those described above for sample 1.

This sample included 277 children between the ages of 3 and 7 years (49.5% girls). The mean age of the children was 59.9 months (*SD* = 17.0; range = 36 – 95). Data on race and ethnicity and family income were provided by 65.0% of mothers and by 40.8% of fathers. Of those, the ethnic composition was as follows: Caucasian/White (77.6%), Hispanic/Latino (7.8%), African American/Black (10.9%), Asian (1.6%), other (3.7%), and bi- or multiracial (5.1%). Yearly family income ranged from less than $10,000 to greater than $100,000; 21.7% reported income less than $41,000. Approximately 79% of the children came from two-parent

[2] Categories do not sum to 100% because participants could endorse multiple categories

households, which is slightly higher than the rate of two-parent households in the surrounding area (Ingham County, Michigan) from which this sample was drawn (71%; U.S. Census, 2014).

2.1.3 Sample 3

Participants were drawn from the Family Transitions Project (FTP), an ongoing longitudinal study of 559 target individuals and their families of destination in adulthood (see Elder & Conger, 2000; Neppl et al., 2010). The children in this sample are the offspring of the original FTP targets, and thus the families include one original FTP target, her/his child, and the other parent of the child. In almost all cases, both parental informants are the biological parents of the child. Data for the FTP is collected regularly by trained interviewers who visit the participants in their home. During these visits targets complete multiple questionnaires spanning a wide range of topics. The parent reported temperament data are from the first administration of the CBQ for each family with an eligible child.

This sample included 222 children between the ages of 3 and 5 years (46% girls). The mean age of the children was 39.7 months (*SD* = 7.89; range = 36 - 60). Mothers were between the ages of 18 and 41 years (*M* = 26.07, *SD* = 2.91), and fathers were between the ages of 18 and 43 (*M* = 28.01, *SD* = 3.75). Data on race and ethnicity and family income were available for 97% of mothers and by 95% of fathers. Of those, the ethnic composition was as follows: Caucasian/White (97%), Hispanic/Latino (1.12%), African American/Black (0.5%), Asian (0.47%), other (0.5%), and bi- or multiracial (0.25%). Yearly family income ranged from less than $10,000 to greater than $100,000; 41% reported income less than $41,000. Approximately

79% of the children came from two-parent households, which is equal to the rate of two-parent households in the state of Iowa[3] (79%; U.S. Census, 2014).

2.1.4 Total Sample

There were a few notable demographic differences between samples. Children from sample 3 were younger than the children from both samples 1 (Cohen's d = 1.64) and 2 (d = 1.52). Mothers (d = 2.73) and fathers (d = 2.21) were also younger on average in sample 3 compared to sample 1 (parent age was not available in sample 2). Compared to samples 1 and 2, sample 3 was also less ethnically diverse, with nearly all parents reporting their ethnicity to be Caucasian (97% versus approximately 77% in samples 1 and 3). The average level of annual family income was similar for samples 2 and 3 (between $40,000 and $60,000), whereas families from sample 1 on average reported higher levels of annual income (between $60,000 and $100,000).

Despite these demographic differences, mean differences between samples on the CBQ scales were generally small in magnitude, and unsystematic. When combined, the total sample included 605 children with CBQ data (47.3% girls) aged 3 to 7 years. The mean child age was 52 months (SD = 10.79; range = 36 - 95), or 4.3 years. The average age of mothers was 31.49 years (SD = 7.66; range = 18 - 49). The average age of fathers was 33.44 years (SD = 7.49; range = 18 - 57). Family incomes ranged from below $10,000 annually, to over $100,000; 27% of all households included reported income of less than $41,000 yearly. In all there were 588 maternal reports of child temperament, and 479 paternal reports of child temperament.

[3] Although they are spread out across the state, most families included here from the FTP still reside in Iowa.

2.2 Data Analytic Strategy

In creating short or revised questionnaires, it is important to have both a calibration sample, and an independent validation sample (Smith, McCarthy, & Anderson, 2000). Here, maternal ratings were used as the initial calibration sample while paternal ratings served as a "semi-independent" validation sample with which to confirm the quality of the revised form. Given evidence suggesting that mothers and fathers generally use the CBQ equivalently (Clark et al., 2016; Clark, Durbin, Donnellan, & Neppl, 2017), there are likely not any major psychometric differences between parents that invalidate the splitting of the sample by informant in this way.

Each of the 16 individual scales was examined in turn using the same general procedure outlined below; the superordinate dimensions were considered after each scale that is included in a given dimension was analyzed. First, the original CBQ scale was evaluated using the graded response model (GRM), an IRT model for polytomous items (Samejima, 1969; Samejima, 2010). The GRM is what is known as a "cumulative logits" model, meaning the threshold value for any given item response category represents the point of theta at which there is a 50% chance of scoring in that category or above (e.g., the second threshold in a five category item would represent the contrast of responses 1 and 2 versus 3, 4, and 5; Embretson & Reise, 2000). The initial GRMs provided baseline information on the functioning of the CBQ scales as they are typically used in practice, while highlighting potentially weak items.

The second step was to evaluate scale dimensionality. One assumption of the GRM is that the scale being tested is unidimensional, or at least essentially unidimensional (see e.g., Slocum-Gori, Zumbo, Michalos, & Diener, 2009; Slocum-Gori & Zumbo, 2011). Furthermore, unidimensionality at the scale level is implied by the nature of the CBQ. Dimensionality was assessed via Item Factor Analysis (IFA; Wirth & Edwards, 2007) and the bi-factor model

(restricted or unrestricted, based on the IFAs; Stucky, Thissen, & Edelen, 2013). The bi-factor model was specifically used when there was evidence of multidimensionality as a means of determining the extent to which items were related to the primary factor of interest versus specific, or "nuisance", factors (items that did not meaningfully load on any factor in the IFAs were not included in the bi-factor models). Item level "explained common variance" (I-ECV) captures the percentage of an item's total shared variance that is explained by the general factor (as opposed to the specific factors), and was used to quantify the extent to which items were related to the primary dimension of interest (e.g., Hansen et al., 2014; Rodriguez et al., 2016). Notably, bi-factor models in which there are specific factors that include only two items are more likely to encounter estimation issues. When this scenario arose, one item from the two-item dimension was retained in the revised scale on the basis of the other selection criterion while the bi-factor model omitted the two-item dimension.

In the third step, the original scale was trimmed and the revised scale was analyzed using the GRM. Items that demonstrated low initial discrimination (defined here as less than .60) and/or failed to load meaningfully on any factor in the IFA (defined here has above .40) were flagged for potential elimination, as were items with I-ECV values below 35%. At minimum, however, at least one item per IFA dimension was included in the revised scale, with priority going to those items with I-ECV values above 35%. This strikes a compromise between content coverage (multidimensionality often being indicative of a specific facet of a scale) and the assumptions of the model[4], while providing a more holistic appraisal of each item (e.g., weak initial discrimination may be due to the presence of distinct sub-factors). It is important to note

[4] A brief set of simulations run prior to analysis suggests that the mild amount of multidimensionality likely to be introduced via this procedure (based on preliminary analyses) is not likely to substantially bias results. The average level of bias in discrimination and intercept values was less than 10%.

that these criteria represented guidelines and heuristics. Flexibility was sometimes necessary in order to maintain adequate content coverage and precision in the revised scales. Additionally, as the overarching goal was to shorten the questionnaire, if every item was exemplary by the standards laid out here, relatively weaker items were trimmed. Indeed, there are diminishing returns on information such that at higher levels, more information does not necessarily confer meaningfully more precision (e.g., 8 logits of information corresponds to a reliability of .86 while 16 logits of information corresponds to a reliability estimate of .94; a 100% increase of information here only corresponds to a 9% increase in reliability), making some items potentially redundant.

There were two targets for the revised scale. The first was to maintain content coverage to ensure that the conceptual richness of each scale was preserved. This was done by including at least one item from each dimension identified in the dimensionality assessment in the revised scale. The second was to maintain an acceptable degree of precision in the revised scales. Specifically, the target was for the revised scales to provide at least 4 logits of information (i.e., standard error of .50; reliability of .75) in the majority of the space between two standard deviations above and below theta's mean. If even the original scale failed to meet this target, the goal was to simply maintain a comparable level of precision in the revised scale. Marginal reliability was also considered when assessing and revising scales, and is presented along with information. Marginal reliability (Thissen, Nelson, & Swygert, 2001) is more similar to the CTT approach to reliability in which a single number is provided as a holistic summary of the amount of information a scale provides. Marginal reliability can be thought of as the "average" reliability across theta, or the average information provided.

In the fourth and final step, the revised scales (and dimensional composites) were compared to the original scales (and dimensional composites) using both maternal and paternal reports. First, the revised scales using *paternal* ratings were analyzed with the GRM. Second, descriptive statistics for each scale, and the correlations between the revised and original scales were, computed. Third, inter-parent agreement was examined for both the revised and original scales. Fourth, the revised and original scales were correlated with two criterion variables: externalizing and internalizing behavior, as measured by a composite score of mothers' and fathers' ratings on the Child Behavior Checklist total externalizing and total internalizing scales (Achenbach & Ruffle, 2000)[5]. Finally, after each individual scale and dimension was considered, exploratory factor analyses were conducted to investigate the dimensional structure of the revised and original CBQ.

All IRT models were run using the flexMIRT software and estimated via full information maximum likelihood (Cai, 2012). All IFAs, bi-factor models, and comparison/follow-up analyses were conducted using Mplus version 8.0 (Muthen & Muthen, 1998-2017) and estimated either via mean and variance adjusted weighted least squares (WLSMV; for the categorical latent variable models; Wirth & Edwards, 2007), or full information maximum likelihood (FIML; for the comparisons and follow up analyses).

[5] This data was collected around the same time as the CBQ data in all samples

3. RESULTS

Results for each individual scale are presented in turn, with the order of presentation following the canonical dimensional structure. The results for the overarching dimensions themselves are presented following the discussion of each scale that is used in the computation of that dimensional composite. In the final section, the dimensional structure of the entire CBQ is evaluated. Throughout, the discussion of information is restricted to the area between two standard deviations below the mean (-2.0) to two standard deviations above the mean (2.0). This range was selected as it includes the majority of children, and it is often the case that the least information is provided at the extreme ends of the theta (i.e., more than 2 standard deviations from the mean). Though precise measurement across the full spectrum of theta is desirable, given the small number of children in the extreme tail ends, and the typical reduction in precision around those levels, information beyond 2 standard was not counted against scales (precision at these levels is however included in the computation of the marginal reliability index).

3.1 Effortful Control

3.1.1 Attentional Focusing

The original 9-item attentional focusing scale consistently provided slightly less than 4.00 logits of information (Table 1; $\alpha = .73$). The dimensionality analysis suggested that a three factor solution was optimal (Table 2), though these factors were conceptually ambiguous. Four items were subsequently trimmed from this scale for either inadequate initial discrimination (Items 160 and 186), failing to load on any factor in the dimensionality assessment (Item 186), and/or being weakly associated with the general factor in the bi-factor analysis (Items 125 and 144).

The revised 5-item attentional focusing scale's information curve was similar to the original's (see Table 3; $\alpha = .73$). All items had discrimination values above .60. The revised scale

was slightly less reliable when paternal ratings were used, generally providing between 3.15 and 3.30 logits of information (α = .70). The original and revised scales correlated at *r* = .85 with maternal reports, and *r* = .82 with paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's q^6 = .07; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's *q*s ≤ .12; Table 12).

3.1.2 Attentional Shifting

The original 5-item attentional shifting scale consistently provided slightly less than 3.00 logits of information (Table 2; α = .66). The dimensionality analysis suggested that a single factor solution was optimal (Table 2). Given the original length and relatively low amount of information provided, only one item (Item 180) was removed, for demonstrating inadequate initial discrimination.

The revised 4-item attentional shifting scale provided a similar amount of information as the original (Table 2; α = .66). All items had discrimination values above .60. The revised scale was slightly less reliable when paternal ratings were used (Table 3; α = .62). The original and revised scales correlated at *r* = .87 for maternal reports, and *r* = .85 for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's *q* =.04; Table 12). The revised scales were also similarly predictive of both

[6] Cohen's q represents the difference between two correlations that have been converted into Z-scores. Conventional standards typically hold that a *q* of .20 represents a small effect.

externalizing and internalizing problems relative to the original scales (Cohen's *q*s ≤.07; Table 12).

3.1.3 Inhibitory Control

The original 13-item inhibitory control scale consistently provided between 4.99 and 5.39 logits of information (Table 2; α = .81). The dimensionality analysis suggested that a two factor solution was optimal (Table 2). The first factor centered on the ability to follow directions, while the second captured the ability to override impulses. Four items were subsequently trimmed from this scale for either inadequate initial discrimination (Items 63, 116, and 162), failing to load on any factor in the dimensionality assessment (Item 63 and 185), and/or being weakly associated with the general factor in the bi-factor analysis (Items 116, 162).

The revised 9-item inhibitory control scale consistently provided slightly less than 5.00 logits of information (Table 3; α = .79). All items had discrimination values above .60. The revised scale was somewhat less reliable when paternal ratings were used, providing between 4.01 and 4.32 logits of information (Table 3; α = .76). The original and revised scales correlated at *r* = .94 for maternal reports, and *r* = .95 for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's *q* = .04; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's *q*s ≤ .04; Table 12).

3.1.4 Low Intensity Pleasure

The original 13-item low intensity pleasure scale provided the most (5.86) information at -2.0, but information declined as scores increases, dropping to 2.97 by 2.0 (Table 2; α = .80). The dimensionality analysis suggested that a three factor solution was optimal, though these factors

did not appear to be characterized by any meaningful conceptual distinctiveness (Table 2). Seven items were subsequently trimmed from this scale for either inadequate initial discrimination (Items 12, 36, and 86), failing to load on any factor in the dimensionality assessment (Items 12 and 111), and/or being only weakly associated with the general factor (Items 36, 86, 113, 164, and 174). Though item 113's I-ECV was above 35% (Table 2), it was still relatively low (46%), and factor one remained adequately represented with its removal.

The revised 6-item low intensity pleasure scale provided information in a similar pattern as the original, ranging from 4.11 logits (-2.00) to 1.93 logits (2.00) (Table 3; $\alpha = .71$). All items had discrimination values above .60. The revised scale provided slightly more information when paternal ratings were used, especially above the mean (Table 3; $\alpha = .72$). The original and revised scales correlated at $r = .78$ for maternal reports, and $r = .75$ for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's $q = .06$; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs $\leq .12$; Table 12).

3.1.5 Perceptual Sensitivity

The original 12-item perceptual sensitivity scale provided exceptionally high levels of information (up to over 20 logits) from -2.0 to 1.0, before dropping precipitously (to 3.13 logits) by 2.0 (Table 2; $\alpha = .92$). This dramatic information function is largely due to the inclusion of item 65, which had a discrimination value of 7.87. Discrimination values of this extreme magnitude however are typically indicative of estimation issues, not exemplary item quality. Thus, this item was not included in subsequent analyses. The dimensionality analysis suggested

that a two factor solution was optimal (Table 2). Factor 1 seemed to reflect the tendency to comment on novel, social-related stimuli, while factor 2 captured the acknowledgement of more subtle stimuli. Seven items were removed from this scale for either being associated with estimation problems (Item 65), inadequate initial discrimination (Items 9, 52, 142, and 154), and/or being only weakly associated with the general factor (Items 9, 52, 84, 105, 142, and 154).

The revised 5-item perceptual sensitivity scale provided the least information at -2.0 (2.82 logits), but then provided consistently adequate amounts of information (between 4.87 and 5.25 logits) from -1.00 to 2.00 (Table 3; $\alpha = .80$). All items had discrimination values above .60. The revised scale provided slightly more information when paternal ratings were used below -1.00, but less information at -1.00 and above (Table 3; $\alpha = .76$). The original and revised scales correlated at $r = .87$ for maternal reports, and $r = .85$ for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's $q = .05$; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs $\leq .08$; Table 12).

3.1.6 Effortful Control Composite

The five original and revised Effortful Control scales demonstrated similar patterns of intercorrleation (Table 4), though the correlations between scales were slightly weaker when the revised versions were considered. Effortful Control dimensional composites were created by averaging together the original (52 items total) and revised (29 items total; Table 17) scales. The original dimension consistently provided around 10 logits of information when maternal reports were used (Reliability of .90), and 9 logits of information when paternal reports were used (Reliability of .89). The revised dimension consistently provided around 8 logits of information

when maternal reports were used (Reliability of .88), and 7.5 logits of information when paternal reports were used (Reliability of 87).The original and revised dimensions correlated at r = .92 for maternal reports, and r = .89 for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised dimensions (Table 11). Parental agreement was similar across the revised and original dimensional composites (Cohen's q = .05; Table 12). The revised dimensions were also similarly predictive of both externalizing and internalizing problems relative to the original dimensions (Cohen's qs ≤ .10; Table 12). Overall, the length of the Effortful Control dimension was reduced by 44% considering all component scales (Table 17).

3.2 Negative Affectivity

3.2.1 Anger/Frustration

The original 13-item anger/frustration scale provided the most information at -2.0 (5.63 logits), with information values decreasing gradually as scores increased (4.59 at 2.00) (Table 5; α = .81). The dimensionality analysis suggested a three factor solution was optimal (Table 5). The first factor was the largest and represented the general anger/frustration construct, while the second factor emphasized bed-time related anger, and the third captured frustration with other children and failure. Five items were subsequently removed from this scale for inadequate initial discrimination (Item 156), failing to load on any factor in the dimensionality assessment (Item 73), being weakly associated with the general factor (Item 120 and 173), or demonstrating consistent mediocrity (for the sake of reducing length; Item 140). The bi-factor model with both minor factors encountered convergence problems, and so only the 3 item minor factor was included. Thus, for the two item minor factor, the item with the highest discrimination value in

the initial GRM was retained in the revised scale (i.e., higher initial discrimination implies a stronger relation to the general factor).

The revised 8-item anger/frustration scale's information curve followed a pattern similar to the original's, falling from 5.19 logits at -2.00 to 3.79 logits at 2.00 (Table 5; $\alpha = .79$). All items had discrimination values above .60. The revised scale provided slightly less information when paternal ratings were used (Table 6; $\alpha = .79$). The original and revised scales correlated at $r = .93$ for maternal reports, and $r = .92$ for paternal reports (Table 11). For both parental informants, means were smaller, and standard deviations were larger, in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's $q = .13$; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs $\leq .06$; Table 12).

3.2.2 Discomfort

The original 12-item discomfort scale consistently provided above 5.00 logits of information, though precision was strongest between -1.00 and 1.00 (Table 5; $\alpha = .86$). The dimensionality analysis suggested that a three factor solution was optimal (Table 5). The three factors appeared to tap into physical pain (factor 1), overstimulation (factor 2), and the propensity to explicitly verbalize discomfort (factor 3). Six items were subsequently removed from this scale for either inadequate initial discrimination (Items 87, 115, 141, and 157), failing to load on any factor in the dimensionality assessment (Items 73, 115, 141, and 157), being weakly associated with the general factor (Items 87 and 190), and/or being the weakest item in a generally strong sub-factor for the sake of reducing length (Item 5). In the bi-factor model factor

3 could not be included without encountering convergence problems. Thus, for factor 3, the item with the highest initial discrimination value was retained in the revised scale.

The revised 6-item discomfort scale provided over 5.00 logits of information until around 2.0 (Table 5; α = .86). Three items (Items 21, 97, and 178) had discrimination values below.60. These items were retained to ensure that each factor was represented, and these discrimination values imply an enduring distinctiveness between factors. The revised scale provided slightly less information when paternal ratings were used (Table 6; α = .80). The original and revised scales correlated at r = .90 for maternal reports, and r = .89 for paternal reports (Table 11). For both parental informants, means were smaller, and standard deviations were larger, in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's q =.03; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs ≤ .06; Table 12).

3.2.3 Soothability

The original 13-item soothability scale provided the most information (up to 7.00+ logits) between -1.0 and 1.0 (Table 5; α = .85). The dimensionality analysis suggested that a three factor solution was optimal (Table 5). The three factors appeared to capture calming down after an exciting activity (factor 1), propensity to be soothed (factor 2), and how quickly it is possible to be soothed (factor 3). Six items were subsequently removed from this scale for either inadequate initial discrimination (Items 14, 27, 42, 85, and 103), failing to load on any factor in the dimensionality assessment (Items 85, 103, and 167), and/or being weakly associated with the general factor (Items 14 and 27).

The revised 7-item soothability scale generally provided at least 6.00 logits of information, though less information was provided around -2.00 (Table 6; α = .85). Three items (Items 53, 92, and 118) had discrimination values below .60. The revised scale was slightly less reliable with paternal ratings from -1.00 to 2.00, but more reliable at -2.00 (α = .86). The original and revised scales correlated at *r* = .90 for maternal reports, and *r* = .90 for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's *q* =.03; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's *q*s ≤ .10; Table 12).

3.2.4 Fear

The original 12-item fear scale provided the most information (up to 7.00+ logits) between -1.00 and 1.00 (Table 5; α = .86). The dimensionality analysis suggested that a two factor solution was optimal (Table 5). The first factor captured fear related to darkness and sleep, while the second factor captured fear of potentially dangerous or startling stimuli. Five items were subsequently removed from this scale for either inadequate initial discrimination (Items 15, 58, 138, 161, and 189), failing to load on any factor in the dimensionality analysis (Items 58, 138, and 189), and/or being weakly associated with the general factor (Items 15 and 161).

The revised 7-item fear scale provided the most information (up to 9.00+ logits) between -1.00 and 1.00 (Table 6; α = .87). Two items (Items 50 and 80) had discrimination values below .60. The revised scale was similarly informative when paternal ratings were used (Table 5; α = .87). The original and revised scales correlated at *r* = .89 for maternal reports, and *r* = .90 for paternal reports (Table 11). For both parental informants, means were smaller, and standard

deviations were larger, in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's q =.04; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs ≤ .05; Table 12).

3.2.5 Sadness

The original 12-item sadness scale consistently provided around 3.30 logits of information (Table 5; α = .70). The dimensionality analysis suggested that a two factor solution was optimal (Table 5). Most items loaded on the first factor, which appeared to represent the general sadness construct. The second factor centered on sadness catalyzed by stories or television shows. Two items were subsequently removed from this scale for inadequate initial discrimination (Item 109), failing to load on any major factor (Item 149), and/or being weakly associated with the general factor in the bi-factor model (Item 109). Only two items were removed given the low level of initial information, and exploratory analyses suggesting that removing more than these two items would cause an unacceptable drop in precision.

The revised 10-item sadness scale consistently provided around 3.20 logits of information (Table 6; α = .69). Two items (Items 112 and 127) had discrimination values below .60. The revised scale was slightly less informative when paternal ratings were used (Table 6; α = .68). The original and revised scales correlated at r = .97 for maternal reports, and r = .96 for paternal reports (Table 11). For both parental informants, means were slightly smaller, and standard deviations were larger, in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's q =.05; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs ≤ .04; Table 12).

3.2.6 Negative Affectivity Composite

The five original and revised Negative Affectivity scales demonstrated similar patterns of intercorrleation (Table 7), though the correlations between scales were slightly weaker when the revised versions were considered. Negative Affectivity dimensional composites were created by averaging together the original (62 items total) or revised (38 items total; Table 17) scales. The original dimension consistently provided around 10.5 logits of information when either maternal or paternal reports were used (Reliability of .90). The revised dimension consistently provided around 8.5 logits of information when either maternal or paternal reports were used (Reliability of .88). The original and revised dimensions correlated at $r = .96$ for both and paternal reports (Table 11). For both parental informants, means were slightly smaller, and standard deviations were larger, in the revised dimensions (Table 11). Parental agreement was similar across the original and revised dimensional composites (Cohen's $q = .04$; Table 12). The revised and original dimensions were also similarly predictive of both externalizing and internalizing problems (Cohen's qs $\leq .04$; Table 12). Overall, the length of the Negative Affectivity dimension was reduced by 39% considering all component scales (Table 17).

3.3 Surgency

3.3.1 Activity Level

The original 13-item activity level scale consistently provided around 6.00 logits of information above -1.0 (Table 8; $\alpha = .83$). The dimensionality analysis suggested that a three factor solution was optimal (Table 8). Factor one captured the tendency to be in a hurry, while factors 2 and 3 included the various reverse scored items that center on sitting quietly, and moving slowly. Six items were subsequently removed from this scale for inadequate initial discrimination (Items 1, 48, 126, and 187), failing to load on any factor in the dimensionality

analysis (Items 48 and 187), being weakly associated with the general factor (Items 1, 88, and 126), and/or being the weakest item in a generally strong sub-factor for the sake of reducing length (Item 41).

The revised 7-item activity level scale consistently provided between 5.00 and 6.25 logits of information above -1.0 (Table 9; $\alpha = .82$). No items had a discrimination value below .60. The revised scale was similarly informative when paternal ratings were used, though paternal ratings provided slightly more information below the mean ($\alpha = .82$; Table 9). The original and revised scales correlated at $r = .89$ for maternal reports, and $r = .88$ for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was largely similar in the revised scales compared to the original scales, though the revised scales had a small advantage (Cohen's $q = .20$; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs $\leq .07$; Table 12).

3.3.2 High Intensity Pleasure

The original 13-item high intensity pleasure scale consistently provided around 6.00 logits of information until above 1.0, where it fell to 3.82 by 2.00 (Table 8; $\alpha = .82$). The dimensionality analysis suggested a two factor solution was optimal, though only two items loaded onto the second factor (Table 8). These two items specifically centered on children's enjoyment of slides. Five items were subsequently removed from this scale for either inadequate initial discrimination (Items 77, 107, 159 and 182), failing to load on any factor in the dimensionality assessment (Items 77, 107, 159, and 182), and/or being weakly associated with the general factor (Item 8). Although item 51 did not load on any factor in the IFAs, it was retained to ensure an adequate degree of reliability given its initial discrimination (and the fact

that it did load substantially in the IFA if a one factor solution was imposed on the data). Notably, the bi-factor model did not converge for this scale, and thus the item from the 2 item sub-factor that demonstrated the most initial discrimination was retained for the revised scale.

The revised 8-item high intensity pleasure scale provided at least 4.00 logits of information below 1.0; above 1.0 the revised scale provided slightly less than 4.00 logits (Table 9; $\alpha = .74$). No items had discrimination values below .60. The revised scale was slightly more informative when paternal ratings were used (Table 8; $\alpha = .75$). The original and revised scales correlated at $r = .93$ for maternal reports, and $r = .93$ for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's q =.15; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs $\leq .04$; Table 12).

3.3.3 Impulsivity

The original 13-item impulsivity scale consistently provided between 5.00 and 5.50 logits of information (Table 8; $\alpha = .81$). The dimensionality analysis suggested that a 3 factor solution was optimal (Table 8). The first factor centered on planning ahead, the second factor captured the propensity to approach novel stimuli, and the third factor captured impulsive behaviors. Six items were subsequently removed from this scale for either inadequate initial discrimination (Items 26, 114, 137, and 155), failing to load on any factor in the dimensionality assessment (Items 26, 90, 114, and 137), and/or being weakly associated with the general factor in the bi-factor model (Items 104 and 155). Although item 46 did not load on any major factor in the IFAs, it was retained to maintain an adequate level of information given its initial discrimination.

Further, though both items of factor 2 demonstrated low I-ECV values, these items also had high initial discrimination values, and contributed to maintaining the conceptual makeup of the scale.

The revised 7-item impulsivity scale consistently provided between 4.50 and 5.00 logits of information (Table 9; α = .80). No items had discrimination values below .60. The revised scale was somewhat less informative when paternal ratings were used (Table 9; α = .75). The original and revised scales correlated at r = .90 for maternal reports, and r = .89 for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was equivalent for the revised and original scale (Cohen's q =.00; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs ≤ .10; Table 12).

3.3.4 Positive Anticipation

The original 13-item positive anticipation scale provided the most information at -2.00 (5.28 logits), with values decreasing as scores increased to 3.75 logits by 2.00 (Table 8; α = .79). The dimensionality analysis suggested that a four factor solution was optimal, though the 4 factors were not clearly distinguishable conceptually (Table 8). Six items were subsequently removed from this scale for either inadequate initial discrimination (Items 69, 175, and 188), failing to load on any factor in the dimensionality assessment (Items 69, 175, 188, and 191), being weakly associated with the general factor in the bi-factor model (Item 35), and/or displaying consistent estimation issues across analyses (Item 82). The only item that was associated with factor 3 was included in the revised scale. Further, although no item from factor 1 had an I-ECV value above 35%, the items with the two highest values were retained for adequate representation of the factor.

The revised 7-item positive anticipation scale had provided between 4.20 and 5.50 logits of information until after 1.0 (Table 9; α = .78). One item (Item 10) had a discrimination value below .60; this item was one of the items with a low I-ECV value from factor 1. The revised scale was slightly less informative when paternal ratings were used until the higher range of the trait, where it provided more information (α = .78; Table 9). The original and revised scales correlated at *r* = .88 for maternal reports, and *r* = .88 for paternal reports (Table 11). For both parental informants, means were slightly smaller, and standard deviations were larger, in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's *q* =.13; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's *q*s ≤ .04; Table 12).

3.3.5 Shyness

The original 13-item shyness scale provided consistently high levels of information (8.50+ logits), though relatively less information was provided above 1.0 (see Table 8; α = .92). The dimensionality analysis suggested that a two factor solution was optimal, though the two factors demonstrated considerable conceptual overlap (Table 8). Six items were subsequently removed from this scale. Importantly, this scale was particularly strong, with only one or two items meeting the initial criteria for removal. Thus, the items that were removed tended to be the weakest from an altogether strong set. Items were thus removed for either relatively low initial discrimination (Items 7, 89, and 143), failing to load on any factor in the dimensionality assessment (Items 37 and 119), and/or being worded similarly to other items (Item 129).

The revised 7-item shyness scale provided above 9.00 logits of information between -2.0 and 1.0, before dropping to 5.25 logits by 2.00 (Table 9; α = .89). No items had discrimination

values below .60. The revised scale was slightly less informative when paternal ratings were used (α = .86; Table 9). The original and revised scales correlated at r = .96 for maternal reports, and r = .95 for paternal reports (Table 11). For both parental informants, means and standard deviations were larger in the revised scales (Table 11). Parental agreement was similar in the revised scales compared to the original scales (Cohen's q =.03; Table 12). The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's qs ≤ .07; Table 12).

3.3.6 Smiling and Laughter

The original 13-item smiling and laughter scale provided above 6.00 logits of information until around 1.0, dropping to 3.16 logits by 2.00 (Table 8; α = .86). The dimensionality analysis suggested that a two factor solution was optimal, though these factors showed considerable conceptual overlap (Table 8). Six items were subsequently removed from this scale for inadequate initial discrimination (Items 43 and 83), failing to load on any factor in the dimensionality analysis (Items 83, 121, and 179), being weakly associated with the general factor in the bi-factor model (Items 43 and 152), and/or demonstrating consistent estimation issues across analyses (Item 56). Although most items demonstrated low I-ECV values, the high degree of overlap in content between the items of both factors suggest that this scale can be treated as effectively unidimensional. Given this, though item 152 did not have the lowest I-ECV, it was still removed for the sake of parsimony as it had a fairly low initial discrimination value.

Like the original, the revised 7-item smiling and laughter scale provided a substantial amount of information –above 6.00 logits – up to 1.0, but then dropped to 2.64 by 2.0 (Table 9; α = .86). One item (Item 11) had a discrimination values below .60. The revised scale was slightly

less informative when paternal ratings were used until the higher range of the theta, where it was more informative (α = .86; Table 9). The original and revised scales correlated at *r* = .92 for maternal reports, and *r* = .90 for paternal reports (Table 11). For both parental informants, means were slightly smaller, and standard deviations were larger, in the revised scales (Table 11). Parental agreement was largely similar in the revised scales compared to the original scales (Cohen's *q* =.22; Table 12), though the revised scale had a small advantage. The revised scales were also similarly predictive of both externalizing and internalizing problems relative to the original scales (Cohen's *q*s ≤ .10; Table 12).

3.3.7 Surgency Composite

The five original and revised Surgency scales demonstrated similar patterns of intercorrleation (Table 10), though the correlations between scales were slightly weaker when the revised versions were considered. Surgency dimensional composites were computed by averaging together either the original (65 items total) or revised (43 items total) scales (Table 17). The original dimension consistently provided around 20 logits of information when either maternal or paternal reports were used (Reliability of .95). The revised dimension consistently provided around 15 logits of information when maternal reports were used (Reliability of .93), and 16 logits of information when paternal reports were used (Reliability of .94). Notably, in both the original and revised dimensions information began to decrease above 1.0, however values remained over 10 logits. The original and revised dimensions correlated at *r* = .97 for maternal reports, and *r* = .97 for paternal reports (Table 11). For both parental informants, means were equivalent, and standard deviations were larger, in the revised dimensions (Table 11). Parental agreement was also equal across the revised the original dimensions (Cohen's *q* =.00; Table 12). The revised and original dimensions were also similarly predictive of both

externalizing and internalizing problems (Cohen's *q*s ≤ .02; Table 12). Overall, the length of the Surgency dimension was reduced by 45% considering all component scales (Table 17).

3.4 Exploratory Factor Analyses

The overarching dimensional structure of the original and revised CBQ was analyzed via EFA (with an oblique, geomin rotation). Maternal and paternal reports were both considered, and results were generally similar across each parental informant (congruence coefficients across informants for the factor solutions discussed below ranged from .87 to .98; Table 16). For both maternal and paternal reports, a 3 factor solution for the original scales appeared optimal, being supported by both an examination of the scree plot and a parallel analysis (Table 15). Conversely, a 4 factor solution appeared to be optimal for both the maternal and paternal revised scales based on an examination of the scree plot and a parallel analysis (Table 15). To further examine the breakdown in factor composition, 3 and 4 factor solutions were extracted for both the original and revised scales. The factor loadings from these solutions are presented in tables 13 and 14.

The 3 factor solution in the original scales was generally consistent with the typical dimensional structure used in the literature (Rothbart et al., 2001). However, the attentional shifting scale failed to load substantially on any factor. Further, the sadness scale loaded most strongly on the Effortful Control factor, as did the smiling and laughter scale. The anger/frustration scale also loaded to a similar degree on both the Negative Affectivity and Surgency factors. Factor intercorrelations were of a trivial magnitude (< .20; Table 16). The 3 factor solution was less consistent with the typical framework when the revised scales were used. Factor one appeared to capture negative emotionality and distractibility, while factor 2 captured impulsivity/reticence, and factor 3 captured general positive emotionality. Inhibitory control, low

intensity pleasure, and perceptual sensitivity all failed to load on any factor. Again, factor intercorrelations were small (<.20; Table 16).

The 4 factor solution in the original scales was similar to the 3 factor solution, however what was the Surgency factor split into two factors. One was more centered on positive emotionality, while the other focused on impulsivity. Both the attentional shifting and high intensity pleasure scales failed to load on any factor in this solution. Most factor intercorrelations were small, but the 2 Surgency factors were correlated around r = .60 (Table 16). With the revised scales, the first factor included all the Effortful Control scales except inhibitory control (or the attentional shifting scale with paternal reports), which did not load on any factor. The second factor included most of the Negative Affectivity scales. However, the anger/frustration scale also loaded almost as strongly on the Effortful Control factor, and the sadness scale instead loaded on factor 3. Factor 3 resembled a positive emotionality dimension, including sadness (reversed), activity level, positive anticipation, and smiling and laughter. The fourth factor again appeared to represent an impulsivity/reticence dimension, including both impulsivity and shyness. The high intensity pleasure scale failed to load on any major dimension. Factor intercorrelations were mostly small, but the positive emotionality and impulsivity factors were correlated to a moderate degree (r ~ .35; Table 16).

4. SUMMARY AND CONCLUSIONS

Item response theory (IRT) and related categorical latent variable modeling techniques were applied to the 195 item[7] Child Behavior Questionnaire (CBQ; Rothbart et al., 2001). This was done in an attempt to both evaluate the CBQ, and improve its efficiency. That is, the CBQ was edited to reduce length while preserving the favorable measurement qualities of the original form as much as possible. Thus, the aim was not to develop another short form of the CBQ (Putnam & Rothbart, 2006) per se, but rather to revise the standard form with an eye toward maximizing efficiency. This allows researchers and practitioners to take advantage of the strengths of the original form while simultaneously reducing participant burden. Indeed, more contemporary psychometric techniques such as those used here have made it easier than ever to pinpoint weak and/or unnecessary items and fine-tune assessments to achieve the desired goal in as brief a form as possible.

4.1 Summary

The CBQ contains 16 individual scales, and the initial IRT models demonstrated that these scales ranged in psychometric quality from somewhat weak (e.g., sadness) to quite strong (e.g., shyness). Regardless of initial quality, however, all scales contained items that were unnecessary (i.e., that contributed little useful information) and could be removed without undercutting the functioning of the scale to a practically significant degree. Importantly, throughout the process of revising the scales weight was given to both general psychometric performance, as well as content coverage. When scales demonstrated multidimensionality, as most did, care was taken to ensure that each dimension was represented in the revised scale, even

[7] Although the CBQ contains 195 items, only 192 items were actually included in the analyses here. This is because items 3, 33, and 49 are not incorporated into any scale in the CBQ scoresheet. As such, these three items have been trimmed by default.

if the best representation of a dimension was not a particularly strong item overall. Overall, this approach helped ensure that both the psychometric *and* substantive strengths of the original CBQ were maintained. Altogether, the length of the CBQ was reduced by 44% (Table 17). Despite this considerable reduction in length, however, the revised scales and dimensions functioned very similarly when compared to the original scales.

The original CBQ scales typically managed to provide at least 4.00 logits of information (i.e., Reliability of .75) across much, if not all, of the range of theta considered (2 standard deviations above and below the mean). Some did not provide this much information, however (e.g., attentional shifting, sadness). Further, it was not uncommon for less, sometimes much less, information to be provided at the more extreme ends of theta. That is, the original CBQ scales tended to be less precise in making distinction between children at higher and lower levels of theta (i.e., 2 standard deviations out). Practically speaking, this means that although the CBQ is generally capable of identifying children who are particularly high or low on some trait, among these children, more precise fine grained distinctions are difficult. In general, however, the original CBQ scales were fairly reliable by both conventional standards, and the standards applied in this study.

As would be expected given the amount of items removed, the revised scales were somewhat less informative than the original scales. However, these differences tended to be quite modest. Indeed, considering marginal reliability as a holistic representation of information across theta, reductions in α ranged from only .00 to .09 (.12 if perceptual sensitivity is included, but the aforementioned estimation issues make the original reliability estimate questionable), with an average difference of .02. Thus, the reduced length of the revised scales does not appear to entail a meaningful sacrifice of precision. To be sure, examining the information values in isolation

ostensibly imply a more substantial drop in precision on occasion (e.g., shyness). However, as noted, there are diminishing returns on reliability as information values increase; as information values increase, greater amounts of information do not tend to provide substantially more precision. Accordingly, larger drops in information per se do not necessarily correspond to analogously large drops in reliability. The information curves of the revised scales also tended to mirror the shape of the original, and thus provided less information at the extreme ends of theta. In general, paternal reports provided slightly less information than the maternal reports. However, the differences in precision across maternal and paternal ratings in the revised scales were again quite small, with differences in α ranging from .00 to .06 (average difference = .01). Furthermore, paternal reports tended to demonstrate a slight advantage in measuring the extreme ends of theta over maternal reports.

The original and revised scales tended to correlate quite strongly (Table 11), with average correlations of r = .90 for maternal reports and r = .89 for paternal reports. This suggests that the rank ordering of children on the temperament scales was largely preserved. Interestingly, the means of the revised scales tended to be slightly higher than the original scale means for "positively valenced" scales (e.g., inhibitory control), and slightly lower for "negatively valenced" scales (e.g., anger/frustration). The revised scales also universally demonstrated greater variability than the original scales (Table 11). Both trends are likely in part due to the nature of the items that were removed. That is, for many of the weakest items parental responses clustered around only one or two response options. For example, with item 26 ("sometimes interrupts others when they are speaking"; positive anticipation), approximately 70% of all responses were either in category 5 or 6. This restriction of range implies that these items are less capable of discriminating between high and low ranking children, and that these items will also

push means up or down while reducing variability. Therefore, removing these items removed a press that was pushing all children in a certain direction, which increases variability.

Agreement between mothers and fathers on the CBQ was also similar across the 16 original and revised scales. The average inter-parent correlation for the original scales was r = .50, while the average for the revised scales was r = .51. Furthermore, the average difference in correlations between the original and revised scales was only .06. Overall then, the revised scales tended to preserve the degree of inter-parent agreement found in the original scales.

The original and revised scales were also similarly predictive of externalizing and internalizing problems. The average correlation between the original scales and externalizing problems was r = .19 for maternal reports and r = .16 for paternal reports. On the other hand, the average correlation between the revised scales and externalizing problems was r = .21 for maternal reports and r = .18 for paternal reports. For internalizing problems, the average correlation with the original scales was r = .12 for maternal reports and r = .13 for paternal reports, while the average correlation with the revised scales was r = .14 for maternal and paternal reports. The average differences in correlations were .06 and .05 for maternal and paternal reports with externalizing problems, and .04 and .05 for maternal and paternal reports with internalizing problems. Thus, the revised and original scales were similarly related to different facets of child adjustment.

The largest discrepancy between the original and revised scales appeared when analyzing the underlying dimensional structure of the entire questionnaire. The EFA of the original scales largely supported the canonical three factor structure, whereas the EFA of the revised scales was slightly more ambiguous. A 4 factor solution appeared optimal, with the Surgency scales loading on two distinct factors, and one factor representing a blend of Effortful Control and Negative

Affectivity. To be sure, the revised scales tended not to correlate as strongly with each other as the original scales did. However, it is worth noting that the traditional 3 factor structure is not universally supported in the original CBQ (e.g., Kotelnikova et al., 2015). Furthermore, as was the case in the present study, even when the 3 major dimensions do largely emerge there are often several prominent cross-loadings, or unexpected major loadings (e.g., smiling and laughter on the Effortful Control dimension; Clark et al., 2016; Rothbart et al., 2001). Accordingly, the traditional 3 factor structure may be more conceptual useful than empirically robust. Indeed, despite these discrepancies, the dimensional composites based on the revised scales still performed similarly to the composites based on the original scales. That is, the revised dimensions provided substantial information, correlated highly with the original dimensions, contained more variance, demonstrated similar inter-parent agreement relative to the original scales, and predicted externalizing and internalizing problems to a similar magnitude as the original scales. Thus, despite the more ambiguous underlying factor structure, the revised dimensional composites appear to be functionally equivalent to original dimensional composites.

4.2 Implications

The standard form CBQ provides considerable information about individual differences in child temperament to both researchers and practitioners. However, its length may be prohibitive under several circumstances, especially when parents are expected to fill out many questionnaires at once, or there are external time pressures. This study has demonstrated that the length of the CBQ can be reduced by almost half without compromising its major strengths. Indeed, despite its considerably shorter length, the revised form was functionally very similar to the original form. The procedure used here therefore appears to have primarily identified particularly weak and/or redundant items, that is, the superfluous items. Accordingly, it is

important to re-emphasize that the revised form here is not meant as a new short form CBQ. Rather, by identifying the items that provide little to no conceptual or psychometric benefit, the present findings can be used as a basis of a more efficient revised CBQ standard form (i.e., CBQ-R) that also happens to be shorter.

To be sure, a specific short-form CBQ has already been developed using classical approaches (Putnam & Rothbart, 2006). This form has a total of 94 items compared to the revised form's 110 (106 without attentional shifting, which is not included in the short form). In the current sample, the average marginal reliability of the short form scales was $\alpha = .76$ compared to an average marginal reliability of $\alpha = .80$ for the revised scales, and an average marginal reliability of $\alpha = .82$ for the original scales. Thus, on average, the revised CBQ is only marginally less reliable than the original form, while being only marginally longer than the short form. If the revised CBQ is adopted as the new standard form given its general maintenance of the original's desirable properties, the development of a new, shorter short form based on the revised CBQ may also be justified (or simply a refinement of the Very Short Form).

This could also be done by taking advantage of the techniques used here. Indeed, the present study also highlights the advantages of Item Response Theory modeling and related latent variable modeling techniques for fine-tuning psychological assessments. Although the original full and short form, developed using more traditional, classic approaches, have proved useful to researchers and practitioners, the advantages of more contemporary approaches are demonstrated here. That is, these methods allow test developers to take a more nuanced approach to both item and scale functioning to craft the most efficient questionnaire possible for the goal at hand. Although in the classical test framework more items tend to be more unequivocally associated with better measurement (e.g., coefficient alpha is partly a function of scale length),

the present approach facilitates a more nuanced approach to test length. Items may contribute little to the measurement of the construct, or may be unnecessary given the diminishing returns of information. This can then be balanced with substantive concerns regarding content coverage. The present study illustrated that it was possible to balance concerns regarding psychometric functioning, length, and content coverage to derive a much more efficient, but similarly comprehensive, CBQ.

4.3 Limitations and Future Directions

There are some limitations to this study that must be acknowledged. Most important is the absence of a truly independent validation sample. Here, paternal ratings were examined after performing the primary analyses and scale reduction using maternal ratings. However, both informants were reporting on the same children, which drastically undermines the claim that these represent independent samples. Of course, one advantage of this approach was the ability to examine inter-parent agreement, which is often an issue of substantive interest (Clark et al., 2017; De Los Reyes & Kazdin, 2005). Along these lines, though the present sample was relatively sizeable with over 600 children, given the complexity of many of the models used it may have been preferable to have an even larger sample. To be sure, this could have helped in the occasions noted above where convergence issues were encountered. Preliminary simulations based on the current sample size and reasonable population values (taken from earlier analyses) however suggested that the present sample was generally adequate for the aims of the study (i.e., estimates were generally unbiased and precise). Finally, the only external criterion variables available were the CBCL externalizing and internalizing scores. To more fully evaluate the functioning of the revised scales it would have been beneficial to have included a more extensive

network of criterion variables, including alternative approaches to assessing temperament (e.g., laboratory assessments)

Future work in this vein should begin by further validating the revised scales in truly independent samples. Furthermore, the relations between the original and revised CBQ, and other external criterion, can be more fully explored. One particularly important direction in this vein may be to examine associations with observational assessments of temperament. Parent report and laboratory assessments of temperament tend not to overlap too highly (Clark et al., 2017). Conceptually this may be problematic as each method purports to measure the same constructs. Thus, it is critical to investigate whether the revised scales are related to laboratory assessments to a similar degree as the original, or if the revised scales actually confer an advantage given the elimination of weaker items, and the general increase in variability. Finally, it would be useful to more fully compare the short form scales with the original and revised scales. To be sure, the revised scales may undermine the advantages of the original short form; however, techniques such as those used here may be used to create a new short form based on the revised form. This could offer an even greater reduction in length, while attempting to minimize the blow to psychometric functioning.

4.4 Conclusion

The literature on child temperament is rapidly expanding as more and more researchers and practitioners come to appreciate the importance of early emerging individual differences in children's dispositions. Thus, it is more important than ever before that there are widely available tools for assessing temperament that are both conceptually and psychometrically robust, and practical. One of the most popular temperament inventories, the CBQ, is extremely comprehensive, but also quite lengthy. The current study used IRT and related techniques to

streamline the CBQ by identifying and eliminating weak or unnecessary items. The result is a more efficient, revised CBQ that functions very similarly to the standard form, but is considerably shorter. This revised CBQ thus provides the advantages of the original form to researchers and practitioners, but with a substantial reduction in participant burden.

REFERENCES

Achenbach, T. M., & Ruffle, T. M. (2000). The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatrics in Review, 21*(1), 265-280.

Borsboom, D. (2009). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.

Cai, L. (2012). *flexMIRT: Flexible multilevel item factor analysis and test scoring* [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Carey, W. B., & McDevitt, S. C. (1989). *Clinical and educational applications of temperament research*. Amsterdam/Lisse: Swets & Zeitlinger.

Caspi, A., Moffitt, T. E., Newman, D. L., & Silva, P. A. (1996). Behavioral observations at age 3 years predict adult psychiatric disorders: Longitudinal evidence from a birth cohort. *Archives of General Psychiatry, 53*(11), 1033-1039.

Clark, D. A., Donnellan, M. B., Robins, R. W., & Conger, R, D. (2015). Early adolescent temperament, parental monitoring, and substance use in Mexican-origin adolescents. *Journal of Adolescence, 41,* 121-131. doi:10.1016/j.adolescence.2015.02.010

Clark, D. A., Durbin, C. E., Donnellan, M. B., & Neppl, T. K. (2017). Internalizing symptoms and personality traits color parental reports of child temperament. *Journal of Personality*. doi: 10.1111/popy.12293

Clark, D. A., Durbin, C. E., Hicks, B. M., Iacono, W. G., & McGue, M. (2017). Personality in the age of industry: Structure, heritability, and correlates of personality in middle childhood from the perspective of parents, teachers, and children. *Journal of Research in Personality, 67*, 132-143. doi: 10.1016/j.jrp.2016.06.013

Clark, D. A., Listro, C. J., Lo, S. L., Durbin, C. E., Donnellan, M. B., & Neppl, T. K. (2016). Measurement invariance and child temperament: An evaluation of sex and informant differences on the Child Behavior Questionnaire. *Psychological Assessment, 28*(12), 1646-1662. doi: 10.1037/pas0000299

Clark, L. A., & Watson, D. (2008). Temperament: An organizing paradigm for trait psychology. In *Handbook of Personality: Theory and Research,* (pp. 265-286). New York, NY: The Guilford Press.

Crede, M., Harms, P., Niehorster, S., Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology, 102*(4), 878-888.

Creemers, H. E., Dijkstra, J. K., Vollebergh, W. A. M., Ormel, J., Verhulst, F. C., & Huizink, A. C. (2010) Predicting life-time and regular cannabis use during adolescence; the roles of temperament and peer substance use: The TRAILS study. *Addiction, 105,* 699-708. doi: 10.1111/j.1360-0443.2009.02819.x

de Ayala, R. (2009). *The theory and practice of item response theory.* New York: Guilford Press.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*, 483-509. doi: 10.1037/0033-2909.131.4.483

Duckworth, A. L., & Allred, K. M. (2012). Temperament in the classroom. In M. Zentner, & R. L. Shiner (Eds.), *Handbook of Temperament* (pp. 607-626). New York: The Guilford Press.

Elder, G. H., & Conger, R. D. (2000). *Children of the Land: Adversity and success in rural*

America. Chicago, IL: University of Chicago Press.

Embretson S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlsbaum Associates, Inc.

Gartstein, M. A., Bridgett, D. J., & Low, C. M. (2012). Asking questions about temperament: Self- and other-report measures. In M. Zentner, & R. L. Shiner (Eds.), *Handbook of Temperament* (pp. 183-208). New York: The Guilford Press.

Goldsmith, H. H., Buss, A. H., Plomin, R., Rothbart, M. K., Thomas, A., Chess, S… McCall, R. B. (1987). Roundtable: What is temperament? Four approaches. *Child Development, 58*, 505-529.

Goldsmith, H., H., & Gagne, J. R. (2012). Behavioral assessment of temperament. In M. Zentner, & R. L. Shiner (Eds.), *Handbook of Temperament* (pp. 209-228). New York: The Guilford Press.

Goldsmith, H. H., Lemery, K. S., Buss, K. A., & Campos, J. J. (1999). Genetic analyses of focal aspects of infant temperament. *Developmental psychology*, *35*(4), 972-985.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Issues and applicants.* Boston: Kluwer Nijhoff.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS smoking item banks. *Nicotine & Tobacco Research, 16*(3), 175-189. doi: 10.1093/ntr/ntt123

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal, 15* (1), 136-153. doi: 10.1080/10705510701758406

Klein, D. N., Dyson, M. W., Kujawa, A. J., & Kotov, R. (2012). Temperament and internalizing disorders. In M. Zentner, & R. L. Shiner (Eds.), *Handbook of Temperament* (pp. 541-561). New York: The Guilford Press.

Kotelnikova, Y., Olino, T. M., Klein, D. N., Kryski, K. R., & Hayden, E. P. (2015). Higher- and lower-order factor analyses of the Children's behavior questionnaire in early and middle childhood. *Psychological Assessment.* doi: 10.1037/pas0000153

Lo, S. L., Vroman, L. N., & Durbin, C. E. (2014). Ecological validity of laboratory assessments of child temperament: Evidence from parent perspectives. *Psychological Assessment, 27*(1), 280-290. doi: 10.1037/pas0000033

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

McClowry, S. G., & Collins, A. (2012). Temperament-based intervention: Reconceptualized from a response-to-intervention framework. In M. Zentner, & R. L. Shiner (Eds.), *Handbook of Temperament* (pp. 607-626). New York: The Guilford Press.

Mervielde, I., & De Pauw, S. S. W. (2012). Models of Child Temperament. In M. Zentner, & R. L. Shiner (Eds.), *Handbook of Temperament* (pp. 21-40). New York: The Guilford Press.

Moffitt, T. E., Arsenault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H. … Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *PNAS, 108,* 2693-2698.

Muthén, L.K., & Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Neppl, T. K., Donnellan, M. B., Scaramella, L. V., Widaman, K. F., Spilman, S. K., Ontai, L. L.,

& Conger, R. D. (2010). Differential stability of temperament and personality from toddlerhood to middle childhood. *Journal of Research in Personality, 44*, 386-396. doi: 10.1016/j.jrp.2010.04.004

Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment, 87*(1), 103-113.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin*, *114*(3), 552.

Revicki, D. A., & Reise, S. P. (2015). Summary: New IRT problems and future directions. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling* (pp. 457-462). New York, NY: Taylor & Francis.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological methods*, *21*(2), 137.

Rothbart, M. K. (2007). Temperament, development, and personality. *Current directions in psychological science*, *16*(4), 207-212.

Rothbart, M. K. (2011). *Becoming who we are: Temperament and personality in development* New York, NY: The Guilford Press.

Rothbart, M. K., Ahadi, S. A., & Hershey, K. L. (1994). Temperament and social behavior in childhood. *Merrill-Palmer Quarterly*, 21-39.

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development, 72*(5), 1394-1408.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34,* 100-114.

Samejima, F. (2010). The general Graded Response Model. In M. L. Nering, & R. Ostini (Eds.), *Handbook of Polytomous Item Response Theory Models* (pp. 77-109). New York, NY: Taylor & Francis.

Schroder, H. S., Clark, D. A., & Moser, J. S. (2017). Screening for problematic worry in adults with a single item from the Penn State Worry Questionnaire. *Assessment*. doi: 10.1177/1079191117694453

Shiner, R. L., & Caspi, A. (2012). Temperament and the development of personality traits, adaptations, and narratives. In M. Zentner, & R. L. Shiner (Eds.), *Handbook of Temperament* (pp. 497-518). New York: The Guilford Press.

Shiner, R. L., & DeYoung, C. G. (2013). The structure of temperament and personality traits: A developmental perspective. In P. Zelazo (Ed.), *Oxford Handbook of Developmental Psychology* (pp. 113-141). New York: Oxford University Press.

Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of quality of life scales: An illustration with the Satisfaction with Life Scale (SWLS). *Social Indicators Research, 92*(3), 489-496.

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research, 102*(3), 443-461.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102-111. doi:10.1037//1040-3590.121.102

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292.

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?. *Journal of Applied Psychology*, *91*(1), 25.

Stautz, K., & Cooper, A. (2013). Impulsivity-related personality traits and adolescent alcohol use: A meta-analytic review. *Clinical Psychology Review, 33*, 574-592. doi: 10.1016/j.cpr.2013.03.003

Stucky, B. D., Thissen, D., & Edelen, M. O. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement, 37*(1), 41-57. doi: 10.1177/0146621612462759

Tackett, J. L., Martel, M. M., & Kushner, S. C. (2012). Temperament, externalizing disorders, and Attention-Deficit/Hyperactivity Disorder. In M. Zentner, & R. L. Shiner (Eds.), *Handbook of Temperament* (pp. 562-580). New York: The Guilford Press.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: development of the Multidimensional Personality Questionnaire. In *The SAGE handbook of personality theory and assessment 2* (pp. 261-292). Thousand Oaks, CA: SAGE Publishing Inc.

The Children's Behavior Questionnaire. (2017). Retrieved from https://research.bowdoin.edu/rothbart-temperament-questionnaires/instrument-descriptions/the-childrens-behavior-questionnaire/

Thissen, D., Nelson, L., & Swygert, K. A. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—approximation methods for scale scores. In D. Thissen & H. Wainer (Eds.), *Test Scoring.* Hillsdale, NJ: Erlbaum.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring.* Hillsdale, NJ: Erlbaum.

Thomas, A., Chess, S., Birch, H. G., Hertzig, M. E., & Korn, S. (1963). *Behavioral individuality in early childhood.* New York, NY: New York University Press.

U.S. Census Bureau. (2014). *Households and families: 2010-2014 American Community Survey 5-year estimates.* Retrieved December 18, 2015, from 25http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58-79. doi: 10.1037/1082-989X.12.1.58

Zentner, M., & Shiner, R. L. (2012). Fifty years of progress in temperament research: A synthesis of major themes, findings, and challenges and a look forward. In M. Zentner, & R. L. Shiner (Eds.), *Handbook of Temperament* (pp. 673-700). New York: The Guilford Press.

APPENDIX A

TABLES

Table 1 Conceptual Definitions of the CBQ scales

Scale	Description
Attentional Focusing	Capacity to maintain attentional focus on task-related channels
Attentional Shifting	Capacity to shift attention between tasks
Inhibitory Control	Capacity to plan and to suppress inappropriate approach responses under instructions or in novel or uncertain situations
Low Intensity Pleasure	Pleasure or enjoyment related to situations involving low stimulus intensity, rate, complexity, novelty, and incongruity
Perceptual sensitivity	Detection of slight, low-intensity stimuli from the external environment
Anger/Frustration	Negative affectivity related to interruption of ongoing tasks or goal blocking
Discomfort	Negative affectivity related to sensory qualities of stimulation, including intensity; rate; or complexities of light, movement, sound, and texture
Soothability	Rate of recovery from peak distress, excitement, or general arousal.
Fear	Negative affectivity, including unease, worry, or nervousness, which is related to anticipated pain or distress and/or potentially threatening situations
Sadness	Negative affectivity and lowered mood and energy related to exposure to suffering, disappointment, and object loss
Activity Level	Gross motor activity, including rate and extent of locomotion
High Intensity Pleasure	Pleasure or enjoyment related to situations involving high stimulus intensity, rate, complexity, novelty, and incongruity
Impulsivity	Speed of response initiation
Approach	Amount of excitement and anticipation for expected pleasurable activities
Shyness	Slow or inhibited speed of approach, and discomfort, in social situations
Smiling/Laughter	Positive affect in response to changes in stimulus intensity, rate, complexity, and incongruity

Note. Taken from Rothbart and colleagues (2001) and CBQ scoring manual.

Table 2 Graded Response Model and Dimensionality Results for Original Effortful Control Scales

Scale/Item	Discrimination	-2	-1	0	1	2	I-ECV
Attentional Focusing	α = .73	3.69	3.73	3.76	3.70	3.55	
16[1]	.84	.22	.22	.23	.22	.22	62%
38[1]	1.17	.41	.42	.43	.40	.39	85%
47[1]	1.39	.58	.58	.61	.60	.57	72%
125[2]	1.18	.43	.44	.43	.41	.35	35%
144[2]	.99	.31	.31	.31	.30	.28	12%
160[3]	.35	.04	.04	.04	.04	.04	41%
171[2]	1.15	.39	.41	.42	.41	.40	45%
186	.30	.03	.03	.03	.03	.03	-
195[3]	.95	.27	.27	.28	.28	.28	60%
Attentional Shifting	α = .66	2.91	2.96	2.9	2.94	2.85	
6[1]	1.17	.42	.43	.41	.40	.39	-
29[1]	.94	.25	.25	.26	.27	.28	-
95[1]	1.91	1.05	1.08	1.03	1.07	.99	-
180	.34	.04	.04	.04	.04	.04	-
184[1]	.72	.16	.16	.16	.16	.15	-
Inhibitory Control	α = .81	5.37	5.39	5.29	5.14	4.99	
4[1]	.98	.31	.31	.29	.28	.23	-
20[1]	.91	.27	.27	.26	.25	.23	-
32[1]	1.56	.73	.73	.76	.67	.66	-
63	.58	.11	.11	.11	.11	.10	-
75[1]	1.41	.62	.63	.59	.57	.55	-
93[1]	1.22	.43	.44	.46	.47	.45	-
108[1]	.95	.27	.27	.28	.29	.28	-
116[2]	.51	.08	.08	.08	.08	.08	14%
136[1]	1.61	.80	.81	.72	.70	.69	-
147[2]	.95	.29	.29	.28	.26	.24	50%
162[2]	.43	.06	.06	.06	.06	.06	24%
168[2]	.73	.16	.16	.17	.17	.17	49%
185	.88	.24	.24	.24	.23	.23	-

Table 2 Continued

Scale/Item	Discrimination	-2	-1	0	1	2	I-ECV
Low Intensity Pleasure	α = .80	5.86	5.77	5.45	4.2	2.97	
12	.58	.11	.11	.10	.09	.08	-
36[2]	.44	.06	.06	.06	.06	.06	12%
54[3]	.79	.19	.18	.17	.13	.09	92%
66[3]	.93	.27	.26	.20	.11	.06	36%
76[1]	1.5	.71	.68	.58	.22	.08	76%
86[2]	.39	.05	.05	.05	.05	.05	8%
111	.66	.14	.16	.14	.14	.13	-
113[1]	1.31	.55	.53	.51	.41	.23	46%
133[1]	1.66	.86	.82	.77	.43	.15	69%
146[3]	1.36	.59	.57	.56	.45	.26	91%
151[3]	1.01	.31	.31	.30	.26	.20	98%
164[1]	1.15	.41	.43	.42	.40	.33	21%
174[1]	1.38	.61	.62	.59	.45	.25	7%
Perceptual Sensitivity	α = .92	9.28	23.58	16.5	11.81	3.13	
9[2]	.47	.07	.07	.07	.06	.06	19%
28[1]	2.45	1.78	1.90	1.75	1.59	.67	62%
31[1]	2.86	1.63	2.54	2.38	2.11	.58	53%
52[2]	.38	.05	.05	.05	.04	.04	19%
65	7.87	3.87	17.15	1.39	6.16	.01	-
84[1]	.64	.12	.13	.13	.13	.13	00%
98[2]	.83	.22	.21	.21	.19	.17	72%
105[1]	.69	.15	.15	.15	.15	.15	00%
122[2]	.58	.11	.10	.10	.09	.08	75%
142[2]	.20	.01	.01	.01	.01	.01	18%
154[2]	.32	.03	.03	.03	.03	.03	28%
170[1]	.86	.24	.24	.23	.22	.19	87%

Note. Item and test information presented at five levels of the latent trait, -2, -1, 0, 1, and 2. Total test information presented in row with scale name. Superscripts denote factor structure supported by IFAs and used in bi-factor models; identical superscripts denote that the items loaded on the same factor (loadings above .4). α = marginal reliability of scale; I-ECV = Item level explained common variance.

Table 3 Graded Response Model Results for Revised Effortful Control Scales

Scale/Item	Discrimination	-2	-1	0	1	2
Attentional Focusing	α = .73/.70	3.52/3.14	3.64/3.30	3.74/3.34	3.65/3.27	3.60/3.19
16	1.04/1.43	.33/.59	.34/.63	.34/.65	.34/.63	.33/.59
38	1.59/1.17	.68/.38	.71/.40	.78/.42	.76/.43	.74/.41
47	1.88/1.68	.98/.78	1.05/.87	1.09/.87	1.03/.84	1.02/.81
171	.72/.46	.16/.07	.17/.07	.17/.07	.16/.07	.16/.07
195	1.09/1.02	.37/.32	.37/.33	.37/.32	.36/.31	.35/.31
Attentional Shifting	α = .66/.62	2.93/2.63	2.99/2.66	2.92/2.58	2.97/2.58	2.85/2.51
6	1.17/1.36	.42/.58	.44/.58	.42/.55	.41/.54	.39/.51
29	.93/.33	.24/.03	.25/.03	.26/.03	.26/.03	.27/.03
95	1.98/1.71	1.13/.87	1.16/.90	1.11/.85	1.16/.86	1.06/.82
184	.67/.69	.14/.15	.14/.15	.14/.15	.14/.15	.14/.14
Inhibitory Control	α = .79/.76	4.91/4.28	4.92/4.32	4.84/4.25	4.69/4.14	4.65/4.01
4	.98/.84	.31/.23	.31/.23	.29/.22	.28/.21	.23/.19
20	.85/.75	.23/.18	.23/.18	.23/.18	.22.17	.20/.17
32	1.64/1.41	.80/.59	.79/.61	.83/.62	.73/.58	.72/.54
75	1.33/1.43	.56/.63	.56/.64	.53/.60	.51/.58	.49/.56
93	1.3/1.27	.49/.49	.49/.49	.52/.51	.53/.51	.51/.49
108	.97/.77	.28/.18	.28/.18	.29/.19	.30/.19	.30/.19
136	1.65/1.56	.84/.75	.84/.77	.75/.72	.73/.67	.73/.66
147	.83/.64	.22/.13	.22/.13	.21/.13	.20/.12	.19/.12
168	.78/.58	.18/.10	.18/.10	.19/.11	.19/.11	.19/.11

Table 3 Continued

Scale/Item	Discrimination	-2	-1	0	1	2
Low Intensity Pleasure	α = .71/.72	4.11/4.12	4.02/3.99	3.73/3.75	2.86/3.14	1.93/2.37
54	.87/.87	.23/.23	.22/.22	.21/.22	.15/.20	.10/.15
66	1.20/1.70	.46/.92	.44/.86	.33/.73	.13/.28	.06/.06
76	1.23/1.19	.47/.43	.46/.42	.40/.40	.20/.30	.09/.14
133	1.25/1.09	.48/.37	.47/.36	.44/.34	.32/.33	.17/.26
146	1.33/1.26	.56/.50	.55/.49	.53/.47	.44/.45	.26/.38
151	1.73/1.46	.91/.67	.89/.64	.82/.60	.62/.58	.25/.39
Perceptual Sensitivity	α = .80/.76	2.82/3.32	4.87/4.14	5.25/4.31	5.54/4.43	4.91/4.22
28	2.52/1.93	.68/.83	1.68/1.06	1.84/1.13	2.01/1.18	1.87/1.10
31	2.46/2.18	.65/.88	1.63/1.37	1.81/1.44	1.91/1.50	1.4/1.35
98	.90/1.10	.19/.27	.23/.35	.24/.36	.25/.37	.25/.38
122	.66/.71	.10/.14	.12/.15	.13/.15	.13/.16	.14/.16
170	.87/.83	.19/.20	.22/.21	.23/.22	.24/.22	.24/.22

Note. Item and test information presented at five levels of the latent trait, -2, -1, 0, 1, and 2. Total test information presented in row with scale name. Results from maternal reports on left side of slash, results from paternal reports on right side of slash. α = marginal reliability of scale.

Table 4 Original and Revised Scale Intercorrelations for Effortful Control

	Original					Revised				
Maternal Reports	AF	AS	IC	LP	SE	AF	AS	IC	LP	SE
AF	-	.15	.13	.18	.03	-	.15	.13	.18	.03
AS	.01	-	.13	.03	.05	-.14	-	.13	.03	.05
IC	.50	.02	-	.09	.01	.59	-.11	-	.09	.01
LP	.42	.09	.40	-	.00	.26	.06	.32	-	.00
SE	.23	.03	.29	.30	-	.20	-.02	.30	.30	-
	Original					Revised				
Paternal Reports	AF	AS	IC	LP	SE	AF	AS	IC	LP	SE
AF	-	.13	.08	**.24**	.06	-	.13	.08	**.24**	.06
AS	-.08	-	.11	.05	.01	-.21	-	.11	.05	.01
IC	.46	-.07	-	.11	.03	.52	-.18	-	.11	.03
LP	.38	-.03	.36	-	.09	.16	.02	.26	-	.09
SE	.12	.07	.18	.34	-	.18	.06	.21	.42	-

Note. AF = Attentional Focusing; AS = Attentional Shifting; IC = Inhibitory Control; LP = Low Intensity Pleasure; Perceptual Sensitivity. Correlations between scales presented in bottom half of tables, Cohen's *q*s comparing original and revised scale intercorrelations presented in top half of table. Cohen's *q*s above .20 bolded.

Table 5 Graded Response Model and Dimensionality Results for Original Negative Affectivity Scales

Scale/Item	Discrimination	-2	-1	0	1	2	I-ECV
Anger/Frustration	α = .81	5.63	5.44	5.48	5.40	4.59	
2[3]	.89	.24	.25	.25	.25	.24	-
19	.87	.24	.24	.24	.23	.23	-
34[1]	1.53	.70	.74	.70	.70	.65	-
62[1]	2.15	1.43	1.16	1.23	1.25	.57	-
73	.71	.16	.16	.16	.16	.16	-
78[1]	1.29	.51	.53	.51	.51	.49	-
120[3]	.64	.13	.13	.13	.13	.13	-
128[1]	.85	.23	.23	.22	.22	.20	-
140[1]	.78	.19	.19	.19	.19	.18	-
156[2]	.51	.08	.08	.08	.08	.08	63%
173[2]	.81	.21	.21	.21	.20	.19	12%
181[1]	.97	.30	.30	.29	.29	.29	-
193[2]	.79	.20	.20	.19	.19	.19	62%
Discomfort	α = .86	6.05	7.12	7.48	7.24	5.23	
5[1]	1.19	.43	.45	.45	.42	.39	90%
21	.61	.12	.12	.12	.12	.11	-
61[1]	3.22	1.97	2.83	3.1	2.96	1.5	96%
87[2]	.32	.03	.03	.03	.03	.03	8%
97[2]	.47	.07	.07	.07	.07	.07	16%
101[1]	2.61	1.77	1.93	2.04	1.97	1.54	95%
115	.30	.03	.03	.03	.03	.03	-
132[3]	1.19	.44	.46	.46	.414	.36	-
141	.28	.02	.03	.03	.03	.03	-
157	.23	.02	.18	.02	.02	.02	-
178[2]	.44	.06	.06	.06	.06	.06	18%
190[3]	.52	.08	.08	.09	.09	.09	-

Table 5 Continued

Scale/Item	Discrimination	-2	-1	0	1	2	I-ECV
Soothability	α = .85	4.49	7.45	7.24	5.98	5.26	
14[1]	.34	.04	.04	.04	.04	.04	14%
27[1]	.24	.02	.02	.02	.02	.02	14%
42[3]	.43	.05	.05	.05	.05	.05	46%
53[1]	.25	.02	.02	.02	.02	.02	2%
68	.73	.17	.17	.17	.16	.16	-
85	.30	.03	.03	.03	.03	.03	-
92[3]	.53	.09	.09	.08	.08	.08	58%
103	.12	.00	.005	.00	.005	.00	-
118[3]	.62	.12	.12	.12	.11	.11	56%
134[2]	2.86	.87	2.58	2.36	1.85	1.52	34%
150[2]	2.91	1.32	2.54	2.58	1.87	1.59	9%
167	.73	.17	.18	.17	.17	.16	-
177[2]	1.39	.60	.62	.60	.58	.48	76%
Fear	α = .86	4.66	6.76	7.68	7.10	5.41	
15[2]	.28	.03	.03	.03	.03	.03	15%
40[1]	1.15	.27	.39	.43	.42	.41	99%
50[2]	.52	.08	.09	.09	.09	.09	44%
58	.55	.10	.10	.10	.09	.09	-
70[1]	2.12	1.07	1.34	1.42	1.32	1.10	25%
80[2]	.46	.07	.07	.07	.07	.07	37%
91[1]	1.26	.41	.50	.51	.50	.47	88%
130[1]	3.38	1.20	2.77	3.55	3.09	1.68	37%
138	.61	.11	.12	.12	.12	.12	-
161[2]	.33	.03	.03	.03	.03	.03	29%
176[1]	.98	.27	.30	.31	.31	.30	54%
189	.29	.03	.03	.03	.03	.03	-

Table 5 Continued

Scale/Item	Discrimination	-2	-1	0	1	2	I-ECV
Sadness	α = .70	3.27	3.37	3.37	3.33	3.28	
18[1]	1.01	.32	.33	.32	.31	.30	-
39[1]	.77	.13	.17	.18	.18	.19	-
44[1]	1.21	.44	.46	.46	.44	.43	-
55[1]	1.10	.35	.38	.39	.38	.38	-
64	.64	.13	.13	.13	.17	.12	-
72	.73	.17	.17	.17	.17	.17	-
81[1]	.93	.27	.27	.27	.27	.26	-
94[1]	.96	.29	.29	.28	.28	.27	-
109[2]	.44	.06	.06	.06	.06	.06	9%
112[2]	.30	.03	.03	.03	.03	.03	4%
127	.42	.05	.06	.06	.06	.06	-
149	.26	.02	.02	.02	.02	.02	-

Note. Item and test information presented at five levels of the latent trait, -2, -1, 0, 1, and 2. Total test information presented in row with scale name. Superscripts denote factor structure supported by IFAs and used in bi-factor models; identical superscripts denote that the items loaded on the same factor (loadings above .4). α = marginal reliability of scale; I-ECV = Item level explained common variance.

Table 6 Graded Response Model Results for Revised Negative Affectivity Scales

Scale/Item	Discrimination	-2	-1	0	1	2
Anger/Frustration	α = .79/.79	5.19/4.92	4.91/4.97	5.01/4.81	4.95/4.70	3.79/4.07
2	.88/1.18	.24/.42	.25/.44	.25/.43	.24/.42	.24/.41
19	.77/.55	.19/.10	.19/.10	.19/.10	.18/.09	.18/.09
34	1.69/1.75	.88/.89	.91/.97	.85/.90	.84/.89	.77/.80
62	2.46/2.09	1.85/1.34	1.44/1.26	1.63/1.21	1.60/1.14	.56/.65
78	1.22/1.22	.47/.45	.47/.47	.46/.47	.45/.45	.44/.44
128	.84/.96	.22/.30	.22/.29	.21/.28	.21/.27	.19/.26
181	.96/1.00	.29/.32	.30/.32	.29/.31	.29/.31	.28/.30
193	.66/.60	.14/.11	.13/.11	.13/.11	.13/.11	.13/.11
Discomfort	α = .86/80	5.39/4.80	7.38/5.15	8.35/5.39	7.88/5.13	4.33/4.64
21	.57/.70	.11/.16	.11/.16	.10/.16	.10/.16	.10/.15
61	4.15/2.59	2.3/1.71	4.16/1.93	5.07/2.05	4.66/1.92	1.44/1.61
97	.46/.58	.07/.11	.07/.11	.07/.11	.07/.11	.07/.11
101	2.31/2.42	1.45/1.55	1.55/1.67	1.62/1.79	1.58/1.67	1.32/1.51
132	1.17/.92	.42/.26	.44/.27	.44/.27	.42/.26	.35/.25
178	.39/.21	.05/.01	.05/.01	.05/.01	.05/.01	.05/.01
Soothability	α = .85/.86	2.55/4.75	7.86/7.77	8.17/7.96	6.22/6.89	6.93/6.32
53	.16/.15	.01/.01	.01/.01	.01/.01	.01/.01	.01/.01
68	.63/.46	.13/.07	.13/.07	.13/.07	.13/.07	.12/.07
92	.43/.27	.06/.02	.06/.02	.06/.02	.06/.02	.06/.02
118	.53/.34	.09/.03	.09/.03	.09/.03	.08/.03	.08/.03
134	3.16/3.26	.28/.98	3.06/3.02	3.11/2.96	2.20/2.38	2.53/2.07
150	3.24/3.42	.49/2.19	2.99/3.15	3.27/3.41	2.25/2.94	2.66/2.71
177	1.29/1.21	.50/.44	.52/.47	.52/.46	.50/.43	.47/.41

Table 6 Continued

Scale/Item	Discrimination	-2	-1	0	1	2
Fear	α = .87/.87	4.12/4.54	7.53/7.32	9.46/9.11	8.05/8.23	4.93/6.38
40	1.07/1.02	.25/.28	.34/.33	.37/.34	.36/.34	.36/.33
50	.44/.57	.06/.10	.06/.11	.06/.11	.06/.11	.06/.11
70	2.22/2.13	1.13/1.13	.05/1.35	1.55/1.43	.05/1.35	1.18/1.21
80	.38/.25	.05/.02	1.46/.02	.05/.02	1.43/.02	.05/.02
91	1.20/.99	.38/.29	.45/.31	.47/.31	.46/.30	.43/.30
130	4.29/4.36	1.00/1.56	3.88/4.05	5.67/5.73	4.39/4.94	1.57/3.25
176	.96/.73	.26/.16	.28/.16	.30/.17	.30/.17	.29/.17
Sadness	α = .69/68	3.18/3.13	3.28/3.17	3.28/3.14	3.25/3.10	3.20/3.05
18	.99/.88	.31/.24	.32/.25	.31/.25	.30/.24	.29/.23
39	.78/.55	.14/.09	.17/.09	.19/.09	.19/.10	.20/.10
44	1.20/1.28	.44/.50	.46/.52	.45/.51	.44/.49	.43/.48
55	1.12/1.08	.36/.36	.39/.37	.40/.37	.39/.36	.39/.35
64	.62/.59	.12/.11	.12/.11	.12/.11	.12/.11	.12/.11
72	.76/.72	.18/.16	.18/.17	.18/.17	.18/.17	.18/.16
81	.88/1.05	.24/.35	.24/.35	.24/.34	.24/.33	.23/.32
94	.97/.94	.30/.28	.30/.28	.29/.27	.29/.26	.28/.26
112	.18/.21	.01/.01	.01/.01	.01/.01	.01/.01	.01/.01
127	.50/.29	.07/.03	.08/.03	.08/.03	.08/.03	.08/.03

Note. Note. Item and test information presented at five levels of the latent trait, -2, -1, 0, 1, and 2. Total test information presented in row with scale name. Results from maternal reports on left side of slash, results from paternal reports on right side of slash. α = marginal reliability of scale.

Table 7 Original and Revised Scale Intercorrelations for Negative Affectivity

	Original					Revised				
Maternal Reports	AN	DS	SO	FR	SD	AN	DS	SO	FR	SD
AN	-	.04	.06	.00	.08	-	.04	.06	.00	.08
DS	.30	-	.12	.04	.05	.26	-	.12	.04	.05
SO	.15	.46	-	.01	.08	.21	.36	-	.01	.08
FR	.48	.43	.41	-	.05	.48	.40	.42	-	.05
SD	.39	.31	.26	.38	-	.32	.26	.18	.34	-
	Original					Revised				
Paternal Reports	AN	DS	SO	FR	SD	AN	DS	SO	FR	SD
AN	-	.07	.07	.04	.01	-	.07	.07	.04	.01
DS	.32	-	.11	.00	.02	.26	-	.11	.00	.02
SO	.15	.43	-	.09	.02	.22	.34	-	.09	.02
FR	.44	.38	.31	-	.03	.47	.38	.39	-	.03
SD	.32	.29	.32	.37	-	.31	.31	.30	.34	-

Note. AN = Anger/Frustration; DS = Discomfort; SO = Soothability; FR = Fear; SD = Sadness. Correlations between scales presented in bottom half of tables, Cohen's *q*s comparing original and revised scale intercorrelations presented in top half of table.

Table 8 Graded Response Model and Dimensionality Results for Original Surgency Scales

Scale/Item	Discrimination	-2	-1	0	1	2	I-ECV
Activity Level	α = .83	3.65	5.44	6.15	6.27	5.96	
1[1]	.58	.11	.11	.11	.11	.11	19%
25[1]	.97	.27	.29	.30	.30	.30	65%
41[2]	.93	.17	.24	.26	.27	.28	41%
48	.06	.001	.001	.001	.001	.001	-
88[2]	.61	.11	.12	.12	.12	.12	23%
102[2]	1.22	.37	.43	.45	.46	.46	58%
123[3]	1.91	.23	.82	1.10	1.15	.98	43%
126[2]	.03	.0004	.0004	.0004	.0004	.0004	8%
145[3]	1.66	.47	.81	.88	.87	.73	84%
153[3]	1.34	.14	.35	.51	.55	.57	81%
172[1]	1.73	.29	.76	.89	.93	.93	98%
187	.56	.10	.10	.10	.10	.10	-
192[2]	1.16	.40	.42	.42	.41	.39	84%
High Intensity Pleasure	α = .82	6.24	6.53	6.34	5.19	3.82	
8[2]	2.27	1.49	1.59	1.54	1.10	.46	-
22[1]	.91	.25	.28	.27	.26	.26	-
30[1]	1.09	.38	.38	.37	.36	.34	-
51	1.26	.49	.51	.50	.48	.45	-
60[2]	2.20	1.36	1.54	1.46	.98	.41	-
67[1]	1.02	.32	.29	.25	.11	.07	-
77	.69	.15	.15	.15	.15	.15	-
100[1]	.59	.11	.11	.11	.11	.11	-
107	.53	.09	.09	.09	.09	.09	-
124[1]	.70	.15	.15	.14	.14	.13	-
139[1]	1.01	.33	.32	.32	.29	.25	-
159	.24	.02	.02	.02	.02	.02	-
182	.57	.01	.10	.10	.10	.10	-

Table 8 Continued

Scale/Item	Discrimination	-2	-1	0	1	2	I-ECV
Impulsivity	α = .81	5.01	5.35	5.41	5.19	5.01	
13[1]	1.01	.30	.33	.33	.32	.32	99%
26	.37	.04	.04	.04	.04	.04	-
46	1.12	.40	.38	.38	.36	.35	-
59[2]	2.29	1.28	1.58	1.63	1.50	1.38	50%
71[2]	1.41	.60	.61	.62	.59	.57	13%
79[1]	.84	.21	.22	.22	.22	.22	99%
90	.89	.25	.25	.25	.25	.24	-
104[3]	.61	.12	.12	.12	.12	.11	28%
114	.55	.10	.10	.10	.09	.09	-
137	.53	.09	.09	.09	.09	.09	-
155[3]	.40	.05	.05	.05	.05	.05	13%
169[2]	.97	.30	.30	.29	.27	.26	1%
183[2]	.97	.29	.29	.30	.29	.29	6%
Positive Anticipation	α = .79	5.28	5.17	5.11	4.54	3.75	
10[1]	.76	.18	.19	.18	.18	.18	28%
24[1]	.86	.22	.22	.22	.21	.19	33%
35[1]	.61	.12	.12	.11	.11	.11	16%
69	.48	.08	.07	.07	.07	.07	-
82[2]	.98	.29	.28	.27	.19	.14	60%
96[2]	1.16	.43	.41	.41	.38	.34	69%
117[2]	1.88	1.13	.97	.99	.88	.52	83%
131[3]	.83	.21	.22	.22	.21	.21	-
148[4]	1.48	.68	.68	.65	.59	.49	45%
166[4]	1.51	.66	.74	.71	.43	.24	42%
175	.41	.05	.05	.05	.05	.05	-
188	.38	.05	.05	.05	.05	.05	-
191	.75	.18	.18	.18	.18	.17	-

Table 8 Continued

Scale/Item	Discrimination	-2	-1	0	1	2	I-ECV
Shyness	α = .92	12.60	13.55	13.37	12.55	8.87	
7[1]	1.22	.47	.48	.47	.44	.32	58%
17[1]	2.01	1.21	1.25	1.21	1.13	.64	53%
23[1]	2.14	1.33	1.43	1.40	1.24	.95	50%
37	1.25	.48	.49	.50	.49	.43	-
45[1]	2.12	1.40	1.39	1.34	1.06	.26	58%
57[1]	2.61	1.84	2.11	1.95	1.88	.77	58%
74[1]	1.38	.56	.56	.60	.57	.38	77%
89[2]	1.55	.71	.73	.76	.73	.65	95%
106[2]	1.77	.73	.95	.98	.99	.90	97%
119	1.23	.48	.49	.47	.45	.37	-
129[2]	1.99	1.12	1.23	1.23	1.14	.85	99%
143[2]	1.56	.62	.73	.75	.77	.73	99%
158[2]	1.49	.65	.72	.70	.67	.62	61%
Smiling and Laughter	α = .86	7.83	9.15	8.58	6.86	3.16	
11[1]	.60	.10	.10	.10	.09	.07	24%
43[1]	.47	.07	.07	.07	.07	.06	8%
56	-	-	-	-	-	-	-
83	.21	.01	.01	.01	.01	.01	-
99[1]	.83	.21	.20	.18	.13	.08	8%
110[2]	2.04	1.25	1.30	1.18	1.12	.52	-
121	.91	.25	.24	.23	.20	.13	-
135[2]	2.56	1.81	2.08	1.82	.89	.09	11%
152[1]	.66	.13	.13	.12	.10	.08	24%
163[2]	1.70	.90	.84	.80	.51	.14	67%
165[2]	2.57	.96	2.06	2.02	1.73	.37	13%
179	.81	.21	.21	.21	.20	.20	-
194[2]	1.70	.91	.89	.85	.79	.41	14%

Note. Item and test information presented at five levels of the latent trait, -2, -1, 0, 1, and 2. Total test information presented in row with scale name. Superscripts denote factor structure supported by IFAs and used in bi-factor models; identical superscripts denote that the items loaded on the same factor (loadings above .4). α = marginal reliability of scale; I-ECV = Item level explained common variance.

Table 9 Graded Response Model Results for Revised Surgency Scales

Scale/Item	Discrimination	-2	-1	0	1	2
Activity Level	α = .82/82	2.86/3.45	5.04/5.64	6.09/6.01	6.25/6.20	5.65/5.87
25	.81/.79	.19/.18	.20/.19	.21/.20	.21/.20	.21/.20
102	.89/.65	.26/.12	.29/.13	.30/.13	.30/.13	.30/.13
123	2.41/2.28	.17/.41	1.14/1.38	1.72/1.57	1.82/1.59	1.38/1.35
145	1.77/1.73	.48/.67	.92/.90	1.01/.95	.99/.94	.81/.83
153	1.52/1.63	.14/.30	.42/.68	.66/.74	.71/.84	.73/.86
172	1.70/1.91	.29/.45	.74/1.02	.86/1.08	.89/1.16	.90/1.16
192	1.04/1.05	.32/.33	.34/.34	.34/.35	.33/.35	.32/.34
High Intensity Pleasure	α = .74/75	4.15/4.40	4.19/4.41	4.08/4.22	3.60/3.85	3.27/3.01
22	1.05/.94	.31/.28	.35/.28	.35/.28	.34/.28	.34/.27
30	1.43/1.48	.64/.69	.66/.70	.64/.67	.61/.63	.56/.49
51	1.28/1.41	.51/.62	.53/.63	.52/.61	.49/.59	.46/.52
60	1.34/1.78	.56/1.01	.57/1.01	.56/.95	.46/.78	.32/.27
67	1.28/1.05	.51/.34	.45/.32	.40/.25	.13/.14	.07/.06
100	.69/.62	.16/.12	.15/.12	.15/.12	.15/.12	.14/.11
124	.73/.70	.16/.15	.16/.15	.16/.15	.15/.14	.14/.13
139	.99/.77	.31/.19	.31/.19	.30/.19	.28/.18	.24/.16
Impulsivity	α = .80/.75	4.61/3.80	4.85/3.92	5.08/4.03	5.01/4.08	4.87/3.97
13	.82/.89	.21/.24	.21/.25	.21/.25	.21/.26	.21/.25
46	1.04/1.18	.29/.37	.31/.38	.32/.41	.33/.43	.34/.44
59	2.41/1.84	1.55/.93	1.65/.98	1.78/1.02	1.72/1.04	1.60/.96
71	1.72/1.57	.79/.68	.85/.72	.90/.75	.89/.75	.87/.72
79	.69/.57	.15/.10	.15/.10	.15/.10	.15/.10	.15/.10
169	1.02/.92	.26/.24	.30/.25	.32/.26	.33/.27	.33/.27
183	1.10/.86	.37/.23	.38/.23	.38/.23	.38/.23	.37/.23

Table 9 Continued

Scale/Item	Discrimination	-2	-1	0	1	2
Positive Anticipation	α = .78/.78	5.36/4.79	5.48/5.06	5.05/4.80	4.22/4.27	2.32/3.12
10	.52/.54	.09/.09	.09/.09	.09/.09	.09/.09	.08/.09
24	.63/.64	.12/.13	.12/.13	.12/.12	.11/.12	.11/.11
96	.83/.74	.21/.17	.21/.17	.21/.17	.20/.16	.17/.16
117	1.59/1.39	.76/.60	.72/.56	.70/.53	.66/.51	.31/.46
131	.74/.84	.17/.21	.17/.22	.17/.22	.17/.22	.17/.21
148	2.31/2.27	1.64/1.42	1.64/1.60	1.36/1.44	1.37/1.24	.38/.85
166	2.19/2.01	1.38/1.16	1.54/1.30	1.41/1.23	.63/.93	.10/.24
Shyness	α = .89/.86	9.57/7.51	10.31/7.79	9.93/7.63	9.14/7.11	5.25/5.49
17	2.32/2.02	1.58/1.21	1.63/1.27	1.58/1.25	1.47/1.09	.68/.88
23	2.77/2.15	2.12/1.36	2.34/1.40	2.26/1.40	1.91/1.30	1.12/1.08
45	2.02/2.03	1.28/1.27	1.26/1.25	1.22/1.19	1.00/1.13	.28/.51
57	2.84/2.39	2.11/1.64	2.50/1.79	2.25/1.68	2.18/1.50	.79/1.03
74	1.30/1.28	.50/.47	.50/.49	.53/.52	.51/.50	.36/.42
106	1.38/1.05	.52/.33	.59/.35	.61/.35	.61/.35	.58/.35
158	1.23/.87	.47/.24	.49/.24	.48/.24	.47/.23	.44/.23
Smiling and Laughter	α = .86/.86	7.13/7.13	8.83/8.57	8.30/7.93	6.62/6.93	2.64/4.19
11	.50/.30	.07/.03	.07/.03	.07/.03	.06/.03	.05/.28
99	.62/.52	.12/.08	.11/.08	.10/.08	.08/.07	.06/.06
110	2.12/1.77	1.34/.95	1.40/1.00	1.27/.93	1.20/.81	.52/.75
135	2.68/3.18	1.93/2.65	2.27/3.18	1.98/2.72	.98/2.09	.09/.16
163	1.46/1.31	.67/.54	.63/.52	.60/.48	.43/.46	.16/.31
165	2.79/2.33	1.04/.79	2.40/1.65	2.37/1.65	2.02/1.54	.34/1.09
194	1.76/1.90	.97/1.10	.95/1.12	.91/1.04	.85/.94	.42/.80

Note. Item and test information presented at five levels of the latent trait, -2, -1, 0, 1, and 2. Total test information presented in row with scale name. Results from maternal reports on left side of slash, results from paternal reports on right side of slash. α = marginal reliability of scale.

Table 10 Original and Revised Scale Intercorrelations for Surgency Scales

	Original						Revised					
Maternal Reports	AL	HP	IM	PA	SH	SL	AL	HP	IM	PA	SH	SL
AL	-	.01	**.29**	.19	**.48**	**.30**	-	.01	**.29**	.19	**.48**	**.30**
HP	.35	-	.08	.00	.09	.12	.34	-	.08	.00	.09	.12
IM	.58	.50	-	.17	.12	.18	.36	.44	-	.17	.12	.18
PA	.50	.29	.43	-	.05	.17	.63	.29	.28	-	.05	.17
SH	.58	.36	.61	.22	-	.20	.18	.28	.68	.17	-	.20
SL	.51	.32	.35	.52	.33	-	.70	.21	.18	.63	.14	-
	Original						Revised					
Paternal Reports	AL	HP	IM	PA	SH	SL	AL	HP	IM	PA	SH	SL
AL	-	.08	**.30**	.19	.13	**.32**	-	.08	**.30**	.19	.13	**.32**
HP	.36	-	.08	.02	.04	.12	.29	-	.08	.02	.04	.12
IM	.60	.55	-	.18	.09	**.21**	.37	.49	-	.18	.09	**.21**
PA	.54	.28	.44	-	.00	.15	.66	.26	.28	-	.00	.15
SH	.35	.32	.54	.13	-	**.44**	.23	.28	.60	.13	-	**.44**
SL	.53	.32	.38	.54	.53	-	.72	.21	.19	.64	.15	-

Note. AL = Activity Level; HP = High Intensity Pleasure; IM = Impulsivity; PA = Positive Anticipation; SH = Shyness; SL = Smiling and Laughter. Correlations between scales presented in bottom half of tables, Cohen's *q*s comparing original and revised scale intercorrelations presented in top half of table. Cohen's *q*s above .20 bolded.

Table 11 Descriptive Statistics for Original and Revised Scales

	Maternal Reports					Paternal Reports				
	Original		Revised			Original		Revised		
	M	SD	M	SD	*r*	M	SD	M	SD	*r*
Attentional Focusing	4.17	.43	4.59	1.08	.85	4.04	.27	4.49	.81	.82
Attentional Shifting	3.85	.17	4.00	.21	.87	3.90	.14	4.09	.19	.85
Inhibitory Control	4.06	.32	4.21	.47	.94	3.98	.22	4.16	.31	.95
Low Intensity Pleasure	4.99	.18	5.39	.19	.78	4.77	.13	5.17	.23	.75
Perceptual Sensitivity	4.73	.37	5.49	.88	.87	4.60	.28	5.27	.77	.85
Effortful Control	4.03	.04	4.31	.06	.92	3.92	.02	4.21	.04	.89
Anger/Frustration	4.35	.73	4.08	.84	.93	4.30	.71	4.06	.82	.92
Discomfort	3.72	.79	3.62	1.06	.90	3.65	.71	3.58	.96	.89
Soothability	4.37	.73	4.51	.95	.90	4.25	.71	4.34	.89	.90
Fear	3.86	.90	3.36	1.11	.89	3.83	.83	3.36	1.03	.90
Sadness	3.58	.71	3.32	.78	.97	3.60	.63	3.37	.70	.96
Negative Affectivity	3.83	.54	3.57	.64	.96	3.83	.49	3.61	.60	.96
Activity Level	4.92	.80	5.17	1.10	.89	4.93	.73	5.08	1.00	.88
High Intensity Pleasure	4.56	.75	5.02	.82	.93	4.53	.71	5.02	.79	.93
Impulsivity	4.60	.72	4.68	.90	.90	4.56	.65	4.70	.82	.89
Positive Anticipation	4.74	.66	4.56	.82	.88	4.61	.63	4.39	.81	.88
Shyness	3.60	1.14	3.82	1.21	.96	3.68	.99	3.90	1.08	.95
Smiling and Laughter	5.57	.70	5.46	.99	.92	5.34	.66	5.20	.93	.90
Surgency	4.78	.57	4.78	.64	.97	4.67	.52	4.68	.60	.97

Note. M = Mean; SD = Standard Deviation; *r* = correlation between original and revised scale.

Table 12 Correlations with the Child Behavior Checklist and Interparent Agreement for Original and Revised Scales

	Maternal Reports					
	Externalizing			Internalizing		
	Original	Revised	q	Original	Revised	q
Attentional Focusing	-.24	-.35	.12	-.01	-.10	.09
Attentional Shifting	.07	.14	.07	.01	.05	.04
Inhibitory Control	-.46	-.43	.04	-.09	-.08	.01
Low Intensity Pleasure	-.24	-.35	.12	-.01	-.10	.09
Perceptual Sensitivity	-.12	-.20	.08	.02	-.04	.06
Effortful Control	-.30	-.36	.07	-.03	-.09	.06
Anger/Frustration	.35	.40	.06	.17	.19	.02
Discomfort	-.03	-.03	.00	.15	.13	.02
Soothability	-.20	-.10	.10	-.22	-.17	.05
Fear	.03	.07	.04	.31	.30	.01
Sadness	.08	.10	.02	.27	.31	.04
Negative Affectivity	.17	.17	.00	.33	.32	.01
Activity Level	.39	.33	.07	.02	.05	.03
High Intensity Pleasure	.13	.17	.04	-.13	-.12	.01
Impulsivity	.29	.20	.10	-.13	-.22	.09
Approach	.24	.20	.04	.12	.09	.03
Shyness	-.14	-.07	.07	.23	.26	.03
Smiling and Laughter	.07	.14	.07	-.05	.05	.10
Surgency	.24	.22	.02	-.14	-.15	.01

Table 12 Continued

	Paternal Reports								
	Externalizing			Internalizing			Parent Agreement		
	Original	Revised	*q*	Original	Revised	*q*	Original	Revised	*q*
Attentional Focusing	-.20	-.29	.10	-.02	-.12	.10	.53	.48	.07
Attentional Shifting	.01	.05	.04	-.06	-.06	.00	.26	.30	.04
Inhibitory Control	-.44	-.41	.04	-.09	-.08	.01	.58	.55	.04
Low Intensity Pleasure	-.20	-.29	.10	-.02	-.12	.10	.36	.31	.06
Perceptual Sensitivity	-.17	-.23	.06	-.09	-.17	.08	.42	.38	.05
Effortful Control	-.33	-.39	.07	-.10	-.20	.10	.53	.49	.05
Anger/Frustration	.28	.33	.06	.15	.16	.01	.40	.50	.13
Discomfort	-.02	-.06	.04	.19	.13	.06	.48	.50	.03
Soothability	-.11	-.03	.08	-.19	-.12	.07	.52	.54	.03
Fear	.03	.08	.05	.34	.31	.03	.55	.52	.04
Sadness	.07	.08	.01	.24	.25	.01	.37	.41	.05
Negative Affectivity	.13	.13	.00	.32	.28	.04	.55	.58	.04
Activity Level	.33	.28	.06	-.02	.02	.04	.62	.73	**.20**
High Intensity Pleasure	.11	.13	.02	-.16	-.14	.02	.62	.52	.15
Impulsivity	.27	.19	.08	-.16	-.24	.08	.61	.61	.00
Approach	.16	.19	.03	.06	.09	.03	.45	.55	.13
Shyness	-.17	-.12	.05	.20	.22	.02	.67	.65	.03
Smiling and Laughter	-.02	.08	.10	-.13	-.03	.10	.54	.68	**.22**
Surgency	.19	.21	.02	-.18	-.17	.01	.65	.65	.00

Note. Externalizing = total externalizing problems scale of CBCL; Internalizing = total internalizing problems scale of the CBCL. Parent Agreement = correlation between maternal and paternal reports; *q* = Cohen's *q*s comparing original and revised scale correlations. Cohen's *q*s above .20 bolded.

Table 13 Exploratory Factor Analytic Results for Original and Revised CBQ Scales Based on Maternal Reports

	Three Factor Solution		
Original	Factor 1	Factor 2	Factor 3
Attentional Focusing	.02	**.59**	-.23
Attentional Shifting	.08	.10	.04
Inhibitory Control	-.03	**.60**	**-.49**
Low Intensity Pleasure	.13	**.76**	-.02
Perceptual Sensitivity	.32	**.44**	.01
Anger/Frustration	**.49**	-.24	**.45**
Discomfort	**.61**	-.01	-.09
Soothability	**.59**	.02	-.21
Fear	**.74**	.002	.05
Sadness	**.42**	**-.53**	-.01
Activity Level	-.02	-.01	**.78**
High Intensity Pleasure	-.21	.03	**.53**
Impulsivity	-.20	-.14	**.80**
Approach	.34	.30	**.61**
Shyness	-.34	.07	**.51**
Smiling and Laughter	.004	**.61**	**.56**
Revised	Factor 1	Factor 2	Factor 3
Attentional Focusing	**-.59**	-.19	.02
Attentional Shifting	.15	-.02	.17
Inhibitory Control	**-.62**	-.33	-.004
Low Intensity Pleasure	-.30	-.07	.26
Perceptual Sensitivity	-.19	-.11	.24
Anger/Frustration	**.70**	.03	.21
Discomfort	**.40**	-.20	-.10
Soothability	.35	-.26	.04
Fear	**.56**	-.23	.09
Sadness	**.62**	-.06	**-.44**
Activity Level	-.001	.21	**.76**
High Intensity Pleasure	.02	**.43**	.19
Impulsivity	-.001	**.91**	.05
Approach	.04	.11	**.72**
Shyness	-.22	**.72**	.002
Smiling and Laughter	-.28	-.02	**.93**

Table 13 Continued

	Four Factor Solution			
Original	Factor 1	Factor 2	Factor 3	Factor 4
Attentional Focusing	**.59**	.01	-.08	-.12
Attentional Shifting	.10	.07	-.01	.07
Inhibitory Control	**.61**	-.03	-.12	-.36
Low Intensity Pleasure	**.75**	.12	-.03	.10
Perceptual Sensitivity	**.44**	.39	.15	-.05
Anger/Frustration	-.26	**.44**	-.05	.53
Discomfort	-.02	**.65**	.04	-.07
Soothability	.01	**.56**	-.11	-.06
Fear	-.01	**.72**	-.07	.17
Sadness	**-.53**	**.43**	-.01	-.03
Activity Level	-.03	-.08	.03	**.83**
High Intensity Pleasure	.03	-.11	.37	.21
Impulsivity	-.18	.01	**.89**	.14
Approach	.28	.35	.18	**.54**
Shyness	.07	-.15	**.67**	-.05
Smiling and Laughter	**.63**	-.05	.01	**.65**
Revised	Factor 1	Factor 2	Factor 3	Factor 4
Attentional Focusing	**.62**	-.12	.02	-.07
Attentional Shifting	-.07	.17	.14	-.02
Inhibitory Control	**.80**	-.02	-.04	-.15
Low Intensity Pleasure	**.51**	.22	.18	.07
Perceptual Sensitivity	**.46**	.31	.15	.05
Anger/Frustration	**-.44**	**.49**	.15	.003
Discomfort	.06	**.55**	-.22	-.03
Soothability	.01	**.49**	-.06	-.14
Fear	-.04	**.74**	-.05	-.09
Sadness	-.28	.34	**-.50**	.01
Activity Level	-.25	-.02	**.82**	.03
High Intensity Pleasure	-.05	.03	.19	.38
Impulsivity	-.04	.01	.01	**.96**
Approach	-.02	.27	**.70**	.05
Shyness	.19	-.04	-.02	**.74**
Smiling and Laughter	.05	-.004	**.94**	-.14

Note. Geomin oblique rotation used; pattern coefficients presented. Factor loadings above .40 bolded.

Table 14 Exploratory Factor Analytic Results for Original and Revised CBQ Scales for Paternal Reports

	Three Factor Solution		
Original	Factor 1	Factor 2	Factor 3
Attentional Focusing	-.002	**.54**	-.22
Attentional Shifting	.10	.004	.15
Inhibitory Control	-.02	**.54**	**-.53**
Low Intensity Pleasure	.19	**.79**	-.01
Perceptual Sensitivity	.28	**.43**	.10
Anger/Frustration	**.45**	-.09	**.52**
Discomfort	**.66**	.11	-.06
Soothability	**.56**	.001	-.18
Fear	**.66**	.002	.12
Sadness	**.41**	**-.49**	.002
Activity Level	-.05	.15	**.77**
High Intensity Pleasure	-.20	.04	**.55**
Impulsivity	-.17	-.04	**.82**
Approach	.36	**.47**	**.55**
Shyness	**-.41**	.02	**.49**
Smiling and Laughter	-.003	**.72**	**.45**
Revised	Factor 1	Factor 2	Factor 3
Attentional Focusing	**-.59**	-.19	.02
Attentional Shifting	.15	-.02	.17
Inhibitory Control	**-.62**	-.33	-.004
Low Intensity Pleasure	-.11	.35	-.09
Perceptual Sensitivity	-.10	.22	-.02
Anger/Frustration	**.69**	.23	.14
Discomfort	**.45**	-.01	-.20
Soothability	**.47**	.01	-.26
Fear	**.66**	.12	-.18
Sadness	**.57**	**-.40**	-.02
Activity Level	.01	**.79**	.15
High Intensity Pleasure	.06	.20	**.46**
Impulsivity	-.01	.10	**.85**
Approach	.14	**.78**	.02
Shyness	-.25	.01	**.67**
Smiling and Laughter	-.17	**.90**	-.10

Table 14 Continued

	Four Factor Solution			
Original	Factor 1	Factor 2	Factor 3	Factor 4
Attentional Focusing	**.53**	-.08	-.22	-.004
Attentional Shifting	.01	.14	.10	.02
Inhibitory Control	**.57**	-.08	-.20	-.36
Low Intensity Pleasure	**.79**	.18	**-.50**	-.002
Perceptual Sensitivity	**.45**	.33	.11	-.06
Anger/Frustration	-.14	**.48**	.04	**.43**
Discomfort	.10	**.65**	-.19	-.01
Soothability	-.04	**.49**	-.31	.02
Fear	-.02	**.66**	-.12	.11
Sadness	**-.48**	**.44**	-.004	-.09
Activity Level	.03	-.14	-.07	**1.02**
High Intensity Pleasure	.07	-.04	**.59**	.04
Impulsivity	-.04	.02	**.79**	.18
Approach	**.41**	.37	.06	**.46**
Shyness	.05	-.27	**.59**	.02
Smiling and Laughter	**.67**	-.003	.10	**.38**
Revised	Factor 1	Factor 2	Factor 3	Factor 4
Attentional Focusing	**.62**	-.12	.02	-.07
Attentional Shifting	-.07	.17	.14	-.02
Inhibitory Control	**.80**	-.02	-.04	-.15
Low Intensity Pleasure	**.59**	.02	.18	.18
Perceptual Sensitivity	**.55**	.003	.05	.25
Anger/Frustration	-.04	**.69**	.11	.15
Discomfort	.34	**.53**	-.18	-.02
Soothability	.06	**.47**	-.06	-.21
Fear	.09	**.67**	.01	-.12
Sadness	-.05	**.53**	**-.46**	-.04
Activity Level	-.31	.02	**.92**	.002
High Intensity Pleasure	.08	.09	.10	**.51**
Impulsivity	-.04	.002	-.02	**.91**
Approach	.004	.19	**.75**	.03
Shyness	.005	-.23	-.05	**.69**
Smiling and Laughter	.04	-.11	**.89**	-.07

Note. Geomin oblique rotation used; pattern coefficients presented. Factor loadings above .40 bolded.

Table 15 Eigenvalues from Exploratory Factor Analyses

	Original Scales				
	E1	E2	E3	E4	E5
Maternal Reports					
Actual Data	3.45	3.10	2.35	1.02	.91
Parallel analysis	1.30	1.24	1.19	1.15	1.11
Paternal Reports					
Actual Data	3.65	3.00	2.15	1.05	.86
Parallel analysis	1.34	1.27	1.21	1.63	1.12

Table 15 Continued

	Revised Scales				
	E1	E2	E3	E4	E5
Maternal Reports					
Actual Data	3.31	2.97	2.08	1.29	1.01
Parallel analysis	1.30	1.24	1.19	1.15	1.11
Paternal Reports					
Actual Data	3.46	2.95	1.90	1.33	.93
Parallel analysis	1.34	1.27	1.21	1.16	1.12

Note. E1…E5 = eigenvalues 1 through 5. Average values from parallel analyses (based on 1000 replications) presented.

Table 16 Factor Correlations and Congruence Coefficients for Exploratory Factor Analyses

	Three Factor Solutions			Four Factor Solutions			
	F1	F2	F3	F1	F2	F3	F4
Original Scales	(.98)			(.95)			
F1	-	-.32	.07	-	-.24	.12	.11
F2	-.18	-	.04	-.15	-	.09	.15
F3	.06	.04	-	.07	-.20	-	.60
F4	-	-	-	-.08	-.01	.56	-
Revised Scales	(.98)			(.87)			
F1	-	-.07	.12	-	-.09	.19	-.24
F2	.07	-	.26	-.27	-	.02	.11
F3	.17	.20	-	.03	.05	-	.36
F4	-	-	-	-.19	-.07	.35	-

Note. F1…F4 = Factors 1 through 4. Factor correlations from analyses with maternal reports reported below diagonal, factor correlations from analyses with paternal reports reported above diagonal. Congruence coefficients comparing maternal and paternal factor solutions presented in parentheses.

Table 17 Number of Items in Original and Revised Scales

	Original	Revised	Reduction
Attentional Focusing	9	5	44%
Attentional Shifting	5	4	20%
Inhibitory Control	13	9	31%
Low Intensity Pleasure	13	6	54%
Perceptual Sensitivity	12	5	58%
Effortful Control	52	29	44%
Anger/Frustration	13	8	38%
Discomfort	12	6	50%
Soothability	13	7	46%
Fear	12	7	42%
Sadness	12	10	17%
Negative Affectivity	62	38	39%
Activity Level	13	7	46%
High Intensity Pleasure	13	8	38%
Impulsivity	13	7	46%
Approach	13	7	46%
Shyness	13	7	46%
Smiling and Laughter	13	7	46%
Surgency	78	43	45%
Total	195	110	44%

Note. Although the full CBQ contains 195 items, only 192 items were actually included in the analyses here. Items 3, 33, and 49 are not incorporated into any scale in the CBQ scoresheet. As such, these three items by default are not included in the revised CBQ.