HHS Public Access

Author manuscript

Psychol Assess. Author manuscript; available in PMC 2017 December 01.

Published in final edited form as:

Psychol Assess. 2016 December ; 28(12): 1646–1662. doi:10.1037/pas0000299.

Measurement Invariance and Child Temperament: An Evaluation of Sex and Informant Differences on the Child Behavior Questionnaire

D. Angus Clark,
Michigan State University

Caitlin J. Listro,
Michigan State University

Sharon L. Lo,
Michigan State University

C. Emily Durbin,
Michigan State University

M. Brent Donnellan, and
Texas A & M University

Tricia K. Neppl
Iowa State University

Abstract

Parent reports of temperament are used to study many important topics in child development, such as whether boys and girls differ in their levels of emotional reactivity and self-regulation. However, questions regarding measurement equivalence in parental reports of temperament are largely unexplored, despite the fact that this issue is critical for drawing the correct conclusions from mean-level comparisons. In the current study, measurement invariance across boys and girls (as targets), and mothers and fathers (as informants), was investigated in the Child Behavior Questionnaire (CBQ; Rothbart et al., 2001) using a sample of children ranging in age from 3 to 7 years (N = 605). Several instances of non-invariance were identified across both girls and boys, and mothers and fathers. An evaluation of effect size indices suggests that the practical impact of this non-invariance ranges from negligible to moderate. All told, this study illustrates the importance of taking a psychometrically informed approach to the use of parent reports of child temperament.

Keywords

temperament; temperament assessment; parent report; Child Behavior Questionnaire; measurement invariance; sex differences

Correspondence regarding this manuscript should be addressed to D. Angus Clark, Department of Psychology, Michigan State University, 244C Psychology Building, East Lansing, MI 48824. clarkd46@msu.edu.

Introduction

Children differ from one another on a variety of 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 represent dimensions of temperament (Rothbart, Ahadi, & Hershey, 1994; Shiner & DeYoung, 2013; Shiner et al., 2012). Individual differences in childhood temperament form the basis of adult personality (Shiner & DeYoung, 2013), and predict short- and long-term developmental outcomes such as psychopathology (e.g., Caspi, Moffitt, Newman, & Silva, 1996; Klein, Dyson, Kujawa, & Kotov, 2012; Tackett, Martel, & Kushner, 2012), self-esteem (e.g., Robins, Donnellan, Widaman, & Conger, 2010), and substance use (e.g., Clark, Donnellan, Robins, & Conger, 2015; Creemers et al., 2010; Stautz & Cooper, 2013). Research on temperament thus contributes to basic knowledge about human development, and informs programs and policies designed to facilitate positive life outcomes (Moffitt et al., 2011). Inferences about temperament, however, hinge upon the measurement properties of the assessments. As the most common approach to measuring child temperament is parent report (Goldsmith & Gagne, 2012), a deeper understanding of the properties of this method is important for both substantive and methodological reasons.

Knowledge regarding measurement invariance - or the degree to which a particular measure functions similarly across groups (e.g., girls and boys, mothers and fathers) - is critical for establishing that scores on a given assessment are analogous and comparable across different groups (Borsboom, 2006; Zumbo, 2007). A lack of measurement invariance makes it difficult to provide a clear interpretation of any observed group differences. One important example of group comparisons in the temperament literature where this issue is relevant concerns sex differences (Else-Quest, Hyde, Goldsmith, & Van Hulle, 2006). It is necessary to test whether measures are invariant across boys and girls so that valid interpretations can be drawn from any observed sex differences. Yet measurement invariance in temperament assessment is largely unexplored. In the current study we therefore tested the equivalence of ratings in the widely used Child Behavior Questionnaire (CBQ; Rothbart, Ahadi, Hershey, & Fisher, 2001) across girls and boys (as targets), and mothers and fathers (as informants). Our specific goals were to a) test for invariance across child sex, and across mother and father raters; b) quantify sex differences once measurement invariance was established; and c) consider mean-level agreement between mothers and fathers regarding child traits.

Temperament

There are a number of different approaches for classifying the multitude of temperamental differences that can be observed in children (Goldsmith et al., 1987; Mervielde, 2012). These approaches are broadly similar to those used to classify traits in the adult personality literature (Rothbart, 2007). Specifically, individual differences are organized hierarchically with narrow dimensions at the bottom (e.g., cheerfulness, feelings of vulnerability) and broad, higher order dimensions (e.g., Extraversion, Neuroticism) at the top. The higher-order dimensions are usually conceptualized as latent variables that capture consistent patterns of covariation among the lower level primary scales or facets. One of the more prominent models of early childhood temperament (Rothbart, 2011) posits that 16 lower order facets

coalesce into three higher order dimensions: Effortful Control, Negative Affectivity, and Surgency (though Rothbart and colleagues do conceptualize the structure of temperament differently across the lifespan).

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 early-life equivalent to the traits of Conscientiousness and Constraint from the Big 5 and Big 3 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 is largely analogous to the Big 5 and Big 3 traits of Extraversion and Positive Emotionality (Clark & Watson, 2008; Shiner & DeYoung, 2013; Tellegen & Waller, 2008).

The Child Behavior Questionnaire

In the adult personality literature, traits are most commonly assessed via self-report. Given the limited feasibility of self-report with young children, research on temperament has instead largely relied on parent report measures (Goldsmith & Gagne, 2012; Lo, Vroman, & Durbin, 2014). That is, rather than ask children to report on their own temperament, parents are asked to report on their child's temperament. The Child Behavior Questionnaire (CBQ; Rothbart et al., 2001), designed for use with children between 3 and 7 years old, is currently the most widely used parent report measure of temperament (Kotelnikova, Olino, Klein, Kryski, & Hayden, 2015). Kotelnikova and colleagues (2015) reported the primary article describing the development of the CBQ (Rothbart et al., 2001) has been cited over 900 times in the past 10 years. This popularity makes the CBQ one of the primary measures of childhood temperament in the literature.

The CBQ contains 195 items, each tapping one of 16 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 Rothbart et al., 2001 for conceptual definitions). Items tend to emphasize observable, relatively specific behaviors (e.g., “gets mad when only mildly criticized”) rather than more global judgments of the child's characteristics. Respondents rate the degree to which these specific behaviors characterized the target child over the past six months. Responses are given on a 7-point Likert-type scale that ranges from “1: extremely untrue of your child” to “7: extremely true of your child”. Item content was derived both rationally and from extensive parental interviews.

The 16 facet scales of the CBQ are subsumed by the superordinate 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). Typically, the

Effortful Control dimension includes the facets of attentional focusing, inhibitory control, low intensity pleasure, and perceptual sensitivity (Rothbart et al., 2001). Negative Affectivity includes the facets of anger/frustration, discomfort, sadness, fear, and soothability (Rothbart et al., 2001). Lastly, Surgency includes the facets of smiling and laughter, high intensity pleasure, impulsivity, shyness, and positive anticipation (Rothbart et al., 2001). Lower-order scale scores are computed by averaging together a scale's individual items. Higher-order dimensions are computed by averaging or summing the scale scores that are included under a given dimensions.

Measurement Invariance and Temperament

Concerns about the validity of parent reports of temperament have been raised in the literature for at least twenty years (Gartstein, Bridgett, & Low, 2012; Richters, 1992; Seifer, 2002). Accordingly, it is important to develop a thorough understanding of how parental reports function in order to interpret existing and future research on temperament. One significant issue that has been historically unexplored is the issue of measurement invariance. That is, how analogous are parental ratings of temperament across distinct groups of informants and targets?

Measurement invariance refers to the extent to which a particular measure functions similarly across groups (de Ayla, 2009; Embretson & Reise, 2000; Millsap & Olivera-Aguilar, 2012; Osterlind & Everson, 2009). This issue is relevant whenever scores on the same measure are integrated, compared, or interpreted across distinct groups or time points (Borsboom, 2006; Millsap & Olivera-Aguilar, 2012; Tay, Meade, & Cao, 2015; Zumbo, 2007). For example, if a girl and a boy have the same observed score, does this actually mean that they have the same standing on the underlying latent trait being assessed?

Measurement invariance is not necessarily an either/or status; a measure can be invariant or not at several different levels (Millsap & Olivera-Aguilar, 2012). Here, the focus is on three of the most fundamental levels of invariance: weak, strong, and strict[1]. These levels of invariance can be defined clearly from a confirmatory factor analytic perspective (Millsap & Olivera-Aguilar, 2012). Weak invariance (or factorial invariance) refers to whether factor loadings are equivalent across groups, and tests whether factor indicators change at the same rate across groups with fluctuations in the latent factor (Brown, 2006; Millsap & Olivera-Aguilar, 2012). The second level of invariance, strong invariance (or scalar invariance), refers to whether the factor indicator intercepts are equal across groups and tests whether scores on the factor indicators are the same across groups when the latent factor is at its zero-point (Brown, 2006; Millsap & Olivera-Aguilar, 2012). The third level of invariance, strict invariance, refers to whether residual variances are equal across groups and tests if factor indicators are measured with the same precision across groups (Brown, 2006; Millsap & Olivera-Aguilar, 2012). Importantly, these levels of invariance may not apply consistently to all indicators of the latent construct (i.e., scales or items). Indeed, it is possible that a measure exhibit only partial invariance such that only a subset of indicators are invariant at a given level of analysis (Millsap & Olivera-Aguilar, 2012).

[1]Configural invariance – the extent to which the basic factor structure is equivalent across groups – was tested and supported before conducting the analyses on display here. These results are available upon request.

Tests of measurement invariance ensure that scores can be legitimately compared across different groups (Borsboom, 2006). Generally, weak and strong invariance must be achieved to justify mean-level comparisons (Brown, 2006; Millsap & Olivera-Aguilar, 2012). If factor loadings and intercepts are not equal, mean level comparisons are invalid. Thus, for example, mean level differences across boys and girls would not be interpretable if the scales being compared did not exhibit weak and strong invariance. When a measure is partially invariant it may still be possible to conduct multiple types of group comparisons, but care should be taken in how scores are calculated (e.g., scores corrected for non-invariance could be generated via a structural equation modeling program; Tay et al., 2015).

The evaluation of measurement invariance is essential for research in temperament. First, measurement invariance is critical for interpreting sex differences. Knowledge of sex differences in childhood temperament are important for understanding differential outcomes for boys and girls (e.g., discrepant rates of direct aggression; Card, Sawalani, Stucky, & Little, 2008), and the childhood origins of sex differences in adult personality (Hyde, 2014; Olino, Durbin, Klein, Hayden, & Dyson, 2012). Else-Quest and colleagues (2006) conducted a meta-analysis on temperamental sex differences and found that girls had higher levels of Effortful Control ($d = 1.01$), and boys had higher levels of Surgency ($d = .55$). The observed sex differences in their meta-analysis could unknowingly have been attenuated or exaggerated by measurement invariance as the issue of measurement invariance can be difficult to address meta-analytically.

Second, measurement invariance is critical for evaluating differences between mothers and fathers as informants of child temperament. Mothers' and fathers' reports of their child's behavior often demonstrate only moderate agreement (Achenbach, McConaughy, & Howell, 1987; Duhig, Renk, Epstein, & Phares, 2000; Rothbart et al., 2001). This could be a function of measurement non-invariance. To the extent that mothers' and fathers' reports are influenced by different biases (De Los Reyes & Kazdin, 2005), or informed by differential access to trait-relevant behaviors (Richters, 1992), they may interpret and use parent report questionnaires differently. Examining measurement invariance across parental informants can reveal specific instances where mothers and fathers use questionnaires differently, which can help to achieve a greater understanding of how parents come to understand and describe child temperament.

Current Study

We are not aware of any study that has comprehensively evaluated the CBQ with respect to measurement invariance across child sex and parent raters. One study demonstrated invariance across sex and ethnicity for CBQ teacher ratings, but only examined the factor-structure of Effortful Control (Sulik et al., 2010). Other studies on invariance in the CBQ have also only focused on the basic factor structure across groups (e.g. Rothbart et al., 2001). We advance these studies by testing whether factor loadings, intercepts, and residual variances are invariant across mother and father reports, as well as the sex of the target child, in a large sample of children. Specially, we evaluated measurement invariance across two groups of targets and raters: boys and girls, and mothers and fathers. Invariance was

additionally tested with child and parent sex crossed (e.g., invariance of mothers' reports only across boys and girls, invariance of mothers' and fathers' reports of girls only).

Method

Participants

The data used in the current study come from three separate samples of children and their parents. Characteristics of each individual sample are provided below, followed by a brief description of the pooled sample used in analyses. By combining samples it was possible to examine the functioning of the CBQ across a more diverse range of children and parents in a larger sample, which has the added advantage of increasing the precision of parameter estimates.

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 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).

Sample 2—Participants were recruited from the greater Lansing 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

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

(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 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).

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).

Total Sample—Prior to pooling the data, differences between samples were examined. 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%). 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 (e.g., mother reported inhibitory control was significantly lower in sample 3 than the other two samples; full results available upon request). Moreover, preliminary exploratory factor analyses (available upon request) supported the same general 3 factor higher-order structure for all three samples. Factor loadings were similar across samples. Congruence coefficients for mothers' and fathers' composite CBQ ratings across samples were .90 or above for Effortful Control (.90-.97),

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

Negative Affectivity (.97-.98), and Surgency (.91-.99). Results were similar examining maternal and paternal ratings separately.

When combined for the main analyses, the total sample included 605 children with CBQ data (47.3% girls) aged 3 to 7 years. The mean 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.

Data Analytic Strategy

All analyses discussed here, with the exception of the Cohen's *d* for Mean and Covariance Structure analysis (dMACS; described below), were conducted with Mplus version 7.2 (Muthen & Muthen, 1998-2012). Parameters were estimated via Full Information Maximum Likelihood, which provides generally unbiased estimates even in the face of missing data (Allison, 2009; Graham & Coffman, 2012). The dMACS analyses were run using the program developed by Nye and Drasgow (2011); this program has also been used to investigate the invariance of personality inventories across cultures and age groups (Nye, Allemand, Gosling, Potter, & Roberts, in press; Nye & Drasgow, 2011).

Testing Invariance—Tests of measurement invariance followed the general procedure outlined by Millsap and Olivera-Aguilar (2012). First, a completely unconstrained measurement model was estimated across groups. In these models, factor loadings, intercepts, and residual variances were allowed to freely vary across groups (e.g., girls and boys); the individual CBQ scales served as indicators. Factor variances were set to 1, and factor means were set to 0. Following the estimation of the unconstrained “baseline model”, factor loadings were constrained to equality across groups (weak invariance). The variance of one of the group's factors was also allowed to freely vary. The model with equal factor loadings was then compared to the fully unconstrained model.

The chi square difference test was used to compare these (and subsequent) nested models. The chi square difference is the most popular approach to testing nested models, however, this statistic can be overly sensitive with increasing sample sizes (e.g., Meade, Johnson, & Braddy, 2008). Therefore, we also evaluate differences in the CFI and RMSEA fit indices. There is evidence that in the context of testing measurement invariance differences in these indices of model fit can be useful for detecting a lack of invariance (Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). Cutoff values of -.002 (Meade, Johnson, & Braddy, 2008) for change in CFI, and .015 for change in RMSEA (Chen, 2007), have been suggested for indicating a noteworthy drop in fit. If there was any disagreement between these three metrics, we followed the implications of the majority (i.e., 2 out of the 3).

If there was evidence of a significant drop in fit when the constrained model was compared to the baseline model, major sources of fit degradation were sought. Residuals, modification indices, and the results of preliminary exploratory factor analyses were all inspected. The

scale with the largest residuals, modification indices, and/or parameter discrepancies across groups was identified as the most overtly non-invariant scale (the majority of these metrics tend to converge on a single scale). This scale was then subsequently allowed to vary across groups (while the rest of the scales remained constrained). This partially constrained model was then compared to the baseline model. These tests represent tests of partial invariance (Millsap & Olivera-Aguilar, 2012). If the partially invariant model still fit worse than the baseline mode, the next most overtly non-invariant scale was identified via the same procedure as above, and then this scale was subsequently freed. If it was not possible to constrain at least 2 scales to equality in a given stage without observing a significant degrade in model fit, invariance testing was halted. Furthermore, if a scale was found to be non-invariant at one stage, that scale's parameters were freely estimated during subsequent stages.

If there was evidence of weak - or partial weak - invariance, all the intercepts of scales with invariant factor loadings were constrained across groups (tests of strong invariance). Additionally, one of the group's factor means was allowed to freely vary (the same group with a freely varying variance). This model with equal (or partially equal) factor loadings and intercepts was compared to the equal (or partially equal) loadings model. If the new constraints produced a significant worsening of fit, tests of partial invariance were conducted using procedures analogous to the above.

Finally, following the constraint of factor loadings and intercepts, the residual variances of scales with invariant factor loadings and intercepts were constrained to equality (strict invariance). This model with equal residual variances, intercepts, and loadings was compared to the model with invariant (or partially invariant) loadings and intercepts. Specific sources of non-invariance were not sought out at this stage because, unlike weak and strong invariance, strict invariance is frequently viewed as unimportant in practical applications (Millsap & Olivera-Aguilar, 2012).

Evaluating the Impact of Non-Invariance—Although the sequential testing of weak, strong, and strict invariance can indicate if there is non-invariance at some level, this process does not reveal the extent to which non-invariance actually impacts reports of temperament. Indeed, the presence of non-invariance will lead to some artifactual group differences, but the effect could be negligible in magnitude. Chi square difference tests, and changes in CFI and RMSEA, are not informative with respect to the magnitude and practical implications of non-invariance. Because of this, we used several effect sizes to gauge the practical significance of any observed non-invariance. Importantly, several of these metrics (all of those based on the dMACS program; described below) are based on the parameter estimates from the fully unconstrained model, which means that the decisions made in the model comparison and selection process do not influence results. In all, two effect size metrics provide information at the individual scale level, and three provide information at the level of the aggregate temperamental traits.

The first measure of effect size indicates - assuming factor loadings are invariant - the proportion of the observed mean difference in scale scores that is attributable to intercept non-invariance (Millsap & Olivera-Aguilar, 2012). This is calculated when there is intercept

non-invariance by dividing the difference in model estimated intercepts by the difference in observed scale means. When a scale is invariant this value will be 0, which means that all of the observed difference is a function of true differences in levels of the latent trait. Values above 0 indicate the degree to which the observed mean difference reflects differences in intercepts as opposed to true latent trait differences. It is possible for this value to be over 1. When that happens, it indicates that the model estimated intercept difference was larger than the observed mean difference, or that that the impact of non-invariance was greater than the observed (or actual) difference (Nye & Drasgow, 2011).

The second effect size metric is essentially a Cohen's *d* for all non-invariance at the level of the individual scales. Referred to as dMACS, this metric indicates the area between groups' regression lines for a given scale (Nye & Drasgow, 2011). dMACS values were calculated using parameter estimates from a fully unconstrained model in which one invariant scale (identified during the initial tests of invariance) served as a marker variable for both groups (i.e., loadings set to 1, intercepts set to 0). These values were input into a computer program (available from Nye & Drasgow, 2011) which calculates the dMACS value for each scale. These values are interpreted in the same manner as Cohen's *d* (e.g., dMACS of .3 indicates a small to moderate effect of non-invariance). The dMACS values are especially informative as they capture the overall effect of all varieties of non-invariance (i.e., factor loading, intercept, and unique variance).

At the level of the temperamental dimensions, it was possible to calculate the proportion of observed mean differences that reflected non-invariance. This was done by dividing the difference in dimension scores due to non-invariance by the observed mean difference in dimension scores. The resulting value indicates the percentage of the observed difference in dimensional scores that is due to non-invariance (Nye & Drasgow, 2011). The difference in dimension scores due to non-invariance (the numerator) was derived via the dMACS program described above. This value indicates the difference in observed means that would be expected due to non-invariance alone, and is itself an effect size of non-invariance (Nye & Drasgow, 2011). For example, a value of .20 suggests that mother rated Effortful Control is expected to be .20 points higher than father rated Effortful Control because of non-invariance alone.

Finally, it was possible to calculate a Cohen's *d* for the overall effect of non-invariance at the level of the higher order dimensions of temperament. This was calculated by subtracting the model-estimated latent trait mean difference when not taking non-invariance into account from the model-estimated latent trait mean difference when non-invariance was taken into account (Tay et al., 2015). The model that does not take non-invariance into account assumes measurement equivalence, thus all factor loadings, intercepts, and unique variances are constrained to equality across groups. The model that takes non-invariance into account is the most justifiably constrained model as determined by the model testing procedure described above. For both of these models, one group's factor mean is set to 0, and their factor variance is set to 1. The other group's latent factor mean and variance are allowed to vary. These specifications ensure that the difference in model mean differences can be interpreted in approximately the same manner as Cohen's *d*. This metric provides a gauge of the extent to which non-invariance influences conclusions regarding trait-level differences

between groups (e.g., to what degree is the mean comparison of Effortful Control across girls and boys tainted by non-invariance). It can thus be thought of as capturing the overall impact of measurement non-invariance, and is referred to below as the overall effect of non-invariance.

Results

Preliminary Analyses

Means and standard deviations for the CBQ scales/dimensions for boys and girls, and for mothers and fathers, are presented in Table 1. Also presented are the correlations between mothers' and fathers' observed reports, and Cohen's *d*s for the observed mean differences between boys and girls, and mothers and fathers. The means and standard deviations listed in Table 1 for girls and boys represent the average of mothers' and fathers' ratings of the same target child. These composite scale scores were used as indicators when initially examining invariance across girls and boys.

There were several mean differences between girls' and boys' observed scores. Cohen's *d*s ranged from .02 (Soothability) to .65 (Fear). At the scale level, most differences centered around attention, self-control, and impulsivity. With regard to the higher order dimensions, girls were rated higher on effortful control and negative affectivity, whereas boys were rated as higher on surgency. Without taking measurement invariance into account, girls and boys differ on many facets and dimensions of temperament in this sample. At the scale level many of the observed sex differences were larger than those reported in Else-Quest and colleagues' (2006) meta-analysis (e.g., *d* of .29 versus .50 for the low intensity pleasure scale). However, at the dimensional level, 2 of the 3 current observed sex differences (Effortful Control and Surgency) were smaller than what was reported by Else-Quest and colleagues (*d*s of 1.01 versus .38 for Effortful Control, .55 versus .47 for Surgency, and .06 versus .22 for Negative Affectivity). On the whole though our effect size estimates tend to fall within the 95% confidence intervals reported in the meta-analysis.

Mean differences comparing mothers' and fathers' observed scores were less pronounced. Cohen's *d*s ranged from 0 (Sadness) to .27 (Smiling and Laughter; Low Intensity Pleasure). Across scales and dimensions, mothers tended to provide higher scores than fathers. At the scale level, most differences centered on attention, self-control, and positive affectivity. At the level of the higher order dimensions, mothers provided higher ratings of Effortful Control and Surgency, but there were no differences in ratings of Negative Affectivity. The correlation between parents' scores ranged from .35 (Approach; Low Intensity Pleasure) to .69 (Shyness). Most correlations were between .4 and .6, indicating moderate agreement between parents, which is typical of other studies of temperament traits and related constructs (e.g., Achenbach et al., 1987). The degree of mother-father correspondence is also consistent with previous studies of interparental agreement for CBQ scales (Rothbart et al., 2001).

Measurement Invariance Across Boys and Girls

Effortful Control—The baseline fit for the Effortful Control model for girls and boys was: χ^2 = 37.70, p < .001; RMSEA = .17; SRMR = .05; CFI = .93; TLI = .78 *(N* = 286 girls, 318 boys). Constraining factor loadings to equality across boys and girls (and allowing boys' factor variance to freely vary) did not result in a significant degrade in model fit ($\Delta\chi^2$ = 2.90, *df* = *3, p* = .41; *ΔCFI* = .001; *ΔRMSEA* = -.041), providing evidence for weak invariance. Constraining intercepts to equality across boys and girls (and allowing boys' factor mean to freely vary) to test strong invariance resulted in a significant degrade in fit ($\Delta\chi^2$ = 9.58, *df* = *3, p* = .02; *ΔCFI* = -.015; *ΔRMSEA* = -.011). Low intensity pleasure was identified as the most likely non-invariant intercept, and was subsequently unconstrained. The partially invariant model in which all intercepts except for low intensity pleasure were constrained to equality did not fit significantly worse than the model with equal loadings only ($\Delta\chi^2$ = 0.47, *df* = 2, *p* = .79; *ΔCFI* = .003; *ΔRMSEA* = -.017), providing evidence for partial strong invariance. It was not possible to constrain residual variances to equality across girls and boys; a model with all residual variances constrained (except for low intensity pleasure) to test strict invariance fit significantly worse than the model with most intercepts equal ($\Delta\chi^2$ = 19.44, *df* = *4, p* <001; *ΔCFI* = -.028; *ARMSEA* = .003). Parameter estimates for the most justifiably constrained Effortful Control model can be found in Table 2.

The impact of non-invariance was largest for the low intensity pleasure scale (dMACS = .3; see Table 8). Non-invariance accounted for 47% of the observed mean difference between boys' and girls' low intensity pleasure scores. Non-invariance was small to non-existent in the other scales. Girls are expected to score .12 points higher than boys on Effortful Control simply because of non-invariance. This suggests that non-invariance accounted for 14% of the observed mean difference between girls' and boys' Effortful Control scores (M_{diff} =.86; Cohen's *d* = .38). Following from all this, the overall effect of non-invariance was small, at . 09.

Examining mother and father ratings separately: Measurement invariance across boys and girls was also tested separately for mother (*N* = 277 girls, 310 boys) and father (*N* = 223 girls, 256 boys) reported Effortful Control. What follows is a brief discussion of the findings from these analyses. Complete results are available from the first author upon request. For maternal reports, it was possible to constrain everything across girls and boys except for the low intensity pleasure scale intercept and the residual variances, supporting partial strong invariance. Again, the largest effect of non-invariance on a single scale was found for the low intensity pleasure scale (dMACS = .3). The overall effect of non-invariance was relatively small (.11). Mothers' reports of girls' Effortful Control are expected to be .18 points higher than mothers' reports of boys' Effortful Control because of non-invariance. In the present sample, non-invariance accounted for 19% of the observed mean difference (M_{diff} =.93, Cohen's *d* = .43) between mother reported Effortful Control between girls and boys. For father reported Effortful Control it was possible to constrain all factor loadings and intercepts to equality. It was not possible to constrain residual variances. Thus father reported Effortful Control demonstrated strong invariance across boys and girls.

Negative Affectivity—The baseline fit of the Negative Affectivity model for boys and girls was: $\chi^2 = 91.98$, $p < .001$; RMSEA = .17; SRMR = .06; CFI = .90; TLI = .80 (N = 286 girls, 318 boys). Constraining factor loadings to equality across boys and girls (and allowing boys' factor variance to freely vary) did not lead to a significant decrement in fit ($\Delta\chi^2 = 8.77$, $df = 4$, $p = .07$; $\Delta CFI = -.005$; $\Delta RMSEA = -.022$), supporting the presence of weak invariance. Constraining intercepts to equality (and allowing the boys' factor mean to freely vary) to test strong invariance significantly reduced fit ($\Delta\chi^2 = 53.86$, $df = 4$, $p < .001$; $\Delta CFI = -.061$; $\Delta RMSEA = .016$). Partial strong invariance was not achievable; it was not possible to constrain any intercepts to equality without incurring a significant drop in model fit. Parameter estimates for the most justifiably constrained Negative Affectivity model can be found in Table 3.

The dMACS values indicated that the effect of non-invariance on the Negative Affectivity scales was substantial. The dMACS analysis was run twice for this dimension. The dMACS program requires that one invariant scale (invariant loadings and intercepts) serve as a marker variable. The dMACS value of the marker scale will be 0. Because all scales evidenced non-invariance at the intercept level, allowing a value of 0 to stand is inappropriate. Initially sadness was chosen as the marker variable because it evidenced the highest loading on the Negative Affectivity dimension. After this dMACS analysis was run, another analysis was conducted with fear serving as the marker variable as it had the lowest dMACS value in the prior analysis. This allowed us to obtain an effect size for all scales. With the exception of fear, all scales had dMACS values above .3[4] (see Table 8). Because no intercepts evidenced invariance, it was not possible to calculate the extent to which observed differences in scale means reflect non- invariance.[5] The overall effect of non-invariance was .32, meaning that non-invariance has a moderate impact when investigating sex differences in Negative Affectivity (to derive this effect size one intercept needed to be constrained, fear was chosen because of its relatively low dMACS value). Boys are expected to score .20 points higher than girls because of non-invariance alone. Here, this meant that 35% of the observed mean difference (M_{diff} =.55, Cohen's d = .21) was due to non-invariance.

Examining mother and father report separately: Non-invariance across girls and boys was examined separately for mothers' (N = 277 girls, 310 boys) and fathers' (N = 223 girls, 256 boys) reports. For mothers' ratings, it was possible to constrain all loadings to equality, as well as the fear and sadness intercepts and residual variances, supporting partial strict invariance. The anger/frustration, discomfort, and soothability intercepts and residual variances were non-invariant. The impact of non-invariance on these scales was moderate in size (dMACS between .2 and .3). The overall effect of non-invariance on comparisons of mother reported girl and boy Negative Affectivity was small (.08). Boys are expected to score .11 points higher than girls due to non-invariance alone; non-invariance accounted for 13% of the observed mean difference in the current sample (M_{diff} =.78, Cohen's d = .28).

[4]dMACS values were similar regardless of which scale served as the marker variable

[5]to calculate these values, there needs to be at least one invariant intercept so that latent means are on the same metric across groups

For fathers' reports, all scales except for sadness and anger/frustration were invariant, supporting partial strict invariance. The sadness and anger/frustration scales had non-invariant loadings, intercepts, and residual variances. The effect of non-invariance was substantial for the anger/frustration scale (dMACS = .51), and moderate for the sadness scale (dMACS = .34). The overall effect of non-invariance was small to moderate (.21). Boys are expected to score .56 points higher than girls because of non-invariance alone. Non-invariance accounted for 169% of the observed mean difference. This implies that the entirety of the observed mean difference (M_{diff} =33, Cohen's d = .03) reflects measurement non-invariance.

Surgency—The baseline fit of the Surgency model for girls and boys was: χ^2 = 138.55, p < .001; RMSEA = .21; SRMR = .06; CFI = .85; TLI = .70; N = 286 girls, 318 boys. The apparent lack of fit here partly stems from the fact that - as observed in preliminary exploratory factor analyses (available upon request) - two scales typically assigned to the Surgency factor (smiling and laughter, positive anticipation) do not hang particularly strongly with the Surgency factor. Smiling and laughter loads highly on the Effortful Control factor, and positive anticipation loads moderately across all three dimensions. As such there was an un-modeled residual correlation between these two scales, the omission of which negatively impacted model fit. These findings are consistent with other reports of the structure of the CBQ (e.g., Rothbart et al., 2001).

Constraining factor loadings to equality across boys and girls (and freeing boys' variance) did not lead to a significant degrade in model fit ($\Delta\chi^2$ = 4.23, df = 4, p = .38; ΔCFI = -.001; $\Delta RMSEA$ = -.031), supporting the presence of weak invariance. Constraint of the intercepts across boys and girls (and allowing boys' factor mean to freely vary) resulted in a significant degradation in fit ($\Delta\chi^2$ = 11.03, df = 4, p = .03; ΔCFI = -.008; $\Delta RMSEA$ = -.017), suggesting a lack of strong invariance. High intensity pleasure was identified as the scale most responsible for this decrement in fit. A partially constrained model in which all intercepts except for high intensity pleasure were set to equality did not fit worse than the model that only contained equal factor loadings ($\Delta\chi^2$ = 2.47, df = 3, p = .48; ΔCFI = .001; $\Delta RMSEA$ = -.017). A final model was tested in which all residual variances except for high intensity pleasure were set to equality. This model did not fit significantly worse than the partially strong invariant model ($\Delta\chi^2$ = 5.18, df = 4, p = .27; ΔCFI = -.001; $\Delta RMSEA$ = -.015), supporting partial strict invariance. The parameters for this model can be found in Table 4.

For the most part, the Surgency dimension is invariant across boys and girls. However, the high intensity pleasure scale had significant non-invariance at the intercept and residual level. Accordingly, its dMACS value is the highest (.2; see Table 8), although the effect is still relatively small. The non-invariance of the high intensity pleasure scale accounted for 43% of the observed difference in boys' and girls' scale scores. The overall effect of non-invariance on the comparison of girls' and boys' Surgency was also small (.04). Boys are expected to score .11 points higher than girls because of non-invariance. Non-invariance accounted for 10% of the observed mean difference in Surgency scores (M_{diff} =1.11, Cohen's d = .46).

Examining mother and father report separately: Invariance across girls and boys was examined separately for mothers' (N = 277 girls, 310 boys) and fathers' (N = 223 girls, 256 boys) reports of Surgency. Mothers' ratings of Surgency demonstrated partial strict invariance; everything was invariant except for the high intensity pleasure intercept and residual variance. The effect for non-invariance in the high intensity pleasure was small (dMACS = .19). Notably, the dMACS value for the positive anticipation scale was roughly equivalent to the value for the high intensity pleasure scale (dMACS = .18), suggesting that non-invariance is just as pronounced for the positive anticipation scale as it is for the high intensity pleasure scale. The overall effect of non-invariance was .04, indicating a small effect. Boys are expected to score .002 points higher than girls because of non-invariance. Non-invariance accounted for .2% of the observed mean difference (M_{diff} =.886, Cohen's d = .31) in boys' and girls' Surgency scores. For fathers, ratings of Surgency were totally invariant across girls and boys (i.e., strict invariance was achievable).

Summary—When examining mothers' and fathers' ratings together it was possible to achieve partial strong invariance across girls and boys for the Effortful Control and Surgency dimensions. For each of these dimensions, only one scale contained a non-invariant intercept. The Negative Affectivity dimension however was largely non-invariant across child sex. Although it was possible to constrain the factor loadings, all intercepts were non-invariant. Considering parental informants separately, fathers' ratings were slightly more invariant than mothers' ratings. Fathers' ratings of Effortful Control and Surgency demonstrated strong invariance across boys and girls, but mothers' ratings demonstrated only partial strong invariance. Both parents' ratings of Negative Affectivity demonstrated partial strong invariance across boys and girls.

Measurement Invariance Across Parents

Effortful Control—The specification of the baseline Effortful Control model for mothers and fathers was similar to what was used when examining invariance across girls and boys. However, because this was a within-group confirmatory factor analysis (i.e., mothers and fathers reported on the same child), the mother and father Effortful Control factors were allowed to correlate. Furthermore, mother and father paired residual variances were allowed to correlate.

The baseline fit of the Effortful Control model for mothers and fathers was: χ^2 = 98.43, p < .001; RMSEA = .10; SRMR = .05; CFI = .93; TLI = .87 (N = 605). Following the estimation of the baseline model, factor loadings were constrained to equality across parents (fathers' factor variance was allowed to freely vary). These constraints did not result in a significant degrade in model fit ($\Delta\chi^2$ =.99, df = 3, p = .80; ΔCFI = .002; $\Delta RMSEA$ = -.01), supporting weak invariance. Intercepts were then constrained across parents (fathers' factor mean was allowed to vary) to test strong invariance. These constraints resulted in a model that fit significantly worse than the model with equal loadings only ($\Delta\chi^2$ = 20.41, df = 2, p <.001; ΔCFI = -.015; $\Delta RMSEA$ = -.002). The low intensity pleasure scale was identified as contributing to model misfit and its intercept was subsequently freed. This model did not fit significantly worse than the equal loadings only model ($\Delta\chi^2$ = 3.54, df = 2, p = .17; ΔCFI = -.002; $\Delta RMSEA$ = -.003). Last, all unique variances -with the exception of the low intensity

pleasure unique variance - were constrained to equality to test partial strict invariance. This model did not fit significantly worse than the model with partially invariant intercepts ($\Delta\chi^2$ =1.65, *df* = *3, p* = .65; *ΔCFI* = .002; *ΔRMSEA* = -.006). The parameter estimates for the most justifiably constrained model can be found in Table 5.

The highest dMACS value for an individual Effortful Control scale was the value for low intensity pleasure (see Table 9). The value of .23 suggests a small to moderate effect. 60% of the observed mean difference in mothers' and fathers' low intensity pleasure scores was a consequence of non-invariant intercepts. The overall effect of non-invariance on the measurement of Effortful Control was only .06. Still, mother reported Effortful Control scores are expected to be around .22 points higher than father reported Effortful Control scores simply because of non-invariance. Measurement non-invariance thus accounted for 43% of the observed mean difference (M_{diff} = .51, Cohen's *d* = .24) in parents' Effortful Control scores.

Examining girls and boys separately: Measurement invariance across parents' reports of Effortful Control was examined separately for boys (*N* = 318) and girls *(N* = 286). For girls, it was only justifiable to constrain factor loadings across mothers and fathers (i.e., only weak invariance was supported). It was not possible to constrain any intercepts or residual variances to equality without observing a significant degrade in model fit. Non-invariance was especially pronounced in the attentional focusing and inhibitory control scales (dMACS = 1.21 and .54, respectively). The overall effect of non-invariance was .36 (to obtain this estimate, one set of intercepts had to be constrained to equality; perceptual sensitivity was selected as its dMACS estimate was the lowest at .07). Mothers' reports of Effortful Control are expected be .44 points higher than fathers' reports because of non-invariance alone; non-invariance here accounted for 65% of the observed mean difference (M_{diff} = .67; Cohen's *d* = .30) in parents' reports of their daughters' Effortful Control. For boys, it was possible to constrain all factor loadings and scale intercepts to equality across parents (i.e., strong invariance was achievable). It was not possible though to constrain residual variances to equality.

Negative Affectivity—The baseline fit of the Negative Affectivity model for mothers and fathers was: χ^2 = 162.86, *p* < .001; RMSEA = .09; SRMR = .05; CFI = .93; TLI = .88 (*N* = 605). Constraining mothers' and fathers' factor loadings to equality (and allowing the fathers' factor variance to freely vary) did not result in a significant degrade in fit from the baseline model ($\Delta\chi^2$ = 3.44, *df* = 4, *p* = .49; *ΔCFI* = .001; *ΔRMSEA* = -.005), supporting weak invariance. Constraining mother and father intercepts to equality (and allowing fathers' factor mean to freely estimate) to test strong invariance led to a significant degrade in model fit ($\Delta\chi^2$ = 23.79, *df* = 4, *p* < .001; *ΔCFI* = -.011; *ΔRMSEA* = .001). Soothability was identified as the scale most likely contributing to misfit. A partially strong invariant model in which all intercepts except for soothability were constrained was tested against the equal factor loadings model; the partially strong invariant model did not fit significantly worse ($\Delta\chi^2$ = 5.50, *df* = 3, *p* = .14; *ΔCFI* = -.002; *ΔRMSEA* = -.003). All unique variances – except soothability – were constrained to equality across parents. This model did not fit significantly worse than the model with partially invariant intercepts ($\Delta\chi^2$ = 4.06, *df* = 4, *p*

= .40; ΔCFI = .000; $\Delta RMSEA$ = -.004), supporting the presence of partial strict invariance. The parameter estimates of the most justifiably constrained model can be found in Table 6.

The dMACS value for soothability was the highest observed for the Negative Affectivity scales (see Table 9). Although not an especially large value at .17, intercept non-invariance still accounted for 118% of the observed mean difference in parent reported soothability. The overall effect of non-invariance was also small, at .04. Despite this, measurement non-invariance was responsible for 100% of the small mean difference in mothers' and fathers' ratings of Negative Affectivity (M_{diff} = .01; Cohen's d = .004).

Examining girls and boys separately: Invariance was tested separately for mothers' and fathers' ratings of girls' (N = 286) and boys' (N = 318) Negative Affectivity. For girls, it was possible to constrain all loadings and most intercepts to equality across parents (partial strong invariance). It was not possible to constrain residual variances to equality. Only the soothability intercept evidenced significant non-invariance (soothability dMACS = .34). The overall effect of non-invariance for the Negative Affectivity scale was .07. Measurement non-invariance accounted for 43% of the observed difference (M_{diff} =.26; Cohen's d = .14) in parents' reported Negative Affectivity means (mothers' ratings are expected to be.11 points less than fathers because of non-invariance). For boys, it was possible to constrain all factor loadings, intercepts, and residual variances to equality across parents (i.e., strict invariance was achievable).

Surgency—The baseline fit of the Sugency model for mothers and fathers was: χ^2 = 243.00, p < .001; RMSEA = .11; SRMR = .08; CFI = .91; TLI = .86 (N = 605). Constraining all factor loadings to equality (and allowing the fathers' variance to freely vary) to test weak invariance resulted in a significant degrade in fit ($\Delta\chi^2$ = 15.82, df = 4, p = .003; ΔCFI = -.006; $\Delta RMSEA$ = -.004). Shyness was subsequently identified as the scale contributing the most to degrade in fit. When the shyness factor loadings were unconstrained across parents, the model did not fit significantly worse than the baseline model ($\Delta\chi^2$ = 6.53, df = 3, p = .09; ΔCFI = -.002; $\Delta RMSEA$ = -.004), supporting partial weak invariance. Constraining all intercepts - except for shyness - to equality (and freeing the fathers' factor mean) to test intercept invariance resulted in a model that fit significantly worse than the model with partial loading invariance ($\Delta\chi^2$ = 42.25, df = 3, p < .000; ΔCFI = -.017; $\Delta RMSEA$ = .004). The smiling and laughter, and positive anticipation, scales ostensibly contributed the most to this degradation in fit. Only when both of these intercepts were freed was there no degradation in fit from the model with partially invariant loadings ($\Delta\chi^2$ = .26, df = 1, p = .61; ΔCFI = .000; $\Delta RMSEA$ = -.002). Constraining the residual variances of the high intensity pleasure and impulsivity scales (the only two scales with invariant loadings and intercepts) did not significantly degrade model fit ($\Delta\chi^2$ = 5.74, df = 2, p = .06; ΔCFI = -.001; $\Delta RMSEA$ = -.002), supporting partial strict invariance. The parameters of the most justifiably constrained model can be found in Table 7.

The majority of the scales that make up the Surgency dimension evidenced non-invariance at some level. According to the dMACS values, there was a moderate degree of non-invariance in the smiling and laughter scale, and a small degree of non-invariance in the shyness and positive anticipation scales (see Table 9). Intercept non-invariance accounted for 94% of the

observed mean difference in smiling and laughter scores, and 82% of the observed mean difference in positive anticipation scores[6]. Despite the prevalence of non-invariance at the scale level, the overall effect of non-invariance on the Surgency dimension was fairly small (.06). However, non-invariance accounted for most (77%) of the observed mean difference (M_{diff} = .39; Cohen's *d* = .16) in Surgency scores; mothers' are expected to report Surgency scores .3 points higher than fathers because of non-invariance alone.

Examining girls and boys separately: Invariance across parents was tested separately for boys (*N* = 318) and girls (*N* = 286). For girls, it was possible to constrain all loadings to equality across parents, and most of the intercepts (partial strong invariance). It was not possible to constrain the smiling and laughter, and positive anticipation intercepts, or the residual variances. The dMACS values for smiling and laughter, and positive anticipation were .30 and .31, respectively, indicating that the degree of non-invariance is moderate. The overall effect of non-invariance was small however, at .08. Non-invariance still accounted for most (73%) of the observed mean difference between parents (M_{diff} = .60 Cohen's *d* = .24); mothers are expected to report girls as .44 points higher on Surgency than fathers because of non-invariance.

For boys, the shyness and impulsivity loadings, and the shyness, impulsivity, and smiling and laughter intercepts and residual variances all evidenced non-invariance. In other words, it was only possible to set three factor loadings (high intensity pleasure, smiling and laughter, and positive anticipation), and two intercepts/residual variances (high intensity pleasure and positive anticipation) to equality across parents (partial strict invariance). The dMACS values indicated that the smiling and laughter, impulsivity, and shyness scales are all moderately affected by non-invariance (dMACS = .24, .28, and .30, respectively). Despite this, the overall effect of non-invariance was only .06. Still, non-invariance accounted for all (300%) of the observed mean difference (M_{diff} = .07, Cohen's *d* = .10); mothers' reports of their sons' Surgency are expected to be .21 points higher than fathers' reports because of non-invariance.

Summary—Across mothers and fathers, all dimensions demonstrated partial strong invariance. There were more non-invariant intercepts in the Surgency dimension than in the Effortful Control and Negative Affectivity dimensions. When split by child sex, mothers' and fathers' ratings of girls demonstrated considerably more non-invariance than their ratings of boys. Ratings of girls' Effortful Control were especially non-invariant across parents. Still, although ratings of boys' Effortful Control and Negative Affectivity were largely invariant across parents, there was substantial non-invariance observed in the Surgency dimension.

Discussion

Measurement invariance was evaluated across boys and girls (as targets), and mothers and fathers (as informants) for the three higher-order dimensions of temperament measured by the CBQ: Effortful Control, Negative Affectivity, and Surgency. In addition to testing for the presence (or absence) of invariance, we calculated a number of effect size indicators to

[6]this metric was not calculated for the shyness scale because it requires loading equality

provide insight into the practical importance of any non-invariance. All told, the current study illustrates how researchers interested in child temperament can evaluate measurement invariance to provide additional insights into the psychometric properties of the CBQ.

Invariance by Child Sex

The existence and size of sex differences in temperament is an interesting and important question in the literature given that these are early emerging individual differences in disposition. Previous studies suggest that girls score higher than boys on Effortful Control, while boys score higher than girls on Surgency (Else-Quest et al., 2006). However, work along these lines often does not consider issues of measurement equivalence when comparing scores for girls and boys. There was evidence in the current study that some CBQ scales work differently for boys and girls when parents' ratings are combined. The extent of this lack of equivalence differed across dimensions of temperament, with it being most noteworthy for Negative Affectivity. The measurement of Effortful Control and Surgency was largely equivalent across boys and girls, thus previous reports of sex differences in these domains are not likely to be explained away as measurement artifacts. In contrast, the scale intercepts for all indicators of Negative Affectivity were different across boys and girls. This makes mean-level comparisons difficult to interpret because the observed scores will not be the same for girls and boys even if they are at the same latent level of Negative Affectivity. This suggests that sex differences in Negative Affectivity should be interpreted with some caution, especially if the results here replicate.

Mothers' and fathers' reports of boys versus girls were also examined separately. Results for maternal reports of temperament were similar to the results obtained when maternal and paternal reports were averaged. The one departure was that maternal reports of Negative Affectivity had partially invariant intercepts. The anger/frustration, discomfort, and soothability intercepts were non-invariant, but the fear and sadness intercepts were equivalent across boys and girls. Paternal reports of Negative Affectivity were also partially invariant across boys and girls. All factor loadings were equivalent, but the sadness and anger/frustration intercepts were not. Thus, when considering Negative Affectivity, mother- and father-reports may exhibit different psychometric properties.

Regarding the practical importance of all these findings, the average overall effect of non-invariance (looking across the composite and each parent individually) was .07 for Effortful Control, .21 for Negative Affectivity, and .03 for Surgency. Because of non-invariance, boys or girls were on average expected to score .1 point higher than the other group on Effortful Control, .29 points higher on Negative Affectivity, and .04 points higher on Surgency (standard deviations range from roughly 2 to 2.5 across dimensions). Measurement artifacts accounted for 11%, 73%, and 3% of the observed mean differences in boys' and girls' Effortful Control, Negative Affectivity, and Surgency scores, respectively. These effect sizes suggest small to moderate impacts of non-invariance.

On more theoretical grounds, our results suggest that previously reported sex differences in temperament may be only slightly distorted. In the current study, the Cohen's *d*s for the observed mean differences between girls and boys on Effortful Control, Negative Affectivity, and Surgency were .38, .22, and .47, respectively (see Table 1). Correcting for

non-invariance[7], these values become .33, .29, and .42. Because of non-invariance, the original effect sizes for the observed mean differences were between11% and 25% distorted. This indicates that failure to control for measurement non-invariance can lead to over- or under-estimated sex differences.

Still, even though the effect size estimates changed here after correcting for non-invariance, the general patterns were more or less consistent with the meta-analytic results. In other words, the current results point out that although measurement non-invariance creates some distortion when evaluating sex differences, there is no reason to question the main meta-analytic conclusions (Else-Quest et al., 2006). Although the origins and interpretations of sex differences are contentious (Hyde, 2014), such differences in temperament in children seem to persist when taking certain psychometric issues into account. Moreover, these results might be important when evaluating theories for which sex differences in traits are a primary source of support, such as models proposing that higher rates of mood and anxiety disorders in females are attributable to their higher levels of NA (e.g., Fanous et al., 2002).

Invariance by Parental Informant

When looking at girls and boys combined, mothers' and fathers' reports were generally equivalent. Across the entire CBQ, only 4 scales had non-invariant intercepts: low intensity pleasure, soothability, smiling and laughter, and approach/anticipation. A slightly more complicated picture emerged when invariance was examined separately for girls and boys. Mothers' and fathers' ratings of their daughters' Effortful Control were non-invariant and therefore not directly comparable. It was possible to achieve partial strong invariance for the dimensions of Negative Affectivity and Surgency however. For boys, it was possible to achieve weak and strong invariance for Effortful Control and Negative Affectivity, meaning that scores on these dimensions are comparable across parents. Mothers' and fathers' ratings of their sons' Surgency were partially non-invariant though.

Although the current results should be replicated on an even larger sample, it is worth speculating why mothers and fathers might use the CBQ scales differently when reporting on their daughters. One possibility is that gender stereotypes could possibly have psychometric consequences for parents' reports of child behavior (e.g., Gartstein et al., 2009). Also, if fathers spend less time with daughters than mothers (Sayer, Bianchi, & Robinson, 2004; Harris & Morgan, 1991), they could have less experience with their daughter's behavior, leading to more discrepant item interpretations.

The average overall effect of non-invariance, for boys and girls examined together and separately, was .14 for Effortful Control, .04 for Negative Affectivity, and .07 for Surgency. These suggest small effects. However, on average, 36%, 48%, and 150% of the observed mean differences in parents' Effortful Control, Negative Affectivity, and Surgency scores were due to non-invariance alone. This is because one parent is on average expected to rate .22, .04, or .32 points higher than the other parent on Effortful Control, Negative Affectivity, and Surgency due to non-invariance (standard deviations for the dimensions are between 2

[7]E.g., girls are expected to score .12 points higher than boys simply as a function of non-invariance, thus .12 points were deducted from the girls' mean and Cohen's *d* was re-calculated

and 3). The apparent discrepancy between these measures of effect size is due to the fact that the mean differences between mothers' and fathers' ratings are generally small (see Table 1). The overall effects of non-invariance may be modest, but because the observed mean difference is likewise small, it is easy for non-invariance to account for most of it. Put differently, mothers and fathers tend to give the same scores when rating the same child, but a substantial part of any difference is likely a measurement artifact.

The results regarding invariance across mothers and fathers are informative for understanding how parents rate child temperament, and why their reports might diverge. To date, there has been much theorizing about why different informants might disagree when rating the characteristics of the same target (e.g., Funder, 1995; Kenny, 1994; De Los Reyes & Kazdin, 2005), but these treatments rarely consider the potential role of measurement invariance. Our results suggest psychometric differences in reports of temperament from mothers and fathers might contribute to some observed differences. Indeed, scales that proved to be the most non-invariant (e.g., low intensity pleasure) generally exhibited the largest observed mean differences, and the smallest correlations between parents (See Table 1).

Yet on the whole the magnitude of non-invariance was not overwhelming. Thus a fair amount of the disagreement between parents' reports represents “true” disagreement. That is, disagreement between parents is more than simply a function of parents using the scales (and presumably the items) in different ways (e.g., what counts as “a lot” of a behavior for mothers might not count as “a lot” of the same behavior for fathers). To illustrate this point, after correcting for non-invariance, the correlations between parents' ratings did not increase dramatically from what was initially observed. The inter-parent correlation for Effortful Control, originally .53, became .54. For Negative Affectivity and Surgency the correlations changed from .54 and .63 to .52 and .67. The fact that mother-father disagreement cannot be reduced purely to measurement non-invariance means that it is necessary to consider other sources of disagreement between parents (e.g., attributional style and psychopathology; De Los Reyes & Kazdin, 2005).

Implications for Parent Report Measurement of Child Temperament

This study contributes to the large and growing body of research evaluating the parent report approach to measuring child temperament. Parental reports are economical, ecologically valid, and predictive of a wide range of important child outcomes (Rothbart & Goldsmith, 1985). Thus, there are good reasons why researchers and clinicians use parent reports. However, the parent report method is not without critics (e.g., Richters, 1992; Seifer, 2002; Vaughn, Bradley, Joffe, Seifer, & Barglow, 1987). Some have worried for example that parent-reports capture variance related to parental psychopathology or distress (e.g., Fergusson, Lynksey, & Horwood, 1993; Muller, Achtergarde, & Furniss, 2011; Durbin & Wilson, 2012).

Although these points do not necessarily undermine the utility and value of parent report measures, they do suggest it is important to critically evaluate parental reports. Examining measurement invariance is one element of such a comprehensive evaluation. The results of this study suggest that measurement invariance cannot be assumed in parents' reports of

child temperament as has historically been done. Although sex differences and inter-parental agreement were not grossly distorted, measurement non-invariance accounted for some of the observed differences. Researchers should therefore test for invariance in future studies. Effect size measures like those used here can be used to assess the actual impact of non-invariance to help quantify the issue. Negligible effects can probably be ignored in many cases. Moreover, it might be possible to adjust observed statistics to account for non-invariance. For example, the observed mean difference between girls' and boys' Effortful Control in our study was .51, but the expected difference due to non-invariance alone was .12, meaning that the actual mean difference was closer to .40 than .50. Alternatively, non-invariant scales/items could be dropped from the final score calculation, or scores that account for non-invariance between groups could be generated (the models presented here demonstrating partial strong invariance are examples of models that could generate scores corrected for non-invariance; Tay et al., 2015).

Finally, it is importantly to briefly consider the value of evaluating the CBQ specifically (Rothbart et al., 2001). We believe a focus on this instrument is important for at least two major reasons. First, the CBQ is the most commonly used parent-report measure in the literature, making the results here widely applicable. Second, the current findings are useful for pinpointing problematic scales and dimensions, which could help in refining this popular questionnaire. For example, across the majority of analyses the low intensity pleasure scale of the Effortful Control dimension evidenced non-invariance. It appears that this scale is especially vulnerable to differential usage, making it a prime candidate for revision. The smiling and laughter, and positive anticipation scales also frequently demonstrated non-invariance. This is notable given that these scales were the least likely to load on their theoretical dimension in this study, and others (e.g. Rothbart et al., 2001). This fact in addition to their recurrent non-invariance indicates that these scales are worth further evaluating and possibly revising.

Limitations and Future Directions

Despite the strengths of this study, it is important to consider some limitations and caveats about our approach. Foremost, we focused on invariance at the level of the “Big Three” higher-order dimensions of the CBQ. Other levels of analysis are possible including an item by item investigation of specific scales. Differential item functioning was not explicitly tested before scale composites were computed. Although the effects of non-invariance tend to wash out when moving up levels of measurement (e.g., in the current study, dMACS values for scales are generally larger than the corresponding Cohen's d for non-invariance at the dimensional level), and the higher order dimensions of the CBQ are typically what are used in research contexts, invariance at the item level should be investigated in future studies. Second, the results are constrained by the samples and measure used in this research. Other measures of child temperament may be more or less invariant than the CBQ. Moreover, by restricting our investigation to the CBQ we were also restricting the age range of the children that could be included in this study. The CBQ is intended for children aged 3 to 7, thus our results cannot speak directly to children at different developmental stages. More work will have to be done to explore measurement invariance in questionnaires developed for other age groups. Finally, in the current study we focused exclusively on

parents as informants of temperament. Reports of temperament can also be collected from other informants such as teachers, and even self-report with older children. The extent that measures of temperament are invariant across these other reporters should be examined in the future.

In addition to addressing the gaps outlined above, future studies can build more directly on the results presented here. Again, future studies should examine measurement invariance at the item level. Such investigations, in addition to providing more information about how the CBQ functions, could reveal exactly the type of content that is non-invariant across groups, which may in turn inform hypotheses regarding the source(s) of non-invariance. Why, for example, are parents' ratings of their daughters' but not their sons' non-invariant, and why are measures of Negative Affectivity especially non-invariant across boys and girls? Examining item content could help to confirm or disconfirm hypotheses regarding the origins of non-invariance. Models that treat factor loadings and intercepts as random effects (e.g., Bauer & Hussong, 2009; Muthen & Asparouhov, 2013), or are able to isolate and predict rater specific variance (e.g., Bauer et al., 2013), could also be used to further explore the origins of non-invariance. Such models allow for more direct investigations into the extent to which various exogenous variables (e.g., parent depressive symptomatology) contribute to non-invariance.

Conclusion

Parental reports are the most common source of information on child temperament, and they have been used to examine a number of substantively meaningful relations, such as how boys and girls differ in their early dispositions. In spite of their popularity, the extent to which parental reports display measurement invariance across theoretically meaningful groups of targets and informants has been relatively underexplored. Knowledge regarding measurement invariance is invaluable when ratings across distinct groups are compared or integrated in some way. In the current study measurement invariance was examined in the popular CBQ across girls and boys as targets, and mothers and fathers as informants. Several instances of non-invariance were revealed, and a variety of effect size metrics indicated that the magnitude of non-invariance ranged from negligible to moderate. Therefore, we believe it is useful to test for measurement invariance in future studies of childhood temperament.

Acknowledgments

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). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. We would also like to thank Christopher D. Nye for his invaluable assistance with the dMACS program.

References

Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. Psychological Bulletin. 1987; 101(2):213–232. [PubMed: 3562706]

Allison, PD. Missing Data. In: Millsap, RE.; Maydeu-Olivares, A., editors. The SAGE Handbook of Quantitative Methods in Psychology. London: SAGE Publications Ltd; 2009. p. 72-89.

Bauer DJ, Howard AL, Baldasaro RE, Curran PJ, Hussong AM, Chassin L, Zucker RA. A trifactor model for integrating ratings across multiple informants. Psychological Methods. 2013; 18(4):475–493. DOI: 10.1037/a0032475 [PubMed: 24079932]

Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. Psychological Methods. 2009; 14(2):101–125. DOI: 10.1037/a0015583 [PubMed: 19485624]

Borsboom D. The attack of the psychometricians. Psychometrika. 2006; 71(3):425–440. DOI: 10.1007/s11336-006-1447-6 [PubMed: 19946599]

Browne, MW.; Cudeck, R. Testing Structural Equation Models. SAGE Publications Ltd; 1993. Alternative ways of testing model fit; p. 136-162.

Brown, TA. Confirmatory Factor Analysis for Applied Research. New York, NY: The Guilford Press; 2006. CFA with equality constraints, multiple groups, and mean structures; p. 236-319.

Card NA, Stucky BD, Sawalani GM, Little TD. Direct and indirect aggression during childhood and adolescence: A meta-analytic review of gender differences, intercorrelations, and relations to maladjustment. Child Development. 2008; 79(5):1185–1229. [PubMed: 18826521]

Caspi A, Moffitt TE, Newman DL, Silva PA. Behavioral observations at age 3 years predict adult psychiatric disorders: Longitudinal evidence from a birth cohort. Archives of General Psychiatry. 1996; 53(11):1033–1039. [PubMed: 8911226]

Clark DA, Donnellan MB, Robins RW, Conger R, D. Early adolescent temperament, parental monitoring, and substance use in Mexican-origin adolescents. Journal of Adolescence. 2015; 41:121–131. DOI: 10.1016/j.adolescence.2015.02.010 [PubMed: 25841175]

Clark, LA.; Watson, D. Handbook of Personality: Theory and Research. New York, NY: The Guilford Press; 2008. Temperament: An organizing paradigm for trait psychology; p. 265-286.

Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal. 2007; 14(3):464–504. DOI: 10.1080/10705510701301834

Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal. 2002; 9(2):233–255. DOI: 10.1207/S15328007SEM09025

Creemers HE, Dijkstra JK, Vollebergh WAM, Ormel J, Verhulst FC, Huizink AC. Predicting life-time and regular cannabis use during adolescence; the roles of temperament and peer substance use: The TRAILS study. Addiction. 2010; 105:699–708. DOI: 10.1111/j.1360-0443.2009.02819.x [PubMed: 20148797]

de Ayala, RJ. The Theory and Practice of Item Response Theory. New York, NY: The Guilford Press; 2009. Differential item functioning; p. 323-343.

De Los Reyes A, Kazdin AE. Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. Psychological Bulletin. 2005; 131(4):483–509. DOI: 10.1037/0033-2909.131.4.483 [PubMed: 16060799]

Duhig AM, Renk K, Epstein MK, Phares V. Interparental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. Clinical Psychology: Science and Practice. 2000; 7(4):435–453.

Durbin CE, Wilson S. Convergent validity of and bias in maternal reports of child emotion. Psychological assessment. 2012; 24(3):647–660. DOI: 10.1037/a0026607 [PubMed: 22149326]

Elder, GH.; Conger, RD. Children of the Land: Adversity and success in rural America. Chicago, IL: University of Chicago Press; 2000.

Else-Quest NM, Hyde JS, Goldsmith HH, Van Hulle CA. Gender differences in temperament: A meta-analysis. Psychological Bulletin. 2006; 132(1):33–72. DOI: 10.1037/0033-2909.132.1.33 [PubMed: 16435957]

Embretson, SE.; Reise, SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2000. IRT applications: DIF, cat, and scale analysis; p. 249-272.

Fanous A, Gardner CO, Prescott CA, Cancro R, Kendler KS. Neuroticism, major depression and gender: A population-based twin study. Psychological Medicine. 2002; 32:719–728. DOI: 10.1017/S003329170200541X [PubMed: 12102386]

Fergusson DM, Lynskey MT, Horwood LJ. The effect of maternal depression on maternal ratings of child behavior. Journal of Abnormal Child Psychology. 1993; 21(3):245–269. [PubMed: 8335763]

Funder DC. On the accuracy of personality judgment: A realistic approach. Psychological Review. 1995; 102:652–670. DOI: 10.1037/0033-295X.102.4.652 [PubMed: 7480467]

Gartstein MA, Bridgett DJ, Dishion TJ, Kaufman NK. Depressed mood and maternal report of child behavior problems: Another look at the depression-distortion hypothesis. Journal of Applied Developmental Psychology. 2009; 30:149–160. DOI: 10.1016/j.appdev.2008.12.001 [PubMed: 20161323]

Gartstein, MA.; Bridgett, DJ.; Low, CM. Asking questions about temperament: Self- and other-report measures. In: Zentner, M.; Shiner, RL., editors. Handbook of Temperament. New York: The Guilford Press; 2012. p. 183-208.

Goldsmith HH, Buss AH, Plomin R, Rothbart MK, Thomas A, Chess S, et al. McCall RB. Roundtable: What is temperament? Four approaches. Child Development. 1987; 58:505–529. [PubMed: 3829791]

Goldsmith, HH.; Gagne, JR. Behavioral assessment of temperament. In: Zentner, M.; Shiner, RL., editors. Handbook of Temperament. New York: The Guilford Press; 2012. p. 209-228.

Graham, JW.; Coffman, DL. Handbook of Structural Equation Modeling. New York, NY: The Guilford Press; 2012. Structural equation modeling with missing data; p. 277-295.

Gretarsson SJ, Gelfand DM. Mothers' attributions regarding their children's social behavior and personality characteristics. Developmental Psychology. 1988; 24(2):264–269.

Harris KM, Morgan SP. Fathers, sons, and daughters: Differential parental involvement in parenting. Journal of Marriage and the Family. 1991; 53(3):531–544.

Hyde JS. Gender similarities and differences. Annual Review of Psychology. 2014; 65:373–398. DOI: 10.1146/annurev-psych-0102130115057

Kenny, DA. Interpersonal perception: A social relations analysis. New York, NY: The Guilford Press; 1994. Consensus; p. 50-81.

Klein, DN.; Dyson, MW.; Kujawa, AJ.; Kotov, R. Temperament and internalizing disorders. In: Zentner, M.; Shiner, RL., editors. Handbook of Temperament. New York: The Guilford Press; 2012. p. 541-561.

Kotelnikova Y, Olino TM, Klein DN, Kryski KR, Hayden EP. Higher- and lower-order factor analyses of the Children's behavior questionnaire in early and middle childhood. Psychological Assessment. 2015; doi: 10.1037/pas0000153

Lo SL, Vroman LN, Durbin CE. Ecological validity of laboratory assessments of child temperament: Evidence from parent perspectives. Psychological Assessment. 2014; 27(1):280–290. DOI: 10.1037/pas0000033 [PubMed: 25330108]

Meade AW, Johnson EC, Braddy PW. Power and sensitivity of alternative fit indices in tests of measurement invariance. Journal of Applied Psychology. 2008; 93(3):568–592. DOI: 10.101037/0021-9010.93.3.568 [PubMed: 18457487]

Mervielde, I.; De Pauw, SSW. Models of Child Temperament. In: Zentner, M.; Shiner, RL., editors. Handbook of Temperament. New York: The Guilford Press; 2012. p. 21-40.

Millsap, RE.; Olivera-Aguilar, M. Handbook of Structural Equation Modeling. New York, NY: The Guilford Press; 2012. Investigating measurement invariance using confirmatory factor analysis; p. 380-393.

Moffitt TE, Arsenault L, Belsky D, Dickson N, Hancox RJ, Harrington H, et al. Caspi A. A gradient of childhood self-control predicts health, wealth, and public safety. PNAS. 2011; 108(7):2693–2698. [PubMed: 21262822]

Muller JM, Achtergarde S, Furniss T. The influence of maternal psychopathology on ratings of child psychiatric symptoms: An SEM analysis on cross-informant agreement. European Child & Adolescent Psychiatry. 2011; 20:241–252. DOI: 10.1007/s00787-011-0168-2 [PubMed: 21416135]

Muthen, BO.; Asparouhov, T. New methods for the study of measurement invariance with many groups. 2013. Retrieved May 20, 2015, from www.statmodel.com.

Muthen, LK.; Muthen, BO. Mplus user's guide. Seventh. Los Angeles, CA: Muthen & Muthen; 1998-2012.

Neppl TK, Donnellan MB, Scaramella LV, Widaman KF, Spilman SK, Ontai LL, Conger RD. Differential stability of temperament and personality from toddlerhood to middle childhood. Journal of Research in Personality. 2010; 44:386–396. DOI: 10.1016/j.jrp.2010.04.004 [PubMed: 20634996]

Nye CD, Allemand M, Gosling SD, Potter J, Roberts BW. Personality trait differences between young and middle-aged adults: Measurement artifacts or actual trends? Journal of Research in Personality. in press.

Nye CD, Drasgow F. Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. Journal of Applied Psychology. 2011; 96(5): 966–980. DOI: 10.1037/a0022955 [PubMed: 21463015]

Olino TM, Durbin CE, Klein DN, Hayden EP, Dyson MW. Gender Differences in Young Children's Temperament Traits: Comparisons Across Observational and Parent-Report Methods. Journal of personality. 2013; 81(2):119–129. [PubMed: 22924826]

Osterlind, SJ.; Everson, HT. Differential item functioning. SAGE Publications; 2009.

Richters JE. Depressed mothers as informants about their children: A critical review of of the evidence for distortion. Psychological Bulletin. 1992; 112(3):485–499. [PubMed: 1438639]

Robins RW, Donnellan MB, Widaman KF, Conger RD. Evaluating the link between self-esteem and temperament in Mexican origin early adolescents. Journal of Adolescence. 2010; 33:403–410. DOI: 10.1016/j.adolescence.2009.07.009 [PubMed: 19740537]

Rothbart MK. Temperament, development, and personality. Current directions in psychological science. 2007; 16(4):207–212.

Rothbart, MK. Becoming who we are: Temperament and personality in development. New York, NY: The Guilford Press; 2011.

Rothbart MK, Ahadi SA, Hershey KL. Temperament and social behavior in childhood. Merrill-Palmer Quarterly. 1994:21–39.

Rothbart MK, Ahadi SA, Hershey KL, Fisher P. Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. Child Development. 2001; 72(5):1394–1408. [PubMed: 11699677]

Rothbart MK, Goldsmith HH. Three approaches to the study of infant temperament. Developmental Review. 1985; 5:237–260.

Sayer LC, Bianchi SM, Robinson JP. Are parents investing less in children? Trends in mothers' and fathers' time with children. American Journal of Sociology. 2004; 110(1):1–43.

Seifer R. What do we learn from parent reports of their children's behavior? Commentary on Vaughn et al.'s critique of early temperament assessments. Infant Behavior and Development. 2002; 25:117–120.

Shiner RL, Buss KA, McClowry SG, Putnam SP, Saudino KJ, Zentner M. What Is Temperament Now? Assessing Progress in Temperament Research on the Twenty-Fifth Anniversary of Goldsmith et al. Child Development Perspectives. 2012; 6(4):436–444.

Shiner, RL.; DeYoung, CG. The structure of temperament and personality traits: A developmental perspective. In: Zelazo, P., editor. Oxford handbook of developmental psychology. New York: Oxford University Press; 2013. p. 113-141.

Stautz K, Cooper A. Impulsivity-related personality traits and adolescent alcohol use: A meta-analytic review. Clinical Psychology Review. 2013; 33:574–592. DOI: 10.1016/j.cpr.2013.03.003 [PubMed: 23563081]

Sulik MJ, Huerta S, Zerr AA, Eisenberg N, Spinrad TL, Valiente C, et al. Taylor HB. The factor structure of effortful control and measurement invariance across ethnicity and sex in a high-risk sample. Journal of psychopathology and behavioral assessment. 2010; 32(1):8–22. [PubMed: 20593008]

Tackett, JL.; Martel, MM.; Kushner, SC. Temperament, externalizing disorders, and attention-deficit/hyperactivity disorder. In: Zentner, M.; Shiner, RL., editors. Handbook of Temperament (pp -). New York: The Guilford Press; 2012. p. 562-580.

Tay L, Meade AW, Cao M. An overview and practical guide to IRT measurement equivalence analysis. Organizational Research Methods. 2015; 18(1):3–46. DOI: 10.1177/1094428114553062

Tellegen, A.; Waller, NG. The SAGE handbook of personality theory and assessment. Vol. 2. Thousand Oaks, CA: SAGE Publishing Inc; 2008. Exploring personality through test construction: development of the Multidimensional Personality Questionnaire; p. 261-292.

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

Vaugh BE, Bradley CF, Joffe LS, Seifer R, Barglow P. Maternal characteristics measured prenatally are predictive of ratings of temperamental “difficulty” on the Carey Infant Temperament Questionnaire. Developmental Psychology. 1987; 23(1):152–161.

Zumbo BD. Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly. 2007; 4(2):223–233.

Psychol Assess. Author manuscript; available in PMC 2017 December 01.

Table 1

CBQ Descriptive Statistics

	Girls		Boys			Mothers		Fathers			
	M	**SD**	**M**	**SD**	***d***	**M**	**SD**	**M**	**SD**	***d***	**r**
Smiling and Laughter	5.65	.63	5.72	.58	.12	5.76	.66	5.58	.70	.27	.51
High Intensity Pleasure	4.89	.73	5.23	.73	.47	5.07	.81	5.03	.81	.06	.62
Approach	5.24	.54	5.32	.53	.15	5.31	.62	5.21	.60	.14	.35
Soothability	4.74	.42	4.73	.60	.02	4.79	.71	4.67	.71	.16	.45
Impulsivity	4.56	.70	4.80	.66	.35	4.69	.76	4.68	.71	.02	.63
Anger/Frustration	4.55	.75	4.65	.75	.13	4.59	.84	4.58	.81	.01	.53
Fear	3.93	.79	3.40	.85	.65	3.81	.94	3.79	.87	.02	.53
Discomfort	4.28	.74	3.97	.72	.43	4.11	.85	4.06	.78	.06	.50
Sadness	4.02	.59	3.89	.62	.21	3.93	.72	3.93	.66	0	.39
Shyness	3.66	1.14	3.28	1.05	.35	3.44	1.24	3.48	1.11	.04	.69
Attentional Focusing	4.57	.68	4.35	.71	.32	4.49	.81	4.41	.77	.10	.48
Inhibitory Control	4.57	.74	4.31	.71	.36	4.46	.82	4.38	.77	.11	.58
Low Intensity Pleasure	5.58	.52	5.31	.56	.50	5.52	.66	5.32	.62	.27	.35
Perceptual	4.91	.65	4.80	.75	.16	4.89	.80	4.76	.76	.16	.43
Effortful Control	19.62	1.95	18.76	2.54	.38	19.37	2.24	18.86	2.08	.24	.53
Negative Affectivity	20.03	2.60	19.48	2.52	.22	19.67	2.92	19.68	2.69	0	.54
Surgency	24.68	2.79	25.79	1.94	.47	25.40	2.98	25.01	2.85	.15	.63

Note. M.Mean; SD.Standard Deviation; *d*.Cohen's *d* for mean difference; r.correlation between mother and father ratings. N (Girls).285; N (Boys).318; N (Mothers).588; N (Fathers).478. Means and standard deviations for girls and boys based on the average of mothers' and fathers' ratings.

Table 2

Final Model for Effortful Control Across Girls and Boys

	Factor Loadings		Intercepts		Residual Variance	
Scale	**EST**	**SE**	**EST**	**SE**	**EST**	**SE**
Attentional Focusing						
Girl	.50	.04	4.57	.04	.21	.03
Boy	.50	.04	4.57	.04	.30	.03
Inhibitory Control						
Girl	.56	.04	4.56	.04	.22	.03
Boy	.56	.04	4.56	.04	.25	.03
Low Intensity Pleasure						
Girl	.33	.03	5.58	.03	.17	.02
Boy	.33	.03	5.45	.04	.21	.02
Perceptual Sensitivity						
Girl	.31	.03	4.92	.03	.34	.03
Boy	.31	.03	4.92	.03	.48	.04

Note. EST.Parameter estimate; SE.Standard Error. Un-standardized solution presented. Girl factor variance set to 1, factor mean set to 0. Boy factor variance and mean free to vary. *df*.9. *N*.604 (286 girls, 318 boys). Model fit: χ^2.41.06, p <.001; RMSEA =.11; SRMR =.07; CFI =.93; TLI =.91.

Table 3

Final Model for Negative Affectivity Across Girls and Boys

	Factor Loadings		Intercepts		Residual Variance	
Scale	**EST**	**SE**	**EST**	**SE**	**EST**	**SE**
Anger/Frustration						
Girl	.54	.04	4.55	.05	.29	.03
Boy	.54	.04	4.65	.04	.27	.03
Discomfort						
Girl	.45	.04	4.27	.04	.30	.03
Boy	.45	.04	3.70	.04	.35	.03
Sadness						
Girl	.49	.03	4.01	.04	.13	.02
Boy	.49	.03	3.89	.03	.14	.02
Fear						
Girl	.44	.04	3.93	.05	.42	.04
Boy	.44	.04	3.70	.05	.54	.05
Soothability						
Girl	.36	.03	3.26	.04	.27	.03
Boy	.36	.03	3.27	.03	.24	.02

Note. EST.Parameter estimate; SE.Standard Error. Un-standardized solution presented. Girl factor variance set to 1, factor mean set to 0. Boy factor variance free to vary, factor mean set to 0. *df*.14. *N*.604 (286 girls, 318 boys). Model fit: χ^2.100.75, p <.001; RMSEA =.14; SRMR =.09; CFI =.90; TLI =.85.

Table 4

Final Model for Surgency Across Girls and Boys

	Factor Loadings		Intercepts		Residual Variance	
Scale	**EST**	**SE**	**EST**	**SE**	**EST**	**SE**
Smiling and Laughter						
Girls	.30	.03	5.63	.03	.29	.02
Boy	.30	.03	5.63	.03	.29	.02
High Intensity Pleasure						
Girls	.50	.04	4.89	.04	.30	.03
Boy	.50	.04	5.03	.05	.32	.03
Impulsivity						
Girls	.62	.03	4.56	.04	.11	.02
Boy	.62	.03	4.56	.04	.11	.02
Shyness						
Girls	.81	.05	4.37	.06	.61	.05
Boy	.81	.05	4.37	.06	.61	.05
Positive Anticipation						
Girls	.22	.03	5.24	.03	.24	.01
Boy	.22	.03	5.24	.03	.24	.01

Note. EST.Parameter estimate; SE.Standard Error. Un-standardized solution presented. Girl factor variance set to 1, factor mean set to 0. Boy factor variance and mean free to vary. *df*.21. *N*.604 (286 girls, 318 boys). Model fit: χ^2.150.44, p <.001; RMSEA =.14; SRMR =.09; CFI =.85; TLI =.86.

Table 5

Final Model for Effortful Control Across Parents

	Factor Loadings		Intercepts		Residual Variance	
Scale	**EST**	**SE**	**EST**	**SE**	**EST**	**SE**
Attentional Focusing						
Mother	.56	.03	4.50	.03	.34	.02
Father	.56	.03	4.50	.03	.34	.02
Inhibitory Control						
Mother	.58	.03	4.47	.03	.32	.03
Father	.58	.03	4.47	.03	.32	.03
Low Intensity Pleasure						
Mother	.41	.03	5.52	.03	.27	.02
Father	.41	.03	5.40	.03	.25	.02
Perceptual Sensitivity						
Mother	.32	.03	5.40	.03	.52	.03
Father	.32	.03	5.40	.03	.52	.03

Note. EST.Parameter estimate; SE.Standard Error. Un-standardized solution presented. Mother factor variance set to 1, factor mean set to 0. Father factor variance and mean free to vary. The mother and father factors were allowed to correlate (r=.54). Mother and father paired scale residuals were also allowed to correlate. *df*.23. *N*.605. Model fit: χ^2.104.60, p <.001; RMSEA =.08; SRMR =.06; CFI =.93; TLI =.91.

Table 6

Final Model for Negative Affectivity Across Parents

	Factor Loadings		Intercepts		Residual Variance	
Scale	**EST**	**SE**	**EST**	**SE**	**EST**	**SE**
Anger/Frustration						
Mother	.58	.03	4.59	.03	.37	.02
Father	.58	.03	4.59	.03	.37	.02
Discomfort						
Mother	.52	.03	4.11	.03	.44	.03
Father	.52	.03	4.11	.03	.44	.03
Sadness						
Mother	.55	.03	3.94	.03	.21	.02
Father	.55	.03	3.94	.03	.21	.02
Fear						
Mother	.49	.04	3.80	.04	.60	.03
Father	.49	.04	3.80	.04	.60	.03
Soothability						
Mother	.41	.03	3.21	.03	.34	.02
Father	.41	.03	3.34	.03	.35	.03

Note. EST.Parameter estimate; SE.Standard Error. Un-standardized solution presented. Mother factor variance set to 1, factor mean set to 0. Father factor variance and mean free to vary. The mother and father factors were allowed to correlate (r=.52). Mother and father paired scale residuals were also allowed to correlate. *df*.40. *N*.605. Model fit: χ^2.175.85, p<.001; RMSEA =.08; SRMR =.06; CFI =.92; TLI =.92.

Table 7

Final Model for Surgency Across Parents

	Factor Loadings		Intercepts		Residual Variance	
Scale	**EST**	**SE**	**EST**	**SE**	**EST**	**SE**
Smiling and Laughter						
Mother	.38	.03	5.76	.03	.33	.02
Father	.38	.03	5.59	.03	.35	.03
High Intensity Pleasure						
Mother	.57	.03	5.07	.03	.36	.02
Father	.57	.03	5.07	.03	.36	.02
Impulsivity						
Mother	.61	.03	4.69	.03	.17	.02
Father	.61	.03	4.69	.03	.17	.02
Shyness						
Mother	.81	.05	4.56	.05	.79	.06
Father	.68	.05	4.53	.05	.74	.06
Positive Anticipation						
Mother	.28	.02	5.31	.03	.29	.02
Father	.28	.02	5.22	.03	.29	.02

Note. EST.Parameter estimate; SE.Standard Error. Un-standardized solution presented. Mother factor variance set to 1, factor mean set to 0. Father factor variance and mean free to vary. The mother and father factors were allowed to correlate (r=.67). Mother and father paired scale residuals were also allowed to correlate. *df*.35. *N*.605. Model fit: χ^2.255.53, p<.001; RMSEA =.10; SRMR =.09; CFI =.90; TLI =.88.

Table 8

Non-Invariance Effect Sizes for Girls and Boys

	% of Observed Difference (Scale)	dMACS	Expected Difference Due to Non-Invariance	% of Observed Difference (Dimension)	Overall Effect of Non-Invariance
Effortful Control			.12	14%	.09
Attentional					
Focusing	0%	0			
Inhibitory					
Control	0%	.02			
Low Intensity Pleasure	47%	.30			
Perceptual Sensitivity	0%	.11			
Negative Affectivity *			-.20	36%	.32
Anger/Frustration	-	.32			
Discomfort	-	.35			
Sadness	-	.36			
Fear	-	.19			
Soothability	-	.41			
Surgency			-.11	10%	.04
Smiling and Laughter	0%	.09			
High Intensity Pleasure	43%	.20			
Impulsivity	0%	0			
Shyness	0%	.16			
Approach/Anticipation	0%	.09			

Note. % of Observed Difference (scale).The portion of the observed mean difference in scale scores that can be attributed to intercept non-invariance; Expected difference due to non-invariance.the degree to which observed mean scores on a dimension of temperament are expected to differ between groups based on non-invariance alone (positive values indicate girls' scores will be inflated, negative values indicate boys' score will be inflated); % of Observed Difference (Dimension).The portion of the observed mean difference in dimension scores that can be attributed to non-invariance. Values presented are from examinations of measurement invariance across girls and boys based on parents' ratings averaged together.

* Because it was not possible to set any intercepts to equality, values for the first column could not be calculated. To estimate dMACS values for all scales, sadness was first used as a marker variable, followed by fear.

Table 9

Non-Invariance Effect Sizes for Mothers and Fathers

	% of Observed Difference (Scale)	dMACS	Expected Difference Due to Non-Invariance	% of Observed Difference (Dimension)	Overall Effect of Non-Invariance
Effortful Control			.22	43%	.06
Attentional					
Focusing	0%	0			
Inhibitory					
Control	0%	.04			
Low Intensity Pleasure	60%	.23			
Perceptual Sensitivity	0%	.11			
Negative Affectivity			-.01	100%	.04
Anger/Frustration	0%	0			
Discomfort	0%	.10			
Sadness	0%	.10			
Fear	0%	.06			
Soothability	118%	.17			
Surgency			.30	77%	.06
Smiling and Laughter	94%	.27			
High Intensity Pleasure	0%	.11			
Impulsivity	0%	0			
Shyness	-	.08			
Approach/Anticipation	82%	.15			

Note. % of Observed Difference (scale).The portion of the observed mean difference in scale scores that can be attributed to intercept non-invariance; Expected difference due to non-invariance.the degree to which observed mean scores on a dimension of temperament are expected to differ between groups based on non-invariance alone (positive values indicate mothers' scores will be inflated, negative values indicate fathers' score will be inflated); % of Observed Difference (Dimension).The portion of the observed mean difference in dimension scores that can be attributed to non-invariance. Values presented are from examinations of measurement invariance across mothers and fathers based on boys' and girls' ratings averaged together.