



## Method Article

# Conceptual comparison of constructs as first step in data harmonization: Parental sensitivity, child temperament, and social support as illustrations



Marije L. Verhage<sup>a,b,\*</sup>, Carlo Schuengel<sup>a,b</sup>, Annaleena Holopainen<sup>a,b</sup>,  
 Marian J. Bakermans-Kranenburg<sup>a,b</sup>, Annie Bernier<sup>c</sup>,  
 Geoffrey L. Brown<sup>d</sup>, Sheri Madigan<sup>e,f</sup>, Glenn I. Roisman<sup>g</sup>,  
 Mette S. Vaever<sup>h</sup>, Maria S. Wong<sup>i</sup>, The Collaboration on Attachment  
 Transmission Synthesis (CATS)<sup>1</sup>

*Abbreviations:* IPD, Individual participant data.

\* Corresponding author.

*E-mail address:* [m.l.verhage@vu.nl](mailto:m.l.verhage@vu.nl) (M.L. Verhage).

<sup>1</sup> The Collaboration on Attachment Transmission Synthesis (CATS) consists of 72 members. Group author members are: Marian J. Bakermans-Kranenburg, Vrije Universiteit Amsterdam, The Netherlands; Lavinia Barone, University of Pavia, Italy; Kazuko Y. Behrens, SUNY Polytechnic Institute, Utica, NY; Johanna Behringer, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany; Annie Bernier, University of Montréal, Canada; Ina Bovenschen, University of Erlangen - Nuremberg, Germany; Geoffrey L. Brown, University of Georgia, GA; Rosalinda Cassibba, University of Bari Aldo Moro, Italy; Jude Cassidy, University of Maryland, College Park, MD; Gabrielle Coppola, University of Bari Aldo Moro, Italy; Alessandro Costantini, University of Bari Aldo Moro, Italy; Mary Dozier, University of Delaware, Newark, DE; Karin Ensink, Université Laval, Quebec City, Canada; R. M. Pasco Fearon, University College London, UK; Brent Finger, Montana State University Billings, MT; Airi Hautamaki, University of Helsinki, Finland; Nancy L. Hazen, University of Texas at Austin, TX; Elena Ierardi, University of Milano-Bicocca, Italy; Inês Jongenelen, Universidade Lusófona do Porto, Portugal; Simo Køppe, University of Copenhagen, Denmark; Francesca Lionetti, d'Annunzio University of Chieti-Pescara, Italy and Queen Mary University of London, UK; Sheri Madigan, University of Calgary and the Alberta Children's Hospital Research Institute, Calgary, Canada; Sarah Mangelsdorf, University of Wisconsin-Madison, WI; Mirjam Oosterman, Vrije Universiteit Amsterdam, The Netherlands; Cecilia S. Pace, University of Genoa, Italy; K. Lee Raby, University of Utah, Salt Lake City, UT; Cristina Riva Crugnola, University of Milano-Bicocca, Italy; Glenn I. Roisman, University of Minnesota, Minneapolis, MN; Carlo Schuengel, Vrije Universiteit Amsterdam, The Netherlands; Alessandra Simonelli, University of Padova, Italy; Gottfried Spangler, University of Erlangen - Nuremberg, Germany; George M. Tarabulsy, Université Laval, Quebec City, Canada; Mette S. Vaever, University of Copenhagen, Denmark; Marije L. Verhage, Vrije Universiteit Amsterdam, The Netherlands; Maria S. Wong, Endicott College, MA.

Non-author collaborators are: Bronia Arnott, Newcastle University, UK; Heidi Bailey, University of Guelph, Canada; Patrick J. Brice, Gallaudet University, Washington, DC; Karl-Heinz Brisch, Paracelsus Medical School, Salzburg, Austria; Germana Castoro, University of Bari "Aldo Moro", Italy; Elisabetta Costantino, University of Milano-Bicocca, Italy; Chantal Cyr, Université du Québec à Montréal and CIUSSS Centre-Sud-de-l'Île de Montréal, Canada; Carol George, Mills College, Oakland, CA; Gabriele Gloger-Tippelt, Ulm University Medical Center, Germany; Sonia Gojman, Research Center of the Seminario de Sociopsicoanálisis, Mexico City, Mexico; Susanne Harder, University of Copenhagen, Denmark; Carollee Howes, University of California at Los Angeles, CA; Heidi Jacobsen, Region Centre for Child and Adolescent Mental Health, Eastern and Southern Norway, Norway; Deborah Jacobvitz, University of Texas at Austin, TX; Mi Kyoung Jin, Soomkyung Women's University, Seoul, South Korea; Femmie Juffer, Leiden University, The Netherlands; Miyuki Kazui, Ibaraki University, Mito, Japan; Esther M. Leerkes, University of North Carolina at Greensboro, NC; Karlen Lyons-Ruth, Harvard Medical School, Cambridge, MA; Catherine McMahon, Macquarie

<https://doi.org/10.1016/j.mex.2022.101889>

2215-0161/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

<sup>a</sup> *Clinical Child and Family Studies, Vrije Universiteit Amsterdam, The Netherlands*

<sup>b</sup> *Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, The Netherlands*

<sup>c</sup> *Department of Psychology, University of Montréal, Canada*

<sup>d</sup> *Human Development and Family Science, University of Georgia, USA*

<sup>e</sup> *Department of Psychology, University of Calgary, Canada*

<sup>f</sup> *Alberta Children's Hospital Research Institute, Calgary, Alberta, Canada*

<sup>g</sup> *Institute of Child Development, University of Minnesota, USA*

<sup>h</sup> *Department of Psychology, University of Copenhagen, Denmark*

<sup>i</sup> *Endicott College, Beverly, MA, USA*

## A B S T R A C T

This article presents a strategy for the initial step of data harmonization in Individual Participant Data syntheses, i.e., making decisions as to which measures operationalize the constructs of interest - and which do not. This step is vital in the process of data harmonization, because a study can only be as good as its measures. If the construct validity of the measures is in question, study results are questionable as well. Our proposed strategy for data harmonization consists of three steps. First, a unitary construct is defined based on the existing literature, preferably on the theoretical framework surrounding the construct. Second, the various instruments used to measure the construct are evaluated as operationalizations of this construct, and retained or excluded based on this evaluation. Third, the scores of the included measures are recoded on the same metric. We illustrate the use of this method with three example constructs focal to the Collaboration on Attachment Transmission Synthesis (CATS) study: parental sensitivity, child temperament, and social support. This process description may aid researchers in their data pooling studies, filling a gap in the literature on the first step of data harmonization.

- Data harmonization in studies using combined datasets is of vital importance for the validity of the study results.
- We have developed and illustrated a strategy on how to define a unitary construct and evaluate whether instruments are operationalizations of this construct as the initial step in the harmonization process.
- This strategy is a transferable and reproducible method to apply to the data harmonization process.

© 2022 The Author(s). Published by Elsevier B.V.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

## A R T I C L E I N F O

*Method name:* Conceptual comparison to assure construct validity in data harmonization

*Keywords:* Data harmonization, Individual participant data meta-analysis, Construct validity, Data pooling

*Article history:* Received 20 July 2021; Accepted 20 September 2022; Available online 26 October 2022

## Specifications Table

Subject Area:	Psychology
More specific subject area:	Developmental psychology
Method name:	Conceptual comparison to assure construct validity in data harmonization
Name and reference of original method:	Not applicable
Resource availability:	Not applicable

University, Sydney, Australia; Elizabeth Meins, University of York, UK; Salvador Millán; Research Center of the Seminario de Siciopsicoanálisis, Mexico City, Mexico; Lynne Murray, University of Reading, UK; Katja Nowacki, Fachhochschule Dortmund, Germany; David R. Pederson, Western University, London, Ontario, Canada; Lynn Priddis, Edith Cowan University, Western Australia; Avi Sagi-Schwartz, University of Haifa, Israel; Sarah J. Schoppe-Sullivan, Ohio State University, Columbus, OH; Judith Solomon, Cambridge University School of Medicine, UK; Anna Maria Speranza, Sapienza University of Rome, Italy; Miriam Steele, The New School for Social Research, New York, NY; Howard Steele, The New School for Social Research, New York, NY; Doug M. Teti, The Pennsylvania State University, State College, PA; Marinus H. van IJzendoorn, Erasmus University Rotterdam, The Netherlands; W. Monique van Londen-Barentsen, Utrecht University, The Netherlands; Mary J. Ward, Weill Cornell Medical College, New York, NY.

## Method details

### Background

In projects that pool individual participant data (IPD) for secondary data analysis, harmonization of the variables that were measured across the various studies is a vital step to ensuring that researchers are not combining different constructs as if they were similar in their analyses. Although there are many methods for harmonization currently described in the literature (e.g., [31,32,40]), these methods primarily focus on the final step of harmonization: the restructuring of measurements resulting from different instruments to measure a single construct in the same format. However, a necessary condition for harmonization is that those measures must operationalize the same underlying construct. This harmonization is done on different levels, as measures could operationalize very similar constructs (e.g., different varieties of apples are all caught under the construct of 'apple'), but measures could also operationalize constructs that are similar on a higher level (e.g., apples and oranges are both fruits). The process of harmonization includes decisions on the level on which constructs are harmonized, and which measures cannot be included on this level (e.g., potatoes are food, but not fruits, which is our level of interest: therefore potatoes are excluded). It is crucial to be precise about this conceptual comparison of measurements, because the validity of study conclusions depends on the validity of the constructs that are measured [24]. We have developed a strategy for defining constructs and comparing measures in order to determine their common denominator as a first step in the harmonization process. This strategy combines a top-down approach to evaluating the theoretical constructs underlying the measures and a bottom-up approach to evaluating measures prior to the recoding of values or categories. In this paper, we outline this strategy and illustrate our harmonization of three constructs (parental sensitivity, child temperament, and social support) as part of the Collaboration on Attachment Transmission Synthesis (CATS) project [64]. The CATS-project is an IPD meta-analytic investigation focused on synthesizing the literature on the association between parental state of mind regarding attachment and the child-parent attachment relationship. The CATS database holds pooled data from a large number of studies on this association, as well as their data on demographics, psychological characteristics, and family functioning that were collected in the original studies. The aim of the project is to examine the mechanisms behind, and the boundaries of, the association between adult attachment state of mind and child-parent attachment relationship quality by leveraging moderating and mediating variables. The variables of parental sensitivity, child temperament, and social support figure centrally in the conceptual model guiding CATS as potential mediators and moderators.

### A 3-Step Strategy of IPD Data Harmonization

1. *Defining a unitary construct.* This step represents the top-down strategy in our approach. In order to harmonize each construct, we began by familiarizing ourselves with the theoretical framework surrounding the construct by conducting a review of the existing literature. Our priority was on handbooks, theoretical articles, position papers, round table discussions, and literature reviews, although in absence of these, empirical papers were used as well. We mapped the theoretical framework and evaluated whether one dominant framework or multiple frameworks existed. If there was no consensus in the literature, we assessed whether there was overlap for parts of the theory in the different frameworks by sifting through the theoretical dimensions and comparing these across frameworks. If there appeared to be overlap, we examined whether the dimensions in the different frameworks might be different names for the same underlying construct. We specified the overlapping dimensions at the lowest possible order of abstraction to avoid losing information, but used higher-order data if overlap could only be seen on a higher level (i.e., when subdimensions differed between theoretical frameworks, but higher dimensions were similar, e.g., that apples and oranges are both fruits).

It was sometimes necessary to go back and forth between this step and the second step (described below) of evaluating the instruments used to measure the construct, because the availability of the data also impacted decision-making in the first step. For instance, when all instruments providing

data for our study were developed based on the same theoretical framework, it was not necessary to assess the overlap with other theoretical frameworks.

2. *Evaluating the instruments used to measure the construct.* This step represents the bottom-up strategy in our approach. For each of the instruments that were used to measure the construct in our dataset, we determined its theoretical basis and which parts of the theory it measured. If only part of the instrument was in line with the framework, only this part was retained. We were again as detailed as possible by assessing which theoretical subdimensions were measured by which subscale or even item(s). However, this process was a balancing act between precision and availability. For example, we gave priority to including a less specific subscale if it led to a substantial increase in the available data, for instance when another often-used instrument could only be harmonized on a higher level of abstraction. This way, we tried to include as much data as possible with as high a convergent validity as possible. Sometimes, a construct is not related to any distinct theoretical framework. In such a case, we focused our conceptual comparison on the instruments to determine whether they measured the same construct. This was done by reading instrument manuals, validation articles, methods sections of empirical articles using the instruments, and review articles comparing the instruments.
3. *Recoding the scores.* Once it was decided which (sub)scales or items could be retained based on steps 1 and 2, the final step of the harmonization process was choosing a method for the recoding of scores to the same metric based on the existing literature [31,32,40]. Previous harmonization efforts often used z-scores, t-scores or category-centered scores for standardizing scores of different instruments [32], but this approach was not feasible, due to the lack of population norm scores for the instruments used to measure the constructs of interest. Standardization based on sample distributions would conflate mean level differences between study samples and individual variance (i.e., a score of 3 on caregiver sensitivity would get a positive z-score in a sample with on average low levels of caregiver sensitivity and a negative z-score in a sample with on average high levels of sensitivity). We therefore chose the simple calibration method [40], which transforms continuous measures to operate on the same scale through a calibration model. Specific calibration models are described per construct below. We compared scale score descriptions between instruments and matched them between instruments. Likert scales were recalculated to the same metric as the 'gold standard' measure or, if there was no gold standard measure, to the largest range used by any instrument measuring this construct.

#### *Illustrating the data harmonizing strategy*

Our three-step strategy for defining the unitary constructs and harmonizing the measures used to operationalize them for three central constructs in the CATS IPD database (parental sensitivity, child temperament, and social support) is illustrated below. Unique data harmonization challenges for all three constructs were encountered. As such, we also describe our mitigation approaches to overcome these challenges.

#### *Parental sensitivity*

*Step 1.* We began by searching for literature reviews and meta-analyses on parental sensitivity, with a focus on papers that solely emphasized this construct and related constructs. Several meta-analyses [21,66] and literature reviews [43] were found. These all indicated that the term 'parental sensitivity' derived from the theoretical framework of attachment theory. It was defined by Ainsworth, Bell, and Stayton [1] as "a parent's ability to (1) notice child signals, (2) interpret these signals correctly, and (3) respond to these signals promptly and appropriately". The scale to score maternal sensitivity/insensitivity was part of the broader Maternal Care scales. It was designed to assess individual differences in parental (originally maternal) behaviors as a predictor of the quality of the mother-child attachment relationship. A key aspect of the scale is, in line with the definition, appropriate responding, where appropriate meant that the distress of the child was effectively alleviated by the response of the parent. The construct of sensitivity was therefore meant as a reciprocal process between the child and the parent, with the child signaling, the parent noticing,

interpreting, and promptly responding, and then the child having his or her needs met. The framework of attachment theory is the dominant framework for this concept. There are no further subdimensions, with sensitivity itself being a subscale of the Maternal Care scales [2].

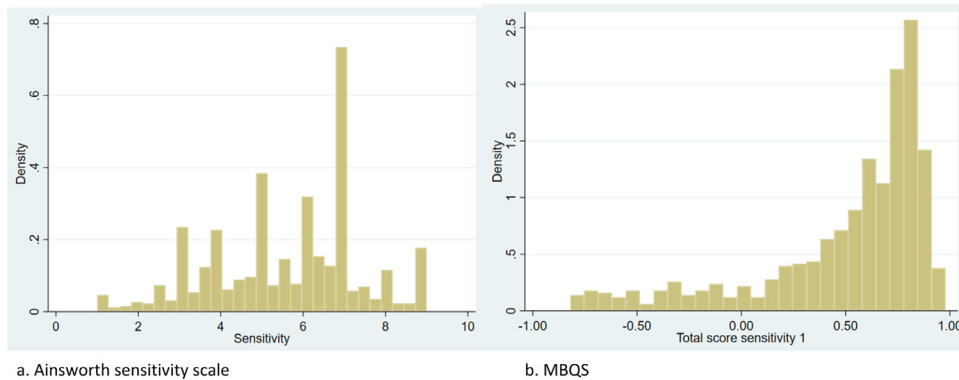
*Step 2.* Sensitivity was measured with eight different instruments across the individual studies included in the CATS IPD database: the original Ainsworth Sensitivity scale, Emotional Availability Scale (EAS 3rd and 4th edition; [9,10]), Maternal Behavior Q-Sort (MBQS; [46]), CARE-Index [17], Parent Child Observation Guide (PCOG; [8]), NICHD Study of Early Child Care and Youth Development (SECCYD) sensitivity scales [45], Infant Caregiving Scales [33], and Coding Interactive Behavior (CIB; [23]). The systematic review of focal measures of “sensitivity” by Mesman and Emmen [43] was instrumental in this step in terms of evaluating the fit between the instruments that were used and the construct of parental sensitivity. Specifically, the review by Mesman and Emmen compared commonly used observational instruments to assess parental sensitivity to the original Ainsworth sensitivity scale. Five of the alternative instruments were included in the review (EAS [9,10], MBQS [46], CARE-Index [17], NICHD SECCYD sensitivity scales [45], and CIB [23]) and these were all developed with Ainsworth’s original sensitivity construct in mind. Most of the instruments have a sensitivity subscale (as was the case for the original sensitivity scale as part of the Maternal Care scales) or a sensitivity composite score made up of several subscales. The MBQS [46] yields a single sensitivity score as the only outcome of the instrument. The review by Mesman and Emmen [43] described the NICHD SECCYD sensitivity scales [45] and the MBQS [46] as the instruments measuring sensitivity most similarly to the original construct definition. The other sensitivity scales (EAS [9,10], CARE-Index [17], and CIB [23]) were slightly broader than the original Ainsworth construct, as these included items on positive affect and warmth displayed by parents. Mesman and Emmen [43] concluded in their review that the instruments all include the main elements of the original sensitivity scale devised by Ainsworth et al. [1]. However, it has also been noted that it is better to separately assess sensitivity and positive affect/warmth, because while expressing warmth is different among cultures, sensitivity as adequately responding is universal [1]. Studies also show mixed results on the association between positive affect/warmth and parent-child attachment. Excluding the items of the EAS [9,10], Care-Index [17], and CIB [23] referring to positive affect/warmth was not possible in our dataset, because either these observational scales provided overall global scores for sensitivity based on scale descriptions or because we did not have data on the item-level. We included the data from these instruments in the CATS dataset, because this makes it possible to perform sensitivity analyses with and without these data to assess if results are similar between the strict definition of sensitivity and the operationalization of sensitivity including positive affect and warmth. Performing sensitivity analyses keeps researcher degrees of freedom constrained. Furthermore, exclusion of these data would have led to the loss of over 500 parent-child dyads, which makes the dataset less generalizable and statistically powerful.

We next evaluated the two instruments that were not included in the review by Mesman and Emmen [43]: the Infant Caregiving Scales [36] and the Parent Child Observation Guide [9]. The Infant Caregiving Scales were developed based on the descriptions of the Ainsworth Maternal Care scales and a sensitivity subscale was derived from this measure with the help of expert judges. The items in the sensitivity subscale do not hold any reference to positive affect or warmth. Given the overlap with the original sensitivity scale by Ainsworth et al. [1], we included this instrument in our harmonized sensitivity construct. The Parent Child Observation Guide was developed as a parent-child interaction assessment instrument, which was not clearly grounded in the theoretical framework of attachment and sensitivity as defined by Ainsworth et al. [1]. The instrument is used to assess both the parent’s and the child’s behavior by coding whether certain behaviors did or did not occur during the observation period. For parents, the items loaded on a sensitivity subscale or a teaching subscale that were both found to predict child cognitive and social competence [8]. Given the lack of grounding in attachment theory, the event coding of behaviors without taking into consideration the appropriateness of the responses that is vital to the Ainsworth sensitivity coding, and the lack of attachment as a reported outcome, we excluded this instrument from our sensitivity construct.

Several studies have measured parental sensitivity on multiple occasions. If this was the case, we used the data from the timepoint that was in between the measurements of our independent and dependent variable, so that parental sensitivity could be tested as a mediator variable. If

**Table 1**  
Conversion formulas for parental sensitivity data.

Likert-scale	Which instruments?	Conversion formula
1-9	Ainsworth scale [2], EAS 3rd edition [9]	none
1-4	NICHD SECCYD sensitivity scale, 15-month system [45]	$(\text{score} - 1) * 2.67 + 1$
1-5	CIB [23]; NICHD SECCYD sensitivity scale, 24-month system [45]	$(\text{score} - 1) * 2 + 1$
1-7	EAS 4th edition [10]; Infant Caregiving Scales [33]	$(\text{score} - 1) * 1.33 + 1$
0-14	CARE Index [17]	$\text{score} * 0.571 + 1$



**Fig. 1.** Distributions of parental sensitivity scores measured by Ainsworth scale (1a) and MBQS (1b).

multiple measurements were done within this timeframe, we prioritized conceptual similarity of the measurements to the Ainsworth sensitivity scale. If the same instrument was used at multiple eligible timepoints, we chose the timepoint at which the data was most complete. If the data were equally complete at multiple timepoints, we chose the timepoint that was closest to the ‘middle’ between the measurements of the independent and dependent variable.

*Step 3.* Ainsworth et al. [2] were the first to conceptualize sensitive responsiveness and operationalize this construct as an observer rating scale. As included in the Maternal Care scales, the rating scale for sensitivity ranged from 1 to 9. This was used as a reference scale for the recoding of the scores measured with other instruments. We recalculated scores on the other instruments with Likert-scales to 9-point scales using the arithmetic equivalents in Table 1.

The MBQS was measured on a continuum from -1 to 1 and therefore could not be recalculated in this way. If we could assume that the same underlying construct of sensitivity was measured both with the Ainsworth sensitivity scale and with the MBQS, as we argued in Step 2 of the harmonization strategy, then the distribution of scores should also be similar between measures, at least if the groups in which they were used were comparable. We therefore set out to compare the populations in which the Ainsworth sensitivity scale and the MBQS were used, and the resulting score distributions. Results showed that the population in which the Ainsworth sensitivity scale was used was less often considered ‘at-risk’ (38%,  $N = 943$ ) than the population in which the MBQS was used (51%,  $N = 759$ ). Populations were thus not entirely comparable, because one would expect lower sensitivity scores in at-risk populations. However, the histograms of each measure (Fig. 1a and b) showed that the distribution of the MBQS was highly negatively skewed, whereas the scores on the Ainsworth sensitivity scale were relatively normally distributed. For all other instruments, distributions were (near) normal as well, making it plausible that ‘sensitivity’ as a construct is normally distributed, while the Q-sort method used in the MBQS left more room for the low end of the scale, thus leading to a large tail in the distribution. We therefore ‘forced’ the distribution of the MBQS scores to look similar to the distribution of the Ainsworth sensitivity scores by transforming the scores.

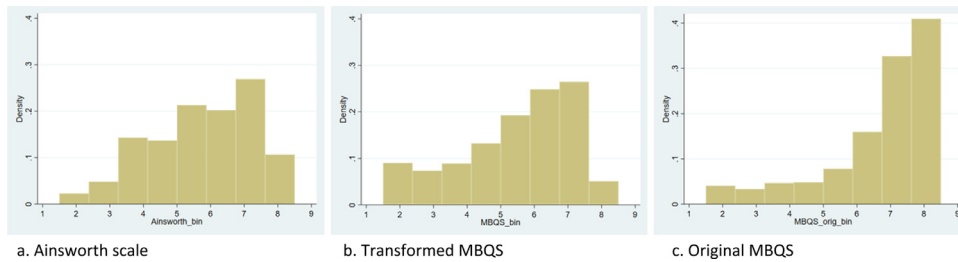


Fig. 2. Distributions of binned parental sensitivity scores.

Several transformation methods were employed and negative reciprocal transformation of the scores ( $-1/\text{score}$ ) proved most successful. Given that the scores of the Ainsworth sensitivity scale were on a scale from 1–9, we then proceeded with ‘binning’ the transformed MBQS scores (ranging from 0.4 to 1) in 8 equally wide bins ranging from 1.5 to 8.5. We did the same with the original MBQS scores and the Ainsworth scores, to facilitate comparison. Fig. 2 shows the distributions of the binned scores. The transformed MBQS scores were a better fit to the Ainsworth distribution of sensitivity scores than the untransformed scores. The distributions showed that the mean (5.85 and 5.53, respectively), standard deviation (1.78 and 1.96, respectively) and the quartiles (4.5, 6.5, 7.5 and 4.5, 5.5, 7.5, respectively) of the Ainsworth sensitivity scores and the transformed MBQS scores were roughly the same, whereas this was not the case with the untransformed MBQS scores ( $M = 6.93$ ,  $SD = 1.87$ , quartiles are 6.5, 7.5, 8.5). This similarity justified harmonizing the MBQS using the bins derived from the reciprocal transformation.

As a final check, we performed sensitivity analyses to compare scores retrieved from strict and broader operationalizations (i.e., including positive affect and warmth) of the parental sensitivity construct. Specifically, scores derived from the Ainsworth scales, NICHD SECCYD sensitivity scale, MBQS, and Infant Caregiving Scales were compared to scores derived from the EAS, Care-Index and CIB. Given that we had no overlap in scores based on different instruments in each study, we could not use correlations between scores to assess their similarity. Furthermore, given that study samples were very diverse (i.e., ranging from normative samples to samples including parents with severe mental health problems, and sampled from 10 different countries), descriptive statistics of scores derived from the strict and broader operationalizations of parental sensitivity could not be directly compared either. We therefore examined the associations between two operationalizations of parental sensitivity and the main outcome measure in the CATS study: child-parent attachment security (operationalized as secure/insecure). To compare odds ratios across categories of strict/broader operationalizations, we tested for interactions between parental sensitivity scores and strict/broad operationalization in a multilevel logistic regression model including attachment security as a dependent variable and study as grouping variable. The results showed no significant interaction effect ( $OR = 0.99$ ,  $p = .926$ ), meaning that there was no empirical reason to differentiate between the strict definition of sensitivity ( $OR = 1.24$ , 95% CI 1.14–1.35) from the operationalization of sensitivity including positive affect and warmth ( $OR = 1.24$ , 95% CI 1.08–1.44). These results justified the harmonization of both operationalizations of parental sensitivity as representing the same construct.

### Child temperament

*Step 1.* Our search started with the Handbook of Temperament [67], a key source of information about temperament. The chapter on models of child temperament [42] distinguishes five theoretical models of child temperament:

- (1) the behavioral styles approach of Thomas and Chess, that focused on ‘the stylistic component of behavior, that is, the *how* of behavior’ ([27], p. 508). They identified nine categories of behavior (activity level, regularity, adaptability, approach-withdrawal, threshold of responsiveness, intensity of reaction, quality of mood, distractibility, attention span) and three types of children based on their temperamental characteristics (easy, difficult, and slow-to-warm-up).

- (2) the criterial approach of Buss and Plomin, that defines temperament as “a set of inherited personality traits that appear early in life” ([27], p.508) This initially led to 4 temperamental dimensions: Emotionality, Activity, Sociability, and Impulsivity.
- (3) the psychobiological approach of Rothbart, that sees temperament as “constitutionally based individual differences in reactivity and self-regulation, in the domains of affect, activity and attention” ([53], p.100). This approach has identified three broad dimensions of temperament: negative affect, surgency, and effortful control.
- (4) the emotion regulation model of Goldsmith and Campos, which defines temperament as “differences in the experience and expression of emotion, including their regulatory aspects” ([28], p. 2). It focuses on individual differences in the primary emotions, such as joy, interest, sadness, anger, and fear.
- (5) the behavioral inhibition model of Kagan [35], that focuses solely on behavioral inhibition, not on the other dimensions of temperament. Importantly, none of the data within our dataset were measured with instruments based on this approach.

In a seminal roundtable with all originators of the theories except Kagan, the theorists agreed that activity level and emotionality were part of temperament, but consensus could not be reached for any other of the dimensions [27]. Each theoretical approach is accompanied by their own instruments to assess temperament. During the decades after this roundtable, factor analytic work was done with these instruments to establish the structure of temperament, both in infancy [26,30] and childhood (e.g., [29,42]). In 2006, Rothbart, original theorist of the psychobiological approach, and Bates, originator of the ICQ instrument based on the Thomas and Chess approach, co-authored a chapter on temperament. They drew two conclusions regarding the structure of temperament from the factor analytic studies: (1) that the “structure of temperament corresponds more to dimensions of reactivity than to a general style” (p. 104), meaning that temperament can differ across dimensions within a person, and (2) that there seem to be three (possibly four) broad factors of child temperament: Surgency/Extraversion, Negative Emotionality, Effortful control/Task persistence, and possibly Agreeableness/Adaptability.

With these four broad factors of temperament determined, we turned to our research questions to determine whether our focus should be on any temperament factors in particular. The CATS project aimed to examine the role of child temperament as a differential susceptibility factor (i.e., factor that makes some children more susceptible to their environment than others) of the association between parental sensitivity and parent-child attachment. Previous studies using temperament as differential susceptibility factor have focused mainly on the general factor of negative emotionality or more specifically on inhibition (fear) and irritability (anger), which are smaller aspects of this broad negative emotionality factor [62]. The meta-analysis by Slagt et al. [57] also showed that children higher on negative emotionality were both more vulnerable to negative parenting as well as profiting from positive parenting during infancy. We therefore focused on this broad factor with the two subcomponents (i.e., fear and irritable distress). Focusing on two distinct subcomponents and one broad factor that combines the two allows for sensitivity analyses comparing the effects of these temperament operationalizations. For the purpose of the data harmonization, we define Negative emotionality as “a general tendency to experience negative emotions” [42].

*Step 2.* Temperament was measured with 9 different instruments across the individual studies included in the CATS project: Infant Behavior Questionnaire (IBQ; [51]), IBQ-Revised (IBQ-R; [26]), IBQ Very Short Form (IBQ-VSF; [49]), Infant Characteristics Questionnaire (ICQ; [6]), Behavioral Style Questionnaire (BSQ; [41]), Early Childhood Behavior Questionnaire (ECBQ; [48]), Childhood Behavior Questionnaire (CBQ; [54]), Short Temperament Scale for Toddlers (STST; [47]), Questionari Italiani del Temperamento (QUIT; [4]), Emotionality Activity Sociability Temperament Survey for Children (EAS; [12]). These instruments were derived from three different theoretical frameworks, see Table 2 below.

The IBQ variants [28,52,54] and the ICQ [6] are most commonly used in the CATS dataset (10 and 9 times, respectively), with all other instruments only being used in one study. We therefore first set out to compare the subscales involved in the broad factor negative emotionality and the components of fear and irritable distress within the IBQ variants and the ICQ (see also Table 2). The IBQ [54] stems from the Rothbart theoretical framework and assesses 6 subdimensions of infant temperament:



**Table 2**  
Comparison of temperament instruments in the CATS dataset.

Instrument	k studies / N participants	Theoretical framework	Scale 'Negative emotionality'	Subscale 'Fear'	Subscale 'Irritable distress'
IBQ	7 / 738	Rothbart	Negative Affectivity composite	Fear / distress to novelty	Distress to limitations
IBQ-R	3 / 108	Rothbart	Average score of Fear and Distress to limitations scale	Fear	Distress to limitations
IBQ-VSF	1 / 200	Rothbart	Negative emotionality	-	-
ICQ	8 / 565	Thomas & Chess	Average score of Unadaptable and Fussy-Difficult	Unadaptable	Fussy-Difficult
BSQ	1 / 40	Thomas & Chess	Average score of Approach and Mood	Approach	Mood
ECBQ	1 / 67	Rothbart	Average score of Fear/Shyness and Frustration	Average score of Fear and Shyness	Frustration
CBQ	1 / 24	Rothbart	Average score of Fear and Anger/Frustration	Fear	Anger/Frustration
STST	1 / 104	Thomas & Chess	Average score of Approach and Reactivity/Cooperation	Approach	Average score of Reactivity and Cooperation
QUIT	1 / 40	Thomas & Chess	Average score of Novelty inhibition and Negative emotionality	Novelty inhibition	Negative emotionality
EAS	1 / 148	Buss & Plomin	Average score of Shyness and Emotionality	Shyness	Emotionality

activity level, soothability, fear (sometimes called 'distress to novelty'), distress to limitations, smiling and laughter, and duration of orienting. The subdimensions of 'Fear' and 'Distress to limitations' (i.e., irritable distress) together constitute a composite measure of "Negative affectivity" [52], which reflects Negative emotionality as defined above.

The revised version of the IBQ, the IBQ-R [26], was expanded to include 14 subscales: approach, vocal reactivity, high pleasure, smile/laughter, activity level, perceptual sensitivity, sadness, distress to limitations, fear, falling reactivity/rate of recovery from distress, low pleasure, cuddliness, duration of orienting, and soothability. The composite measure of "Negative affectivity" includes four subscales (Sadness, Distress to Limitations, Fear, and Falling Reactivity/Rate of Recovery from Distress) and was thus too broad for our definition of negative emotionality. We therefore combined the subscales 'Distress to limitations' and 'Fear' to reflect negative emotionality as we defined it. The same subscales were used separately as measures of irritable distress and fear.

The IBQ-VSF [52] has three subscales: negative emotionality, positive affectivity/surgency, and orienting/regulatory capacity. The "Negative emotionality" subscale reflects the broad dimension we aim for (although slightly broader as it also includes items included in the 'Sadness'-scale of the IBQ-R), but this short questionnaire does not provide separate scale scores for fear or irritable distress. Therefore, we used this slightly broader subscale in our analyses.

The ICQ [6], in line with the Thomas and Chess theoretical framework, assesses infant difficultness. It has 4 subscales, namely 'Fussy-difficult', 'Unadaptable', 'Dull', and 'Unpredictable'. The 'Fussy-difficult' subscale measures the general mood, intensity and frequency of fussing and crying, how easily children are upset, and soothability. This subscale reflects the 'irritable distress' component of negative emotionality. The 'Unadaptable' subscale measures reactions to novelty and disruptions and adaptation in general. This reflects the 'fear' component of negative emotionality.

Comparing the content of the IBQ and ICQ subscales, the ICQ scales seem slightly broader. The ICQ Fussy-Difficult scale focuses on more general distress and includes soothability, whereas the IBQ Distress to Limitations scale limits itself to distress in episodes where the child is confronted with restrictions. The ICQ Unadaptable scale focuses on more general adaptation, whereas the IBQ Fear scale is mostly aimed at fear in novel situations.

Given that we do not have ICQ item-level data in the CATS dataset, we went ahead with these subscales as reflecting the same underlying constructs of fear and irritable distress, but we will perform sensitivity analyses with the IBQ data and ICQ data separately to assess whether results are similar with the narrow operationalization of the IBQ and the slightly broader operationalization of the ICQ.

One study used both IBQ and ICQ to measure child temperament at the same timepoint. In this case, we used the IBQ data, because this is slightly more in line with the dimension of Negative emotionality and the subdimensions as mentioned in the theoretical framework.

Six studies in the CATS dataset used other instruments than the IBQ and the ICQ, though mostly from the same theoretical frameworks. Discussion to develop consensus about which instruments could and could not be harmonized took place between MLV and AH, after which these decisions were discussed again with CS. The ECBQ [51] is a Rothbart measure with similar structure as the IBQ, but for use with toddlers. It contains 18 subscales. The Frustration subscale is similar in content to the Distress to limitations subscale in the IBQ. The Fear subscale seems to be more limited, as it only includes inhibition/unease in situations of novelty and threat, but not in social situations. These situations seem to be captured in the Shyness subscale. Because these situations are included in the IBQ and the ICQ, we chose to average the scores on the Fear and Shyness subscales to include as our Fear component. The Negative Affectivity composite scale contains 8 subscales and is thus broader than the construct of our interest, we therefore combined the two components resembling Fear and Irritable distress to reflect our negative emotionality construct.

The CBQ [58] is also a Rothbart measure with similar structure, designed for use with children age 3 to 7 years. It has a Fear subscale and an Anger/Frustration subscale that assesses the same construct as the Distress to limitations subscale in the IBQ [54]. The Shyness subscale is also present in this measure, but, contrary to the factor analytic findings with the ECBQ, this subscale loaded on the Surgency dimensions and not on the Negative Affectivity dimension [54]. We therefore chose to use the Fear subscale without the Shyness subscale to best reflect our Fear component. The proposed Negative Affectivity composite encompassed more subscales than the two of our interest, therefore we combined the subscales 'Anger/Frustration' and 'Fear' to reflect negative emotionality as we defined it.

The STST [50] is an Australian adaptation of the Temperament Scale for Toddlers [25], which derives from the Thomas and Chess theoretical framework. It has six subscales: approach, reactivity, rhythmicity, cooperation, persistence, and distractibility. The 'Approach'-subscale measures the child's response to unfamiliar people and can be seen as similar to the Fear construct. The 'Reactivity'-subscale measures fussiness and frustration level when restricted, but also more general activity level. The 'Cooperation'-subscale measures how upset the child gets during routine activities (e.g., waiting for food, brushing hair/teeth). Although these two subscales include two items on general activity level, and are thus slightly broader than the irritable distress component of temperament, we combined them for use in our study.

The BSQ [44] is another temperament scale for use with older children. The 9 dimensions theorized by Thomas and Chess form the basis for the subscales of the questionnaire. The subscale 'Mood' measures the tone of overall affect (positive or negative) and is thus slightly broader than the IBQ subscale of distress to limitations to resemble the Irritable distress component of temperament. The subscale 'Approach' refers to the child's response to the environment, similarly to the STST described above, and thus reflects the Fear component.

The QUIT [4] is also derived from the framework of Thomas and Chess and has six subscales. The 'Negative emotionality' subscale describes the child's tendency to express negative emotions, which is similar to the component of irritable distress. The subscale 'Novelty inhibition' describes the child's tendency to respond with inhibition to novel stimuli, which reflects the component of fear.

The EAS [14] is the only instrument derived from the theoretical framework of Buss and Plomin. It contains four subscales (Emotionality, Activity, Sociability, and Impulsivity), one for each temperament dimension. The Emotionality subscale measures how easily the child gets upset or cries, which reflects the same underlying construct as the irritable distress component. The Shyness subscale measures social inhibition, which is similar to the Fear component, but limited to social situations (same as in the IBQ). However, we decided to use this slightly narrower subscale in our analyses.

**Table 3**  
Conversion formulas for temperament data.

Likert-scale	Which instruments?	Conversion formula
1-7	ICQ [6], IBQ [54], IBQ-R [28], IBQ-VSF [52], ECBQ [51], CBQ [58]	none
1-6	BSQ [44], STST [50], QUIT [4]	$(\text{score} - 1) * 1.2 + 1$
1-5	EAS [14]	$(\text{score} - 1) * 1.5 + 1$

One study used two instruments to measure temperament, the QUIT and the ITQ-R, which derive from the same theoretical framework. We ended up using the data collected with the QUIT, because the ITQ-R data was available for less than half of the sample.

As an assessment of the associations between subscales of these instruments, several studies have used multiple temperament instruments in one sample. One such study compared scales on the IBQ, ICQ, and RITQ ([30], as described in [53]). It found that distress to novelty (which we defined as ‘inhibition/fear’) was assessed by all these instruments (IBQ Fear, ICQ Unadaptable, RITQ Approach-Withdrawal scales). Intercorrelations were high: for mother average  $r=.64$ , for daycare teachers  $r=.63$ . Irritable distress was also measured by all instruments with the scales IBQ Distress to limitations, ICQ Fussy/Difficult, and RITQ Negative Mood (which includes positive mood at one pole) scales. Average intercorrelation for M was  $r=.54$ , for daycare teachers  $r=.71$ .

Another study by Lemery et al. [39] used principal components analyses with different questionnaires and showed the existence of four composite temperament dimensions, among which distress-anger and fear. For infants, IBQ Distress to limitations, ICQ Fussy, and ITQ Mood loaded on the distress-anger component. IBQ Distress to Novelty (fear), ICQ Adaptability, and ITQ Approach loaded on the fear component. For toddlers, the scales ICQ Difficult and BSQ Mood loaded on the distress-anger component, whereas ICQ Adaptability and BSQ Approach loaded on the fear component.

Mervielde & de Pauw [42] found clear correspondence between the scales EAS Emotionality and TTS Mood in early childhood ( $r = .60$ ) and between BSQ Adaptability/Mood, EAS Emotionality, and CBQ Negative Affect in middle childhood (range  $r = .48$  to  $r = .54$ ). This is in line with how we structured the ‘Irritable distress’ component.

The results of these studies underline the overlap between the measures and support our choices in the harmonization process.

*Step 3.* The IBQ and ICQ as most often used instruments measured temperament on a 7-point scale. We chose to convert the scores on the other scales to a 7-point scale, also because this was the largest scale in use (other instruments had 5- or 6-point Likert scales). Arithmetic equivalents can be found in Table 3.

Data from two studies had to be excluded, because the available scores were not in line with the rating scales of the instruments. One study used the ICQ, but had only ICQ dichotomous difficult vs easy temperament categorizations. The other study used the BSQ, but scores ranged from 0-2 instead of 1-6, and an explanation for these discrepancies could not be resolved.

Finally, we performed sensitivity analyses for different operationalizations of negative emotionality. First, we examined the correlations between the broad factor of negative emotionality, and the smaller subcomponents of fear and irritable distress. The correlations between negative emotionality and fear and negative emotionality and irritable distress were high, both  $r = .83$ . However, the correlation between fear and irritable distress was  $r = .38$ , suggesting that these two subcomponents assess overlapping as well as unique parts of the negative emotionality construct in the CATS dataset. This indicates that results may differ depending on the choice of the subcomponent. Therefore in subsequent analyses we always performed analyses with the broad factor of negative emotionality as well as with the two subcomponents fear and irritable distress separately.

As a second series of sensitivity analyses, we compared scores retrieved from the IBQ and the ICQ on the subcomponents fear and irritable distress, because the operationalization of both scales was slightly broader in the ICQ than in the IBQ (see above). To compare odds ratios across measures, we tested in two separate analyses the interaction between fear and distress and IBQ/ICQ in a multilevel logistic regression model including attachment security as a dependent variable and study as grouping

variable. For fear, the results showed no significant interaction effect ( $OR = 1.24, p = .174$ ), meaning that there was no empirical reason to differentiate between the IBQ ( $OR = 0.91, 95\% CI 0.75, 1.11$ ) and the ICQ ( $OR = 1.08, 95\% CI 0.87, 1.34$ ). For distress, results did not show a significant interaction effect either ( $OR = 0.95, p = .733$ ), with ORs not differing between IBQ ( $OR = 1.03, 95\% CI 0.84, 1.26$ ) and ICQ ( $OR = 0.96, 95\% CI 0.77, 1.19$ ). Given these results, it is justified to include fear and distress derived from IBQ and ICQ as operationalizations of the same subcomponents.

### *Social support*

*Step 1.* In searching for a theoretical framework for the concept of social support, we found several overview papers that described that social support is mostly used in research as a buffering factor to health outcomes during stressful events [22,60,65]. In most of the empirical work that examines effects of social support, it is not well-integrated with the research and theory on close relationships, but measured with simple measures, such as 'general satisfaction with support' and 'number of close contacts'. As such, the concept of social support as used in research in the field of developmental psychology is not embedded in one theoretical framework; instead, researchers draw on different theoretical frameworks for the embedding of social support. For instance, some researchers link the importance of social support to attachment theory by stating that individuals seek proximity to close others in times of distress [22], whereas other researchers link it to coping theory, social learning theory, or attribution theory [65].

Not only does social support lack a unified theoretical framework, it also lacks agreement in its definition. A concept analysis identified 30 distinct definitions of the concept [65]. Social support may be better viewed as a meta-construct that can be subdivided into distinct conceptual components, such as support network resources, supportive behavior, and subjective appraisal of support [63]. Support network resources are defined as the contacts a person regards as supporting them. Supportive behavior refers to the specific behaviors that can be offered as support. Subjective appraisal of support is the subjective evaluation that a person makes about the support they receive from their support network. Most studies focus on narrow operationalizations of social support, probably due to the complex nature of the concept. This often causes a focus on one of the three conceptual components of support. Several studies have compared the valence of actual received support and the subjective appraisal of support (also termed perceived support or satisfaction with support) and found that perceived support was a better predictor of health outcomes than actual received support (e.g., [22,36,50,59]). Perceived support also seems to be more commonly measured in studies, as reported by Feeney and Collins [22], but also evident in the CATS dataset. We therefore focused on satisfaction with support in the CATS study.

In our search for theoretical framing and defining a unitary construct of social support, we learned that two more distinctions are to be considered: sources of support and type of support. Regarding the sources of support, we set out to review the literature on whether it mattered who provided the support in the context of parenting. Distinctions can be made between partner, family, friends, acquaintances, professionals, and strangers, but comparisons are rarely made. It is suggested that support from close others is more influential than support from less familiar others [22]. A meta-analysis examined the association between support by family (spouse or mother), support by a broader network (friends and community) and maternal interactions of adolescent mothers with their infants [14]. Analyses revealed similar medium-to-large effect sizes for support by family and support by a broader network. A study by Coan et al. [15] showed that adults who were subjected to a stressor showed less brain activation in threat-related areas when holding their partner's hand, and to a lesser extent when holding hands with a stranger, compared to when alone. These findings are in line with the premise of the Social Baseline Theory that the human brain assumes close proximity to a network of familiar others and that this proximity to others diminishes stress responses in the brain [7]. We therefore focused on support given by familiar persons, without differentiating any further.

Finally, we sought literature on types of support given in the context of parenting. The research distinguishing between types of support is scarce, but several types of support are mentioned (with slightly varying terms): emotional support, informational support, material support, and appraisal support [3,11,37]. All types of support had significant negative correlations with post-partum depression in first-time mothers [37], and emotional support, informational support, and material

support were positively associated with parenting competence and parenting satisfaction in African-American teenage mothers [11]. A meta-analysis on the association between social support types and maternal behaviors and attitudes showed similar (moderate) effect sizes for each type of support [3]. All types of support appear to be equally important. Our harmonization thus focused on satisfaction with support by familiar persons, not making distinctions between types of support provided.

*Step 2.* Social support was measured with 14 different instruments across the individual studies included in the CATS project: Who Does What questionnaire (WDW; [16]), Social Support Questionnaire (SSQ; [56]), short form of the SSQ (SSQ6; [55]), Social Support Questionnaire (SSQ; [38]), Multidimensional Scale for Perceived Social Support (MSPSS; [68]), Social Provisions Scale (SPS; [20]), Dyadic Adjustment Scale (DAS; [58]), Secure Base Scoring System (SBSS; [19]), Family and Friends questionnaire [34], support and stress questionnaire [61], two satisfaction items [5], emotional support scale [13], observational assessment of co-parenting and spousal support [44] social support interview [18].

Instruments were all used in one or two studies. Given that there was not one or two instruments that were more common than other instruments, we compared each instrument to the criterion of measuring satisfaction with support by familiar persons.

Six instruments were explicitly focused on measuring satisfaction with support. The WDW [18] asks participants questions about the task divisions in three domains of family life, namely household tasks, family decisions, and childcare responsibilities. For each domain, questions are asked about how the task division is, how the ideal task division would be and how satisfied the participants were with the task division. The satisfaction items together formed a satisfaction scale, measured on a 5-point scale from very dissatisfied to very satisfied. The original SSQ [56] contains 27 items asking who a participant can count on for support. After each item, participants rate on a 6-point scale how very satisfied to very dissatisfied they are with the total support provided. The satisfaction items are averaged to obtain a total satisfaction score. The 6-item short form of the SSQ, the SSQ6 [59], has a similar structure, but with only 6 items. The SSQ developed by Leerkes and Crockenberg [38] has only 4 items and was designed to assess satisfaction with support by partners and others on several domains related to parenting. Satisfaction was measured on a 5-point scale ranging from very dissatisfied to very satisfied. We averaged the score to obtain a total satisfaction score. The support and stress questionnaire [66] consists of 15 items that present potential sources of support. Participants were asked whether they receive support from these sources. If so, participants filled out their satisfaction with the support from this source on a 5-point scale ranging from not satisfied to very satisfied. Satisfaction scores were averaged to obtain a total satisfaction score. The two satisfaction items used in the Bailey et al. study [5] measured satisfaction with support from the baby's father and current partner and satisfaction with support from parents and others. The 7-point scale ranged from not satisfied to very satisfied, and was averaged to obtain a total satisfaction score.

Two instruments were not developed to operationalize satisfaction with support, but a rational interpretation identifies the items as belonging to the universe of satisfaction with support. The MSPSS [74] consists of 12 items measuring the perceptions of social support by family, friends, and significant others. The items do not ask directly for the satisfaction with the provided support, but they aim to assess the perceived quality of support, e.g. "I get the emotional help and support I need from my family". We therefore included the measure in our harmonization. The 7-point scale ranged from very strongly disagree to very strongly agree, and was averaged to obtain a total satisfaction score. The emotional support scale [15] was used for coding the quality and the extent of social support available to mothers based on interview questions. Interview questions were about the satisfaction with support and the consistency and quality of help given to the mothers. The measure was thus not aimed solely at measuring satisfaction with support, but it was prominently included in the interview questions, thus we included the measure in our harmonization. The emotional support scale was double-coded by trained coders on a 7-point scale, ranging from almost non-existent to excellent.

The other six instruments were excluded because these did not measure satisfaction with support. Two instruments, the SPS ([22]; 24 items on 6 subscales) and the Family and Friends questionnaire ([37]; 15 items), were excluded because the formulation of the items (e.g., "There are people I can depend on to help me if I really need it") was more indicative of whether there was a support network (i.e., quantity) than of how satisfied the participant was with the support provided. Two other

**Table 4**

Conversion formulas for social support data.

Likert-scale	Which instruments?	Conversion formula
1-7	Emotional support scale [15], two satisfaction items [5], MSPSS [74]	none
1-6	SSQ [60], SSQ6 [59]	$(\text{score} - 1) * 1.2 + 1$
1-5	WDW [18], SSQ [41], support and stress questionnaire [66]	$(\text{score} - 1) * 1.5 + 1$

instruments, the SBSS [21] and the observational assessment of co-parenting and spousal support [44] were observational measures of support rated by independent observers. As such, these measured whether support was given and not perceived satisfaction with support. Furthermore, the DAS [63] measured dyadic adjustment of the couple and was therefore more a measure of relationship quality than of more general social support. Finally, the social support interview [20] asks about sources of support within the family and outside of the family. This instrument is aimed at measuring the quantity of support and was therefore excluded from the harmonization.

*Step 3.* Before deciding what scale to recode the scores to, we discussed whether the descriptions of the scale values all covered the same continuum of dissatisfaction to satisfaction with support. These descriptions varied from very dissatisfied to very satisfied (or vice versa), from very strongly disagree to very strongly agree, from not satisfied to very satisfied, and from almost non-existent to excellent (quality of support). First, we discussed how to compare the scale of disagree/agree with the scale of dissatisfied to satisfied. The disagree/agree answers were answers to questions regarding satisfaction, thus meaning 'agree with satisfied', which can be seen as similar to answering 'satisfied' to a question on the level of satisfaction. In a similar vein, the scale ranging from almost non-existent to excellent answered questions about the perceived quality of support, which was therefore also comparable to the continuum from dissatisfied to satisfied. Furthermore, 'not satisfied' on the lower end of a 'not satisfied to very satisfied'-scale cannot compare to the statement 'not dissatisfied/not satisfied' as the neutral midpoint of a scale spanning the continuum from very dissatisfied to very satisfied. We thought it more likely that someone answering 'not satisfied' on the first scale, would answer at least 'dissatisfied' on the latter scale. We therefore decided that these various descriptions could all be considered to be on the same continuum.

We recoded scale scores to the scale with the largest score span (7 points), given that there was not one predominantly used scale. Arithmetic equivalents can be found in [Table 4](#).

### Conclusion

This article illustrated the detailed strategies involved when harmonizing datasets in IPD syntheses. This initial step of defining a unitary construct and evaluating whether measurements derived from different instruments are operationalizations of this construct is crucial to ensure the validity of the measured constructs, and thus for the validity of study conclusions [24]. The paper provides researchers with an aid for conducting data pooling of studies, while also filling a gap in the literature on data harmonization. We recommend using the strategies provided herein, along with existing articles focused on the restructuring of measurements from different instruments into the same format (i.e., the final step in the data synthesis process) to effectively and accurately derive an IPD dataset for analysis.

### Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgements

This work was supported by a grant from Stichting tot Steun Nederland to M. Oosterman and C. Schuengel, a grant from the Social Sciences and Humanities Research Council Canada

(No. 430-2015-00989) to S. Madigan, a grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD; R01 HD102035) to G. I. Roisman, and a Veni grant by the Dutch Research Council (No. 451-17-010) to M. L. Verhage. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the grant agencies.

## References

- [1] M.D.S. Ainsworth, S.M. Bell, D. Stayton, Infant-mother attachment and social development, in: M.P. Richards (Ed.), *The Integration of A Child Into A Social World*, Cambridge University Press, London, 1974, pp. 99–135.
- [2] M.D. Ainsworth, M.C. Blehar, E. Waters, S. Wall, *Patterns of Attachment: A psychological study of the Strange Situation*, Erlbaum, Hillsdale NJ, 1978.
- [3] P.A. Andresen, S.L. Telleen, The relationship between social support and maternal behaviors and attitudes: a meta-analytic review, *Am. J. Community. Psychol.* 20 (6) (1992) 753–774, doi:10.1007/BF00942236.
- [4] G. Axia, *Questionari Italian del Temperamento*, 2002 Editore Erickson, Trento.
- [5] H.N. Bailey, G.M. Tarabulsky, G. Moran, D.R. Pederson, S. Bento, New insight on intergenerational attachment from a relationship-based analysis, *Dev. Psychopathol.* 29 (2) (2017) 433–448, doi:10.1017/S0954579417000098.
- [6] J.E. Bates, C.A.B. Freeland, M.L. Lounsbury, Measurement of infant difficultness, *Child Dev.* 50 (1979) 794–803, doi:10.2307/1128946.
- [7] L. Beckes, J.A. Coan, Social baseline theory: the role of social proximity in emotion and economy of action: Social Baseline Theory, *Soc. Personal Psychol. Compass* 5 (12) (2011) 976–988, doi:10.1111/j.1751-9004.2011.00400.x.
- [8] V.J. Bernstein, E.J. Harris, C.W. Long, E. Iida, S.L. Hans, Issues in the multi-cultural assessment of parent-child interaction: An exploratory study from the starting early starting smart collaboration, *J. Appl. Dev. Psychol.* 26 (3) (2005) 241–275, doi:10.1016/j.appdev.2005.02.002.
- [9] Z. Biringen, *Emotional Availability (EA) Scales manual (4th ed.): Part 1. Infancy/Early Childhood version (child aged 0–5 years)*, 2008 Unpublished manuscript.
- [10] Z. Biringen, J.L. Robinson, R.N. Emde, Appendix B: The Emotional Availability Scales (3rd ed.; an abridged infancy/early childhood version), *Attach. Hum. Dev.* 2 (1998) 256–270, doi:10.1080/14616730050085626.
- [11] S.G. Brown, D.B. Hudson, C. Campbell-Grossman, K.A. Kupzyk, B.C. Yates, K.M. Hanna, Social support, parenting competence, and parenting satisfaction among adolescent, African American, mothers, *West. J. Nurs. Res.* 40 (4) (2018) 502–519, doi:10.1177/0193945916682724.
- [12] A. Buss, R. Plomin, *Temperament: Early Developing Personality Traits*, Lawrence Erlbaum, Hillsdale, NJ, 1984.
- [13] E.A. Carlson, D. Jacobvitz, L.A. Sroufe, A Developmental investigation of inattentiveness and hyperactivity, *Child Dev.* 66 (1995) 37–54, doi:10.1111/j.1467-8624.1995.tb00854.x.
- [14] D. Clemmens, The relationship between social support and adolescent mothers' interactions with their infants: a meta-analysis, *J. Obstet. Gynecol. Neonatal Nurs.* 30 (4) (2001) 410–420, doi:10.1111/j.1552-6909.2001.tb01560.x.
- [15] J.A. Coan, H.S. Schaefer, R.J. Davidson, Lending a hand: social regulation of the neural response to threat, *Psychol. Sci.* 17 (12) (2006) 1032–1039, doi:10.1111/j.1467-9280.2006.01832.x.
- [16] C.P. Cowan, P.A. Cowan, L. Coie, J.D. Coie, *Becoming a family: The impact of a first child's birth on the couple's relationship*, in: *The First Child And Family Formation*, 1978, pp. 296–324.
- [17] P.M. Crittenden, *CARE-index manual*, Family Relations Institute, Miami, FL, 2001.
- [18] S.B. Crockenberg, Infant irritability, mother responsiveness, and social support influences on the security of infant-mother attachment, *Child Dev.* 52 (3) (1981) 857, doi:10.2307/1129087.
- [19] Crowell, J., Pan, H., Gao, Y., Treboux, D., & Waters, E. (1998). *Scoring adults' secure base use and support: overview of the Secure Base Scoring System (SBSS)*. Retrieved May 6, 2021 from <http://www.psychology.sunysb.edu/attachment/measures/content/sbss.pdf>.
- [20] C.E. Cutrona, D.W. Russell, The provisions of social relationships and adaptation to stress, *Adv. Personal Relationsh.* 1 (1987) 37–67.
- [21] M.S. De Wolff, M.H. Van Ijzendoorn, Sensitivity and attachment: a meta-analysis on parental antecedents of infant attachment, *Child Dev.* 68 (4) (1997) 571–591, doi:10.1111/j.1467-8624.1997.tb04218.x.
- [22] B.C. Feeney, N.L. Collins, Thriving through relationships, *Curr Opin Psychol* 1 (2015) 22–28, doi:10.1016/j.copsyc.2014.11.001.
- [23] R. Feldman, *Coding Interactive Behavior (CIB) manual*, Bar-Ilan University, 1998 Unpublished manuscript.
- [24] J.K. Flake, E.I. Fried, Measurement schmeasurement: questionable measurement practices and how to avoid them, *Adv. Methods Pract. Psychol. Sci.* 3 (4) (2020) 456–465, doi:10.1177/2515245920952393.
- [25] W. Fullard, S.D. McDevitt, W.B. Carey, *Toddler Temperament Scale*, Temple University, Philadelphia, 1978 Unpublished scale.
- [26] M.A. Gartstein, M.K. Rothbart, Studying infant temperament via the Revised Infant Behavior Questionnaire, *Infant Behav. Dev.* 26 (1) (2003) 64–86, doi:10.1016/S0163-6383(02)00169-8.
- [27] H.H. Goldsmith, A.H. Buss, R. Plomin, M.K. Rothbart, A. Thomas, S. Chess, R.B. McCall, Roundtable: what is temperament? Four approaches, *Child Dev.* (1987) 505–529, doi:10.2307/1130527.
- [28] H.H. Goldsmith, K.S. Lemery, N. Aksan, K.A. Buss, Temperamental substrates of personality, in: V.J. Molfese, D.L. Molfese (Eds.), *Temperament and Personality Development Across the Life Span*, Lawrence Erlbaum Associates Publishers, 2000, pp. 1–32.
- [29] H.H. Goldsmith, L.A. Rieser-Danner, S. Briggs, Evaluating convergent and discriminant validity of temperament questionnaires for preschoolers, toddlers, and infants, *Dev. Psychol.* 27 (4) (1991) 566, doi:10.1037/0012-1649.27.4.566.
- [30] H. Goldsmith, L.A. Rieser-Danner, Variation among temperament theories and validation studies of temperament assessment, in: G.A. Kohnstamm (Ed.), *Temperament Discussed: Temperament and Development In Infancy And Childhood*, Swets & Zeitlinger Publishers, 1986, pp. 1–9.

- [31] L.E. Griffith, E. Van Den Heuvel, I. Fortier, N. Sohel, S.M. Hofer, H. Payette, C. Wolfson, S. Belleville, M. Kenny, D. Doiron, P. Raina, Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported, *J. Clin. Epidemiol.* 68 (2) (2015) 154–162, doi:[10.1016/j.jclinepi.2014.09.003](https://doi.org/10.1016/j.jclinepi.2014.09.003).
- [32] L.E. Griffith, E. Van Den Heuvel, P. Raina, I. Fortier, N. Sohel, S.M. Hofer, D. Doiron, Comparison of standardization methods for the harmonization of phenotype data: an application to cognitive measures, *Am. J. Epidemiol.* (2016) 1–9, doi:[10.1093/aje/kww098](https://doi.org/10.1093/aje/kww098).
- [33] N. Hazen, *Infant Caregiving Scales (ICS)*, University of Texas at Austin, Human Development and Family Sciences, 1997 Unpublished manuscript.
- [34] S. Henderson, P. Duncan-Jones, H. McAuley, K. Ritchie, The patient's primary group, *Br. J. Psychiatry* 132 (1) (1978) 74–86, doi:[10.1192/S0007125000283001](https://doi.org/10.1192/S0007125000283001).
- [35] J. Kagan, Biology, context, and developmental inquiry, *Annu. Rev. Psychol.* 54 (2003) 1–23, doi:[10.1146/annurev.psych.54.101601.145240](https://doi.org/10.1146/annurev.psych.54.101601.145240).
- [36] B. Lahey, P.B. Cassidy, Cognitive processes in perceived social support, *J. Pers. Soc. Psychol.* 59 (2) (1990) 337.
- [37] P. Leahy-Warren, G. McCarthy, P. Corcoran, First-time mothers: Social support, maternal parental self-efficacy and postnatal depression: first-time mothers, *J. Clin. Nurs.* 21 (3–4) (2012) 388–397, doi:[10.1111/j.1365-2702.2011.03701.x](https://doi.org/10.1111/j.1365-2702.2011.03701.x).
- [38] E.M. Leerkes, S.C. Crockenberg, The development of maternal self-efficacy and its impact on maternal behavior, *Infancy* 3 (2) (2002) 227–247, doi:[10.1207/S15327078IN0302\\_7](https://doi.org/10.1207/S15327078IN0302_7).
- [39] K.S. Lemery, H.H. Goldsmith, M.D. Klinnert, D.A. Mrazek, Developmental models of infant and childhood temperament, *Dev. Psychol.* 35 (1) (1999) 189.
- [40] Maelstrom Research (n.d.). Data processing methods. <https://www.maelstrom-research.org/page/data-processing-methods>
- [41] S.C. Mcdevitt, W.B. Carey, The measurement of temperament in 3–7 year old children, *J. Child Psychol. Psychiatry* 19 (3) (1978) 245–253 [10.1111/j.1469-7610.1978.tb00467.x](https://doi.org/10.1111/j.1469-7610.1978.tb00467.x).
- [42] I. Mervielde, S.S.W. de Pauw, Models of child temperament, in: M. Zentner, R.L. Shiner (Eds.), *Handbook of Temperament*, The Guilford Press, New York, NY, 2012, pp. 21–40.
- [43] J. Mesman, R.A. Emmen, Mary Ainsworth's legacy: A systematic review of observational instruments measuring parental sensitivity, *Attach. Hum. Dev.* 15 (5–6) (2013) 485–506, doi:[10.1080/14616734.2013.820900](https://doi.org/10.1080/14616734.2013.820900).
- [44] S.E. Murphy, M.I. Gallegos, D.B. Jacobvitz, N.L. Hazen, Coparenting dynamics: mothers' and fathers' differential support and involvement, *Personal Relationships* 24 (4) (2017) 917–932, doi:[10.1111/per.12221](https://doi.org/10.1111/per.12221).
- [45] M.T. Owen, *The NICHD Study of Early Child Care Mother-Infant Interaction Scales*, Timberlawn Psychiatric Research Foundation, Dallas, TX, 1992 Unpublished manuscript.
- [46] D.R. Pederson, G. Moran, Appendix B: Maternal behavior Q-set, *Monogr. Soc. Res. Child Dev.* 60 (2–3) (1995) 247–254.
- [47] M. Prior, A. Sanson, F. Oberklaid, The Australian temperament project, in: G.A. Kohnstamm, J.E. Bates, M.K. Rothbart (Eds.), *Temperament in Childhood*, Wiley, Chichester, UK; New York, NY, USA, 1989, pp. 537–554.
- [48] S.P. Putnam, M.A. Gartstein, M.K. Rothbart, Measurement of fine-grained aspects of toddler temperament: the Early Childhood Behavior Questionnaire, *Infant Behav. Dev.* 29 (3) (2006) 386–401, doi:[10.1016/j.infbeh.2006.01.004](https://doi.org/10.1016/j.infbeh.2006.01.004).
- [49] S.P. Putnam, A.L. Helbig, M.A. Gartstein, M.K. Rothbart, E. Leerkes, Development and assessment of short and very short forms of the infant behavior questionnaire-revised, *J. Pers. Assess.* 96 (4) (2014) 445–458, doi:[10.1080/00223891.2013.841171](https://doi.org/10.1080/00223891.2013.841171).
- [50] S. Robinson, J.A. Weiss, Examining the relationship between social support and stress for parents of individuals with autism, *Res. Autism Spectrum Disord.* 74 (2020) 1–10, doi:[10.1016/j.rasd.2020.101557](https://doi.org/10.1016/j.rasd.2020.101557).
- [51] M.K. Rothbart, Measurement of temperament in infancy, *Child Dev.* (1981) 569–578, doi:[10.2307/1129176](https://doi.org/10.2307/1129176).
- [52] M.K. Rothbart, Longitudinal observation of infant temperament, *Dev. Psychol.* 22 (3) (1986) 356–365, doi:[10.1037/0012-1649.22.3.356](https://doi.org/10.1037/0012-1649.22.3.356).
- [53] M.K. Rothbart, J.E. Bates, Temperament, in: W. Damon, R. Lerner, N. Eisenberg (Eds.), *Handbook of Child Psychology: Vol. 3. Social, Emotional, and Personality Development*, 6th ed., Wiley, New York, 2006, pp. 99–166.
- [54] M.K. Rothbart, S.A. Ahadi, K.L. Hershey, P. Fisher, Investigations of temperament at 3–7 years: the children's behavior questionnaire, *Child Dev.* 72 (2001) 1394–1408, doi:[10.1111/1467-8624.00355](https://doi.org/10.1111/1467-8624.00355).
- [55] I.G. Sarason, B.R. Sarason, E.N. Shearin, G.R. Pierce, A brief measure of social support: practical and theoretical implications, *J. Soc. Pers. Relat.* 4 (4) (1987) 497–510, doi:[10.1177/0265407587044007](https://doi.org/10.1177/0265407587044007).
- [56] I.G. Sarason, H.M. Levine, R.B. Basham, et al., Assessing social support: the social support questionnaire, *J. Pers. Soc. Psychol.* 44 (1983) 127–139, doi:[10.1037/0022-3514.44.1.127](https://doi.org/10.1037/0022-3514.44.1.127).
- [57] M. Slagt, J.S. Dubas, M. Deković, M.A.G. van Aken, Differences in sensitivity to parenting depending on child temperament: a meta-analysis, *Psychol. Bull.* 142 (10) (2016) 1068–1110, doi:[10.1037/bul0000061](https://doi.org/10.1037/bul0000061).
- [58] G.B. Spanier, Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads, *J. Marriage Family* (1976) 15–28, doi:[10.2307/350547](https://doi.org/10.2307/350547).
- [59] E. Szkody, C. McKinney, Stress-buffering effects of social support on depressive problems: perceived vs. received support and moderation by parental depression, *J. Child Fam. Stud.* 28 (8) (2019) 2209–2219, doi:[10.1007/s10826-019-01437-1](https://doi.org/10.1007/s10826-019-01437-1).
- [60] B.N. Uchino, Understanding the links between social support and physical health: a life-span perspective with emphasis on the separability of perceived and received support, *Perspect. Psychol. Sci.* 4 (3) (2009) 236–255, doi:[10.1111/j.1745-6924.2009.01122.x](https://doi.org/10.1111/j.1745-6924.2009.01122.x).
- [61] D.C. Van den Boom, *Neonatal Irritability And The Development OF Attachment: Observation And Intervention*, Leiden University, Leiden, The Netherlands, 1988 Unpublished doctoral dissertation.
- [62] Van IJzendoorn, Bakermans-Kranenburg, Integrating temperament and attachment: the differential susceptibility paradigm, in: M. Zentner, R.L. Shiner (Eds.), *Handbook of Temperament*, The Guilford Press, New York, NY, 2012, pp. 403–424.
- [63] A. Vaux, An ecological approach to understanding and facilitating social support, *J. Soc. Pers. Relat.* 7 (4) (1990) 507–518, doi:[10.1177/0265407590074007](https://doi.org/10.1177/0265407590074007).
- [64] M.L. Verhage, C. Schuengel, R. Duschinsky, M.H. van IJzendoorn, R.M.P. Fearon, S. Madigan, The Collaboration on Attachment Transmission Synthesis, The Collaboration on Attachment Transmission Synthesis (CATS): a move to the level of individual-participant-data meta-analysis, *Curr. Dir. Psychol. Sci.* 29 (2) (2020) 199–206, doi:[10.1177/0963721420904967](https://doi.org/10.1177/0963721420904967).



- [65] P. Williams, L. Barclay, V. Schmied, Defining social support in context: a necessary step in improving research, intervention, and practice, *Qual. Health Res.* 14 (7) (2004) 942–960, doi:[10.1177/1049732304266997](https://doi.org/10.1177/1049732304266997).
- [66] M.A. Zeegers, C. Colonnese, G.J.J. Stams, E. Meins, Mind matters: a meta-analysis on parental mentalization and sensitivity as predictors of infant–parent attachment, *Psychol. Bull.* 143 (12) (2017) 1245, doi:[10.1037/bul0000114](https://doi.org/10.1037/bul0000114).
- [67] M. Zentner, R.L. Shiner, *Handbook of Temperament*, The Guilford Press, New York, 2012.
- [68] G.D. Zimet, N.W. Dahlem, S.G. Zimet, G.K. Farley, The multidimensional scale of perceived social support, *J. Pers. Assess.* 52 (1) (1988) 30–41, doi:[10.1207/s15327752jpa5201\\_2](https://doi.org/10.1207/s15327752jpa5201_2).