Reliability and validity of the Cardiff Infant Contentiousness Scale in a Dutch sample

Rinske J. Windig, Bsc
s0945013
August 2014

Research Master thesis
Supervisor: Dr. Evelien Platje
Second reader: Prof. Dr. Hanna Swaab
Developmental Psychopathology in Educational and Child studies
Faculty of Social and Behavioral Science

Table of contents

Abstract

The Cardiff Infant Contentiousness Scale aims to measure early signs of aggression. In the 'Een Goed begin' longitudinal study, aggressive traits of 152 infants (85 of them boys) were assessed using the CICS. In this study, reliability and validity of the instrument were examined. Internal consistency was .44 for children aged 6 months, and .54 for children around 12 months old. For children aged 12 months old in the low-risk part of our sample however, this number increased to .68. Test-retest reliability between these two time points was .20. In validity analysis, a correlation was expected between CICs ratings and infant temperament as well as physical aggression. At 6 months old the CICS correlated with several aspects of infant temperament (r = -.32, -.24, .37, p <.01) while at 12 months old there were no correlations between the CICS and infant temperament. At 12 months of age, a correlation was found with physical aggression scores (r = .34, p >.01) . Since reliability and validity were not found to be sufficient, caution must be exercised when using the CICS as a measure of infant aggression in research and practice.

Introduction

Aggression is part of normative development, and aggressive behaviors emerge at an early age (Alink et. Al., 2006). In fact, more than half of the 12-month old children show some form of aggression according to their mother, and percentages rise up to 80% for 2 to 3 year olds. Some children however display aggressive behaviors to such an extent that it becomes problematic. 6.6% of preschoolers suffer from Oppositional Defiant Disorder and 3.3% are diagnosed with conduct disorder (Egger & Angold, 2006). In both these disorders, aggressive behavior becomes such a problem that these children are hard to handle at home and in school. These children could be on a trajectory towards lifelong behavioral problems (Hay et. al., 2010). Interest has therefore grown in being able to identify aggression problems at a very young age in order to allow for early intervention. This study focuses on a measurement instrument designed to do just that.

The Cardiff Infant Contentiousness Scale (CICS) is a questionnaire aimed at identifying early manifestations of aggressive behavior in infants. The scale was developed in 2010 and consists of five questions embedded in a larger checklist of normative developmental milestones (The Cardiff Child Development Study Milestones Questionnaire, CCDSMSQ). The CICS can be used to identify infants between 6 months and 3 years old who are on a trajectory towards aggression. The scale has been translated in Dutch by H.J.A. Smaling and J. Suurland and is currently being used in research. However, very little is known about the validity and reliability of this scale, other than research conducted by the original publishers of the scale.

Hay et. al. (2010) provided a first indication of psychometric properties of the original English version of the CICS. The CICS was found to be both reliable and valid. A sample of 310 children between 5-8 months of age was assessed in a home visit. Several questionnaires were filled out including the CICS, and several assessments of learning, attention and emotion-regulation were conducted, including the Car Seat procedure. In the Car seat procedure the infant is strapped in a car seat and distress is observed.

An internal consistency of .65 was reported for the CICS. Inter-rater reliability between parents was .51. In support of validity, infants with high CICS scores were also more likely to obtain a high score on observed distress; the CICS scores had a small but significant correlation with distress scores from the Car Seat procedure ($r(261) = .15$, $p < .01$).

Of these children, 289 were assessed again 6 months later in a laboratory setting, when they were between 11 and 15 months of age. During this lab visit, the CICS was completed again and behavior of the infant was observed during a simulated birthday party. Similarly to wave one, an internal consistency of .68 was reported for the CICS of this wave. In support of validity, CICS scores were found to be correlated to children's observed use of force in peer interaction during the birthday party ($r(254) = .21$, $p < .005$). Test-retest reliability was .44 over these six months ($r(253) = .44$, $p < .001$). In further support of validity, CICS ratings at age 5-8 months were also predictive of use of force in peer interaction during the lab visit birthday paradigm at 11-15 months ($r(248) = .15$, $p < .05$).

To be able to interpret results from research using the CICS, research on validity and reliability using the translated version in a Dutch sample is needed. The goal of this study is therefore to replicate the earlier research by Hay et. al. (2010) in a Dutch sample, as well as expand knowledge about reliability and validity of the CICS.

The current study was performed in a sample of children of which half are at high-risk for aggression. Since the Dutch sample differs from the original sample in this regard, we would also like to know how these differences affect the results. Previous research has shown that parental depression can be important moderators when examining reliability (Parade & Leerkens, 2008). Depressed mothers may misinterpret their child’s behavior and therefore report less reliably. Maternal depression was therefore assessed separately as a background variable.

This leads us to our research question: To what extent is the Cardiff Infant Contentiousness Scale (CICS) a reliable and valid measure of infant aggression? We are interested in internal consistencies at ages 6 months and 12 months, as well as test-retest reliability over these six months, and the extent to which maternal depression affects these results. We expect to replicate an internal consistency of around .65 at both time points, as reported in Hay et. al (2010). We also expect to replicate the test-retest correlation of around .44. Concurrent, divergent and predictive validity will be examined. Hay et al. (2010) reported a predictive validity of .15 with observed aggression. We similarly expect the CICS to be significantly related to physical aggression and infant temperament with a correlation around or higher than .15. We also expect the CICS to be predictive of externalizing, but not internalizing problem behaviors at a later age.

Method

Procedure

Participants included in this study were recruited as part of the 'Een goed begin' study. This longitudinal study assesses children, their mothers, and their fathers, at five times in the child's life. Mothers were approached through their obstetrician. Figure 1 summarizes which measures were included at what ages.

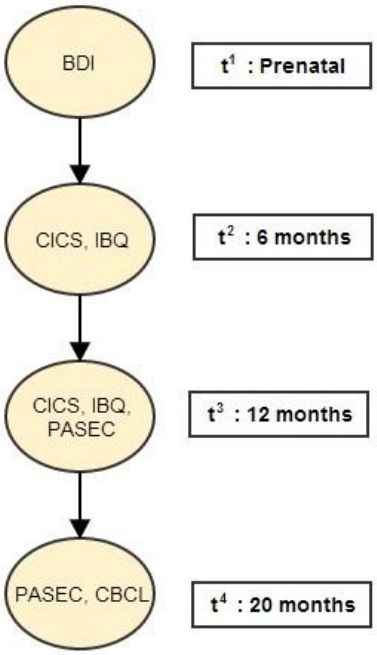

Figure 1: Measurement instruments

The first time point (t^1) assesses mother during pregnancy in a home visit. Several questionnaires including The Beck Depression Inventory (BDI) are filled out assessing background variables and an interview assessing reflective functioning of the mother is conducted.

The second time point (t^2) is a home visit when the baby is 6 months old. Physiology measures of the baby are recorded during free play and more frustrating tasks. Mother also fills out several questionnaires of interest to this study, including the CICS and the Infant Behavior Questionnaire (IBQ).

At the third time point (t^3) mother and baby visit the university lab around the child's first birthday. Again physiology measures are recorded while the child performs several tasks measuring mother child interaction, sustained attention, social referencing and joint attention. Saliva samples are also taken throughout to measure cortisol levels. Mother again completes several questionnaires, including the CICS, IBQ and the Physical Aggression Scale for Early Childhood (PASEC).

The fourth time point (t^4) is the final time point included in this study. When the child is around 20 months old, mother and child participate in another home visit. The Child Behavior Checklist, as well as other questionnaires are administered at this point. In addition to this, several tasks are administered measuring mother-child interaction, theory of mind, working memory, language development, inhibition and frustration.

Participants

At t^2, the sample consisted of 152 children, of which 85 are boys and 68 are girls. Children were between 5 and 8 months of age, with an average of 6.01 months old (*SD* = .46). At t^3, data was available for 71 children, of which 35 were boys and 36 girls. Ages ranged between 11 and 13 months of age, with an average of 11.73 months (*SD* = 3.59). Finally, at t^4, 43 children between 18 and 23 months old were included, with a mean age of 20.42 (*SD* = .71). 18 were boys and 25 were girls.

All children came from urban areas in the southwest part of the Netherlands. In this sample, 85 children were considered at high-risk for aggression. Children were considered at risk for aggression for a variety of reasons, listed in Table 1. The at-risk and control groups were matched for age and gender.

Table 1: Factors leading to inclusion into the at-risk group

One of the following:
Current psychological problems
Alcohol use during pregnancy
Smoking during pregnancy
Other substance use during pregnancy
Two of the following:
Single parent
Mother younger than 20
No income
No education above primary school
Financial problems
Small social network

Measurement instruments

Reliability Analysis

- Cardiff Infant Contentiousness Scale

The Cardiff Infant Contentiousness Scale (CICS) is the subject of this study and reliability and validity of this scale were investigated. As stated above, the CICS is a fairly new measure to assess early manifestations of childhood aggression. The measure consists of the following five items: *'Pulls hair' 'Hits out at you or other people', 'Bites you or other people', 'Has angry moods'* and '*Has temper tantrums*'. These questions are embedded in a larger, more neutral list of questions, such as *'Has taken two steps'*. Parents could answer this question with 0 = never, 1 = sometimes and 2 = often, adding up to a total score between 0 and 10. This measure was administered at both t^2 and t^3. Hay et. al. (2010) reported a reliability of .65 and concurrent validity of .15. In that study *'Pulls hair'* was not deemed a reliable question and deleted from the scale. We opted to include this question in analysis and assess the reliability of it again.

- Beck depression inventory

Depression measured by the Beck depression inventory (BDI) was included in reliability analysis as moderator. The BDI is a questionnaire used to measure depression in adults. Mothers filled out the questionnaire at t^1. The questionnaire has an internal consistency of .81 (Beck, Steer & Garbin, 1988), and is significantly related to clinical ratings of depression.

Validity Analysis

- Infant Behavior Questionnaire

The Infant Behavior Questionnaire (IBQ) is widely used as a parent report measure of infant temperament. Parents answer questions such as *'Did your child seem sad without a clear reason?'* on a 7-point Likert scale. The IBQ was administered at t^3 and $t^{3,}$ and used for both concurrent and predictive validity analyses. The revised version used here contains fourteen different sub-scales consisting of different aspects

of temperament. In this study the sub-scales Soothability, Falling reactivity (rate of recovery from distress) and Distress to limitations were used.

Reliability and validity of the IBQ has been firmly established. In the initial research of the revision (Garthland & Rothshield, 2003), internal consistency of the sub-scales ranged from.70 to .89 for children between 3 and 12 months old. For the sub-scales of interest to this paper average internal consistency across different age groups was .82 for all three subscales. Subsequent research showed that mothers and fathers were similarly reliable (Parade & Leerkens, 2008).

Inter-rater reliability was sufficient for some sub-scales, but not all, as correlations ranged from .06 to .75 (Garthland & Rothshield, 2003). Inter-rater reliability was not significant for Soothability ($r(26) = .06$, $p > .05$), but highly significant for both Falling reactivity ($r(26) = .69$, $p < .01$) and Distress to limitations ($r(26) = .57$, $p < .01$).

Support for validity of the IBQ has been found in research investigating the links between the IBQ and observed behavior. Father ratings on the fear sub-scale were related to observed fear, and Distress to limitations mother ratings were related to observed anger (Parade & Leerkens, 2008). However, these correlations were only significant for mothers and fathers low on depressive symptoms ($B = .69$, $p < .001$).

- Physical Aggression scale for early childhood

The Physical aggression scale for early childhood (PASEC) was developed by Alink et. al. (2006) and contains 11 items on physical aggression, scored on a 3-point Likert scale. It was administered at t^3 and t^4, and used to determine both concurrent and predictive validity.

In a random sample of 2253 children in three different age groups, internal consistency of mother and father ratings respectively was .65 for 12 month old children, .80 for 24-month-olds and .83 for 3 year olds (Alink et. al., 2006). Mother and father ratings were very similar. In a follow-up sample one year after the original questionnaire was completed, test-retest reliability was .65 for 12 month old children, .80 for 24-month-olds and .83 and for 3 year olds.

- Child Behavior Checklist

The Child behavior Checklist (CBCL) is a widely used parent report method of measuring a wide range of children's behavioral problems. The CBCL distinguishes between externalizing and internalizing problems. In this study it was administered at t^4. Correlations with the externalizing subscale were used for predicitive validity while the internalizing scale was used to determine divergent validity. The Dutch version for children aged 1½-5 years old was used, of which both reliability and validity have been firmly established (for a summary see Verhulst, Koot Akkerhuis & Veerman, 1990). Test-retest reliability was high (ICC = .78). In more recent research, Ang et. al. (2012) reported an internal consistency of .96 for the total scale, .89 for the internalizing scale and .91 for the externalizing sub-scale.

Strong indications of validity were also found. In longitudinal research in a large Dutch sample (n = 2033), Verhulst, Koot & Van der Ende (1994) found the CBCL to be predictive of the child encountering

any academic, behavioral or emotional disturbances in the six years following the measurement. The CBCL total problem score was also able to differentiate between children referred for psychiatric help and non-referred children, with a sensitivity of 64.9%, and a specificity of 90.5% . In total 80.1% of children were correctly predicted to be referred or not referred.

Data-Analysis

Before starting data-analysis, descriptive analyses were performed to check for normality, detect outliers and provide descriptive data. Since the sample is large, small deviations from normality do not violate the assumptions of tests used. For large deviations transformation of the data was considered. No outliers were detected.

Reliability was determined in two ways. For the test-retest analysis a sum-score was calculated by adding the responses to all five questions on the CICS, creating a variable ranging from 0-10. The CICS at t^2 and t^3 were considered as test and retest respectively, and a correlation was calculated. For determining internal consistency of the questionnaire, a reliability analysis was conducted, with Chronbachs alpha as the outcome measure.

Next, several analyses were performed to determine if reliability is influenced by maternal depression or whether children belonged to the at-risk group. Depression was examined as a possible moderator for test-retest reliability using a regression analysis. Risk group was included as covariate in an ANOVA. The reliability analysis was repeated separately for those at-risk and not at-risk, as well as for mothers below the clinical cutoff score of the BDI to establish if reliability differs between these groups.

As has become apparent in the research questions, many types of validity were considered in this paper. Concurrent validity was examined by considering the correlation between the CICS and IBQ, as well as the PASEC, at t^3. Predictive validity will be examined in a variety of ways. First, the CICS at t^2 will be considered as a predictor for the CICS, the PASEC, and the IBQ, at t^3. Second, the CICS at both time points will be considered as a predictor of the PASEC and the CBCL externalizing sub-scale at t^4. Finally, divergent validity will be examined by correlating the CICS to the CBCL internalizing scale.

Results

Descriptive statistics

The CICS administered at 6 months had a mean score of 3.55 (*SD* = 1.45) out of a possible 10. The CICS administered at 12 months had a mean score of 4.61 (*SD* = 1.80). The correlations between the individual questions at both time points are displayed in Table 2. It can be seen that all questions significantly correlate to the total score at both time points. Especially at t^2 however, not all questions are inter-related. *'Angry moods'* and *'Temper tantrums'* are the only two questions at t^2 predictive of the same question at t^3.

Table 2: Correlations between CICS questions at t^2

		t^2					t^3				
		Angry moods	Hits	Bites	Temper tantrums	Total score	Pulls hair	Angry moods	Hits	Bites	Temper tantrums	Total score
t^2	Pulls hair	-.02	.05	.18*	.07	.37**	.21	.09	.06	.06	.12	.18
	Angry moods		.05	.14	.47**	.64**	-.10	.32**	-.16	-.17	.27*	.04
	Hits			.19*	.07	.49**	.15	.04	.12	-.12	.00	.05
	Bites				.10	.62**	.16	.00	.05	.19	-.13	.10
	Temper tantrums					.62**	.01	.26*	-.14	.00	.32**	.14
	Total score						.14	.29*	-.02	-.01	.23	.20
t^3	Pulls hair							.32**	.12	.36**	.23*	.66**
	Angry moods								-.06	.09	.56	.61**
	Hits									.38**	-.04	.50**
	Bites										.00	.65**
	Temper tantrums											.56**

* $p < .05$
** $p < .01$

Reliability

The CICS was administered at t^2 and t^3. At t^2 data was collected for 152 participants. At t^3, data was available for 71 out of these 152 participants. For one person, data was missing at t^2 but not t^3, making the total sample size 152 for calculating internal consistency at t^2, 72 at t^3 and 71 for the test-retest analysis.

A reliability analysis was executed to determine internal consistency of the CICS. The results are summarized in Table 3. At t^2 the CICS has an internal consistency of Cronbach's $\alpha = .44$ when including all five questions. Consistent with what was reported in Hay et. al (2010), deleting the question on hair pulling resulted in a slightly higher consistency of .45. However, since this increase was minimal, it was decided to keep all five questions in the analysis. The CICS at t^3 had an internal consistency of .54 when including all five questions. Deleting the question on hitting from analysis increased the Cronbachs alpha to .57, however, this was again deemed too slight of an increase and all five questions were kept in.

To determine whether mothers' wellbeing influenced reliability, internal consistency was assessed separately without mothers with a moderate to high depression score. Consistency did not improve for the CICS at t^2 (Cronbach's $\alpha = .40$) or t^3 (Cronbach's $\alpha = .56$). Internal consistency was also assessed separately for children at-risk and children not at risk. Consistency did improve slightly at t^2 (Cronbach's $\alpha = .48$) and at t^3 (Cronbach's $\alpha = .67$) with at-risk children excluded.

Table 3: Internal consistency of the CICS

	t^2	t^3
Overall	.44	.54
Non-depressed	.40	.56
Not at-risk	.48	.68

Test-retest reliability was calculated using the total score of t^2 and t^3. Both of the sum-scores were normally distributed. The test-retest correlation was not significant ($r(71) = .20$, $p > .05$).

In a regression analysis, influence of maternal depression on test-retest reliability was tested. The interaction was not significant ($B = -.003$, $p = .908$). In an ANOVA, influence of risk group on test-retest

reliability was tested. Again, the interaction was not significant ($F(5,71) = 10.97$, $p < .627$) In conclusion, we can state that reliability did not differ for the normal population and those high on depression, while risk group does influence internal consistency but not test-retest reliability.

Concurrent Validity

The results for the concurrent validity analysis are summarized in Table 4. At t^2 data was available for 152 children on the CICS, and 146 children on the IBQ. Correlations between CICS total score at t^2 and the three relevant IBQ subscales were all significant: Soothability ($r(146) = -.32$, $p > .01$), Falling reactivity ($r(146) = -.24$, $p > .01$) and Limitations to distress ($r(146) = .37$, $p > .01$).

At t^3 71 children had available CICS and PASEC data, 69 did so for the IBQ. None of the correlations between the CICS and the IBQ at t^3 were significant. The CICS at t^3 was however significantly positively correlated to the PASEC at t^3 ($r(71) = .187$, $p > .01$).

Table 4: Correlations between the CICS and other measures of aggression at t^2 and t^3

	IBQ Soothability	IBQ Falling reactivity	IBQ Limitations to distress	PASEC
t^2	.-32**	-.24**	.37**	X
t^3	.01	-.11*	.23	.34**

* $p < .05$
** $p < .01$

Predictive validity

As reported under test-retest reliability, the CICS at t^2 was not predictive of scores on the CICS at t^3. Similarly, CICS ratings at t^2 were not related to any of the three relevant IBQ subscales at t^3: Soothability ($r(146) = .16$, $p > .01$), Falling reactivity ($r(146) = .19$, $p > .01$) and Limitations to distress ($r(146) = .10$, $p > .01$) CICS ratings at t^2 were not predictive of PASEC ratings ($r(70) = .15$, $p > .05$). CICS ratings were also not predictive of CBCL externalizing ratings (n=39), neither from t^2 to t^4 ($r(39) = .04$, $p > .05$) or from t^3 to t^4 ($r(39) = .09$, $p > .05$).

Divergent validity

Finally, as a measure of divergent validity, correlations with the internalizing scale of the CBCL were calculated. Correlations were not significant at t^2 ($r(37) = -.13$, $p > .05$) or t^3 ($r(37) = -.06$, $p > .05$).

Discussion

The CICS was developed as a measure of infant aggression. In this study we examined reliability and validity of the Dutch version of this instrument. Our results indicate that, in our sample, the CICS is not a reliable measure for aggression in children aged 6 months old or 12 months old. Internal consistency at both time points, as well as test-retest reliability, was insufficient. Although some evidence of validity was found, results in that regard were inconsistent.

Our results are inconsistent with the reliability reported by Hay et. al. (2010). Although the CICS was found to be sufficiently reliable in a British sample, in our Dutch sample it was not. Especially at 6 months of age, the difference between Cronbach alpha's reported was quite large. The cause of the

inconsistency between the study by Hay et. al. (2010) can be somewhat explained by differences in the sample. Children were assessed at around the same ages in both studies. However, In this study around half of the children were at high risk for developing aggression. For children 12 months old the CICS was reliable in the low-risk group. It seems that in at-risk families, mothers were less able to reliably report on the behavior of their child. It is possible that the various problems these families have to deal with leave mothers preoccupied and less able to interpret their child's signals (Parade & Leerkens, 2008). Controlling for important risk variables in our analysis did not, however, change the conclusion on reliability of the CICS at age 6 months. Possibly the difference between reported reliability at 6 months represents a cultural difference. Dutch parents might view and report their child's behavior in a different way from British parents. A study by Achenbach, Verhulst, Baron, and Akkerhuis (1987) in older children shows, however, cross-national standardized research on children's emotional and behavioral problems is feasible. Research comparing British and Dutch children has however not been done.

Especially at 6 months old, correlations between the individual questions were low, as is expected with a low internal consistency. It seems that the questions are inter-related in very different ways at different ages. At age 6 months, *'Pulls hair'* did not correlate to most of the other questions, and had the lowest correlation to the total score. The same question asked when infants were 12 months old, was correlated to almost every other question and had the highest correlation with the total score. The same can to a certain extent be said about *'Hits'* and *'Bites'.* This seems to indicate that hair-pulling, hitting and biting at 6 months old are not reliable indicators of aggression, while at 12 months old they are. It is possible that these three behaviors are part of the normative display of aggressive behaviors in young infanthood and therefore are not an indication of later problem behavior when measured at 6 months old.

Test-retest reliability was also low. This is consistent with the finding that *'Angry moods'* and *'Temper tantrums'* are the only two questions at t^2 significantly correlated with the same question at t^3. This is again an indication that the CICS at t^2 an t^3 might not measure the same construct. Unfortunately the correlations between individual questions could not be compared to the study of Hay et. al. (2010), since that information was not available.

In addition to reliability, validity was also assessed in this study. Although in some cases the CICS was correlated to other, similar instruments as expected, the results were inconsistent. At 6 months old, the CICS was significantly correlated with infant temperament, while at 12 months old, the CICS was not related to infant temperament at all. However, a correlation between the CICS and physical aggression was found. This might be an indication that at a younger age the CICS is a measure of child temperament as opposed to childhood aggression, and at an older age, the CICS is more a measure of physical aggression. This is also consistent with our finding that the CICS at 6 months old was not predictive of the CICS at 12 months old. It is possible that *'Angry moods'* and *'Temper tantrums'* are questions more related to temperament. These questions do seem to inquire more toward irritability of the child as opposed to aggressive behavior. As stated above, these two questions were the two strongest items in the scale at 6 months old. The other three questions might represent physical aggression.

These results thus provoke the question: Is it at all possible to measure precursors of aggressive behaviors in infants? So far the earliest age at which aggression in children has been reliably reported is 12 months (Alink et. al., 2006). It is possible that 6 months is too young of an age for infants to exhibit aggressive behavior other than those behavioral aspects already covered by the construct of infant temperament. In our sample at least, aggressive behavior could not reliably be reported for children aged 6 months old, and CICS ratings at this age were not predictive of aggressive behavior later in life. Consistent with Alink et. al. (2006), we did find that at 12 months old, aggressive behavior was reliably reported among the low-risk part of our sample. CICS ratings at this age were also related to ratings of physical aggression. These ratings were however not predictive of externalizing problem behavior later in life.

There are several limitations that need to be taken into account when interpreting these results. First of all, the reliability of the CICS forms an important limitation for interpreting the results from our validity analysis. Our results might be a product of the unreliability, not the (lack of) validity, of the CICS. Second, only questionnaires were used in our validity analysis, while observation measures might provide more firm conclusions. Finally, the sample size at t^4 was quite small. It is possible that when more of the children in this study reach the required age for t^4, CICS ratings turn out to be more predictive of problem behavior at this age.

Hay et al (2010) state that the CICS might be used to identify children on a trajectory towards angry aggressiveness later in life. We have found no evidence to support this claim in our sample. The CICS fails to provide reliable and valid results especially the earliest age and in a high-risk sample. Based on this evidence, currently the Dutch version of the CICS can only be used as a measure of infant aggression at 12 months old in low-risk samples.

References

Achenbach, T. M., Verhulst, F. C., Baron, G. D., and Akkerhuis, G. W. (1987). Epidemiological Comparisons of American and Dutch Children: I. Behavioral/emotional Problems and Competencies reported by Parents for ages 4 to 16. *Journal of the American Academy of Child and Adolescent Psychiatry, 26(3),* 317-325.

Alink, L. R. A., Mesman, J., Van Zeijl, J., Stolk, M. N., Juffer, F., Koot, H. M., . . . Van IJzendoorn, M. H. (2006). The Early Childhood Aggression Curve: Development of Physical Aggression in 10- to 50-month-old Children. *Child Development, 77(4),* 954-966.

Ang, R. P., Rescorla, L. A., Achenbach, T. M., Phaik Ooi, Y., Fung, D. S. S., & Woo, B. (2012). Examining the criterion validity of CBCL and TRF problem scales and items in a large Singapore sample. *Child psychiatry & Human development, 43,* 70-86.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression inventory: Twenty-five years of evaluation. *Clinical Psychology review, 8,* 77-100.

Buss, A.H., & Perra, M. (1992). The Aggression Questionnaire. *Journal of Personality and Social Psychology, 63(3),* 452-459.

Garthstein, M., & Rothbart, M K. (2003). Studying infant temperament via the Revised Infant Behavior Questionnaire. *Infant Behavior & Development, 26,* 64-86.

Hay, D. F., Perra, O., Hudson, K., Waters, C. S., Mundy, L., Phillips, R., . . . the CCDS Team (2010). Identifying early signs of infant aggression: Psychometric properties of the Cardiff Infant Contentiousness Scale. *Aggressive Behavior, 36,* 351-357.

Kabachoff, R. I., Segal, D. L., Hersen, M. & Van Hasselt, V. B., (1997). Psychometric properties and diagnostic utility of the Beck anxiety inventory and the state-trait anxiety inventory with older adult psychiatric outpatients. *Journal of Anxiety disorders, 11(1),* 33-47.

Mullen, M., Snidman, N., & Kagan, J. (1993). Free-play behavior in inhibited and uninhibited children. Infant Behavior and Development, 16, 383–389.

Parade, S. H., & Leerkes, E. H. (2008). The reliability and validity of the Infant Behavior Questionnaire - Revised. *Infant Behavior & Development, 31,* 637-646.

Van Bakel, H. J. A., & Riksen-Walraven, J.M. (2004). Stress reactivity in 15-month old Infants: Links with infant temperament, cognitive competence and attachment security. *Developmental Psychobiology, 44,* 157-167.

Verhulst, F.C., Koot, J. M., Akkerhuis, G. W., & Veerman, J.W. (1990). *Praktische handleiding voor de CBCL.* Assen/Maastricht: Van Gorcum.

Verhulst, F. C., Koot, H. M., & Van der Ende, J. (1994). Differential predictive values of parents' and teachers' reports of children's problem behaviors: A longitudinal study. *Journal of Abnormal child psychiatry, 22(5),* 531-546.