Current Psychology
https://doi.org/10.1007/s12144-021-01871-9

# Psychometric analysis of the Children's Behavior Questionnaire (CBQ) in Chile

Carolina Caffarena Barcenilla[1] · Benjamín Lira Luttges[2] · Cristian A. Rojas-Barahona[3] · Anna Lucía Campos[4]

Accepted: 18 May 2021


## Abstract

Measuring temperament is an important, yet challenging matter. The Children Behavior Questionnaire (CBQ) is a widely used measure, yet its psychometric properties have not been established in Latin America, and few studies have analyzed its structure at the item level. We analyzed the factor structure and reliability of the CBQ-short form in 998 Chilean children ($M_{age}$ = 5.95 years) in central Chile. Confirmatory factor analysis revealed that the 15 factors proposed by the theory unsatisfactorily fit the data (CFI = .58, TLI = .56, RMSEA = .046, SRMR = .080), and reliability was lacking (range of $\alpha$ = .30–.74). We extracted the 36 items of the CBQ-vsf and it performed better (CFI = .82, TLI = .80, RMSEA = .078, SRMR = .074). Exploratory analyses suggested that the surgency factor was composed of two latent variables, and separating them generated a model with better validity and reliability (CFI = .77, TLI = .75, RMSEA = .076, SRMR = .077, range of $\alpha$ = .68–.77). We suggest the CBQ-vsf provides more validity, reliability and parsimony than the CBQ-sf. Finally, we discuss the functioning of the CBQ in Chilean culture and child-rearing patterns, and issues with the wording of questions in the Spanish translation.

**Keywords** Temperament · Children's Behavior Questionnaire (CBQ) · Psychometric analysis · Validation · Short form · Very short form

A diverse research tradition on child temperament has relied upon parent-report likert scales to ask important questions about the core of children's personality development. Measuring temperament is complex, and a number of different questionnaires have been developed to this end. Of these, the most widely used in modern research is Rothbart's Child Behavior Questionnaire (Putnam & Rothbart, 2006; Rothbart et al., 2001). In this investigation we set out to evaluate its psychometric properties in the Chilean context.

According to Rothbart (2007) understanding temperament enables us to understand where personality emerges from, since it develops on the basis of temperamental characteristics and experience. Therefore, temperament has long been a topic of research interest in the attempt to improve scientific explanations of the individual differences observed in people's behaviour. Our genetic makeup and everyday experience shape our behaviour as temperament begins to manifest itself.

From a neurobiological perspective, Rothbart and Bates (2007) define temperament as constitutionally-based, individual differences in reactivity and self-regulation. In this view, such constitutional differences indicate that temperament has a biological basis influenced by genetics, contextual or experiential factors, and maturation. Reactivity refers to the wide range of ways in which people can react to external and internal changes, such as heart rate or manifestations of fear. Likewise, these authors regard self-regulation as a set of processes for controlling effort and orientation that operate as a modulator of reactivity. In other words, this perspective suggests that temperament is a tendency or disposition to react in

✉ Cristian A. Rojas-Barahona
c.rojas@utalca.cl

Carolina Caffarena Barcenilla
ccaffare@uc.cl

Benjamín Lira Luttges
blira@upenn.edu

Anna Lucía Campos
acampos@child-development-lab.org

[1] Faculty of Education, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Macul, Santiago, Chile

[2] Psychology Department, University of Pennsylvania, 425 S. University Ave., Philadelphia, PA 19104-6018, USA

[3] Faculty of Psychology/Faculty of Education Sciences, Universidad de Talca, Av. Lircay S/N, Campus Lircay, Talca, Chile

[4] Child Development Lab IDEA, Avenida Del Pinar 110 of 1206. Chacarilla, Surco, Lima, Peru

Published online: 25 May 2021

a certain way, with its expression varying depending on environmental conditions.

Based on this definition of temperament, a number of instruments aimed at various age groups have been developed: the Infant Behavior Questionnaire (IBQ, Rothbart, 1981), aimed at infants aged 3 – 12 months, which has a revised version (Gartstein & Rothbart, 2003) and short and very short version (Putnam et al., 2014); the Early Childhood Behavior Questionnaire (ECBQ, Putnam et al., 2006), aimed at children aged 18 - 36 months, which also has short version (Putnam et al., 2010) the Child Behavior Questionnaire (CBQ, Rothbart et al., 2001) aimed at children between 3 and 7 years old, which has standard, short and very short version. The Temperament in Middle Childhood Questionnaire (TMCQ, Simonds & Rothbart, 2004), aimed at children aged 7 – 10, of which only one version exists; the Early Adolescent Temperament Questionnaire - revised (EATQ-r, Ellis & Rothbart, 2001), aimed at adolescents and children between 9 and 15 years of age, which has only one version; and the Adult Temperament Questionnaire (ATQ, Evans & Rothbart, 2007), which focuses on adult temperament and of which a standard and a short version exist. All these questionnaires have been translated from English into several languages[1].

In the present article, we focus on the Children's Behavior Questionnaire (CBQ), which Rothbart et al. (2001) developed to evaluate temperament in children between 3 and 7 years of age. The CBQ has three higher order dimensions: Negative Affectivity (NA), Surgency/Extraversion (SU), and Effortful Control (EC). Within each of these, it includes lower-order more specific subdimensions. Negative Affectivity is associated with responses such as fear, anxiety, and sadness along with behavioural traits that enable individuals to respond in certain ways to rewards or punishments. Its component subdimensions are anger/frustration, discomfort, soothability, fear, and sadness. Surgency/Extraversion refers to affective responses associated with desires, positive emotionality, and sociability. In addition, it is also linked to the search for novelty and a high activity level. Its subdimensions are activity level, high intensity pleasure, approach/positive anticipation, impulsiveness, shyness, and smiling/laughter. Finally, Effortful Control refers to a person's ability to inhibit an impulsive response and instead select a non-dominant response. Its subdimensions are low intensity pleasure, perceptual sensitivity, inhibitory control, and attentional focusing (de la Osa et al., 2014; Putnam & Rothbart, 2006).

The CBQ standard version (CQB-sv) consists of 195 items rated with a Likert scale ranging from 1 to 7 points (from "*extremely untrue*" to "*extremely true*", with "*not applicable*" being a possible choice). These 195 items are grouped into 15 first-order factors that include 12 to 15 items per factor. The short form (CBQ-sf) and the very short form (CBQ-vsf) are derived from the original questionnaire. The CBQ-sf resulted from several steps: eliminating questions with a high rate of "not applicable" answers and low psychometric quality and analysing items by content, while ensuring the internal consistency of the instrument. For the CBQ-vsf, the authors calculated scores for the 3 second-order factors by averaging the standard scale scores for each factor, and correlated these scores with individual items. Then, 2-3 highly correlated items for each subdimension were retained (for more details, see Putnam & Rothbart, 2006).

The CBQ is one of the most widely used instruments for measuring child temperament. As of this writing, the article reporting its creation has been cited in 2766 published studies, whereas the short and very short versions of the questionnaire have been cited 1047 times. The questionnaire has been administered in more than 50 countries and in more than 200 databases. To illustrate the variety of cultures where this instrument has been used, we could highlight the studies conducted by Carranza et al. (2013) in Spain, Gouze et al. (2012) in the United States, and Gagne et al. (2015) in Israel, along with the psychometric study carried out by Najarpourian et al. (2017) in Iran.

The CBQ is a questionnaire used in different cultures, however, few studies have analysed its validity and reliability, especially in Latin America. Slobodskaya et al. (2019) state that a good number of studies have shown how cultural differences can be found when measuring child temperament. Despite the described cross-cultural differences being established by the ECBQ, they show how culture has an impact on temperament traits in at least two ways. On the one hand, it highlights the differences between Eastern and Western cultures. For example, Chinese and South Korean children are more behaviorally inhibited than Australian, Canadian, and Italian children. Another cited example mentioned that the US and Spanish children have more similarities between them, than in comparison to Chinese toddlers. On the other hand, it points out differences in the individualism/collectivism perspective by showing how temperamental traits might vary. After analyzing data from 18 different countries, the authors mentioned that the east-west effect is less clear in Latin America. The Negative Affectivity dimension has the highest impact on individualism/collectivism. The authors suggested that children from collective cultures such as Chile, Korea, and China are more distressed than toddlers from individualist cultures (e.g. the US, Finland, the Netherlands, and Italy). The Cross-cultural differences need to be better understood in the Latin American population to identify how the CBQ captures temperament in children.

In addition, psychometric analyses of this instrument are not wholly conclusive. A number of studies have managed to replicate the factor structure of the test, such as the one conducted by Sleddens et al. (2011), who worked with the CBQ-

[1] These are available at Mary Rothbart's website at https://research.bowdoin.edu/rothbart-temperament-questionnaires/

sv of the questionnaire and its CBQ-vsf. They found that the items had adequate internal consistency in both versions of the questionnaire and preserved the original 3-factor model. However, it should be noted that, when examining the CBQ-sv, they did not run factor analyses at the item level but focused on the 15 subscales instead. Similarly, Najarpourian et al. (2017) report that the internal consistency of the CBQ-vsf ranges from 0.71 to 0.79. The authors also state that the construct validity of the instrument is adequate and that all items have factor loadings higher than .30.

Few studies have examined the factor validity of the questionnaire at the item level. Kotelnikova et al. (2016) analysed the CBQ-sv and, based on item-level factor analyses, proposed an alternative 4-factor structure. Although they do not label these four tentative factors, the first contains subscales that refer to anger, frustration, and low inhibitory control; the second comprises the factors of adventure and silent play; the third includes smiling and becoming aware of appearances; and the fourth refers to the ability to stay calm, feel fear, and be sociable.

Likewise, Frohn (2017) used the CBQ-sv and the CBQ-sf, to analyse the functioning of the inhibitory control subscale to compare age groups. His results indicate that the scale needs a revision because there are major differences between the age groups (3-4 and 6-7 years of age) on several items. This author indicates that certain items are more relevant for only one age group; therefore, it is necessary to review aspects of the development of children's temperament and its evolution between 3 and 7 years of age to better define the construction of the subscales and items. In addition, Frohn (2017) proposes revising the 7-point Likert scale since it appears to fail to discriminate the informants' responses.

Other authors have reported different results. For instance, Allan et al. (2013) administered the very short form of the questionnaire (36 items) to the parents and teachers of 277 preschoolers. First, they ran a confirmatory factor analysis (CFA) of the reports delivered by parents and teachers. These analyses show that the 3-factor structure does not have good fit indexes and that the reliability indexes of the scales vary among parents' and teachers' responses, with the former's reports being less reliable. Then, the authors conducted an exploratory factor analysis (EFA) of the teachers' data and another of the parents' data. In both cases, better indexes are obtained with a 5-factor model and not a 3-factor model as the original publication suggests. In consequence, this study adds a fourth factor called "Shyness" and a fifth factor labelled "Sensitivity" to the three original factors of the scale.

To date, little evidence has been published about the psychometric properties of the Spanish-language short form of the CBQ, let alone at the item level, when administered to Latin American children. De la Osa et al. (2014) report that the short form of the CBQ in Spanish is moderately reliable and valid for preschool-age children. This study employed a sample of 622 children aged 3. Their parents answered the short CBQ in Spanish or Catalan, since they were able to use either languages indistinctly. In their study, both the short form and the very short form replicate the 3 major dimensions proposed by the original model. The scores on each scale display acceptable validity and reliability. Specifically, the analyses conducted to determine the structure and internal consistency of the instruments indicate that the short form fits the 3-factor model to a satisfactory degree ($\chi 2(63) = 165.4$, CFI = .94, TLI = .90, RMSEA = .052). One important caveat is that they did not analyze the 94 items, but factored the mean scores of the 15 subscales, and thereby assumed the validity of the subscales a priori. As for the very short form of the CBQ, the CFA did not replicate the results obtained by Putnam and Rothbart (2006) because the fit indexes and factor loadings were inadequate.

## Current Investigation

Given the limited information about how the Spanish-language version of the CBQ behaves in Latin America, this study analysed the psychometric properties of the short form and very short form of the CBQ in a group of Chilean children. We hypothesize that the original 15 factor structure of the short form will be replicated in our data. Additionally, we also plan to replicate the factor model of the very short form, by extracting the relevant items.

## Methods

### Participants

The participants were recruited from 9 different schools in central Chile. Even though we employed a convenience sample, we collected data from differing socioeconomic strata. 998 children, boys ($n$ = 475) and girls ($n$ = 516) from 4 to 7 years old were invited to participate as well as their parents or guardians. Seven questionnaires did not report the gender of the child. Children's average age was 5.95 years old. Sampled children with developmental disorders previously diagnosed were excluded from the analysis. In relation to informants, mean age was 32.89 years old, SD=7.10. Where possible, both parents would jointly complete the questionnaire. When this was not possible, typically mothers responded to the questionnaire ($n$ = 307), then fathers ($n$ = 38) and finally other kinds of guardians ($n$ = 12). Regarding scholarship, 21 people completed eight grades, 144 completed secondary school, 92 achieved higher technical education, 95 people finished university studies and 5 did not state their qualifications. This information was obtained only from a subsample ($n$ = 357) and represent 35.8% of the total sample.

Approximately, 2100 questionnaires were delivered to selected schools including children from prekindergarten to second grade. Of those, 1031 participants chose to participate. 33 participants were excluded for failing to respond to 10 items or more. Figure 1 shows how the final sample was obtained.

Table 1 shows the socio-demographic characteristics of the families participating in the study. To describe our sample, we used the *Índice de Vulnerabilidad Escolar* (IVE), or school vulnerability index. This index is calculated annually by the *Junta Nacional de Auxilio Escolar y Becas* (JUNAEB) a governmental institution that economically supports schools when their students are at risk of school desertion or poverty. This index is calculated using household data (e.g., income, years of schooling, number of people living together) and indicates the percentage of vulnerability of Chilean schools ranging from 0% to 100%. Higher percentages imply higher vulnerability. In our data, schools from low SES have, on average, 87.15% of vulnerability, schools from middle SES have 49.50% of vulnerability. Schools belonging to high SES are rated under 20% of vulnerability. They are privately funded and they do not receive economic support from the government. Additionally, we have included the average IVE in each county in order to position the school within its context. Some counties have a higher social mixture, therefore, the IVE of the schools does not match the county average.

## Materials

We used the 94-item CBQ-sf (Putnam & Rothbart, 2006; Rothbart et al., 2001). It is a parental report questionnaire rated with a 7-point Likert scale (*extremely untrue – extremely true*). The theoretical structure of the test comprises three second-order superfactors and fifteen first-order factors. The following are some of the items related to the second-order superfactor labelled "Surgency/Extraversion": "Often rushes into new situations", "Enjoys activities such as being chased, spun around by the arms, etc", "Is comfortable asking other children to play". The superfactor "Negative Affectivity" includes items such as the following: "Is afraid of loud noises", "Becomes quite uncomfortable when cold and/or wet", "Gets quite frustrated when prevented from doing something s/he wants to do". As for the superfactor called "Effortful Control", items include: "Can wait before entering into new activities if s/he is asked to", "Enjoys looking at picture books", "When practicing an activity, has a hard time keeping her/his mind on it".

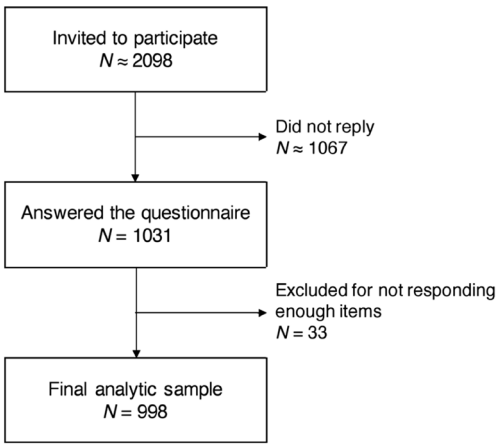


**Fig. 1** Sampling procedure

## Procedure

As in the original study (Putnam & Rothbart, 2006), the parents or caregivers of the participating children answered the questionnaire. In order to ensure their support and consent, we first contacted the principals of the participating schools and then the preschool educators and primary school teachers working there. Two strategies for administering the questionnaires were used. The first of them consisted in handing the families the questionnaire, consent form, and instructions in a sealed envelope. The second strategy consisted in inviting the parents to attend a meeting where they received the necessary instructions and answered the questionnaire in the same place. All participants signed informed consent forms before completing the instrument. When possible, parents jointly responded to the child's questionnaire. Otherwise, only one of them (primarily mothers) responded. Aside from the questionnaire, they answered additional questions about their demographic characteristics, such as family structure and the educational level of the respondent(s). The project was reviewed and approved by the ethics committee of the Pontificia Universidad Católica de Chile.

## Analysis

Data analyses were conducted using R (R Core Team, 2018). We manipulated data and generated figures with the tidyverse packages (Wickham et al., 2019). Questionnaires with more than 10 unanswered items were eliminated. The remaining missing data was multiply imputed ($m = 5$) using the mice algorithm (Van Buuren & Groothuis-Oudshoorn, 2011), which provides better estimates of missing data than traditional methods like mean substitution. Items were descriptively analysed using averages, standard deviations, and other descriptive statistics to examine item distribution and the possibility that some of them may be biased toward positive or negative answers. To analyse reliability, we used Cronbach's $\alpha$ and the average variance extracted (AVE). To analyse item quality, corrected item-test correlations were used. To

**Table 1** Socio demographics characteristics of the sample

| County | N | SES | IVE | County | Mean IVE Region |
|---|---|---|---|---|---|
| *Fifth region* | *197* | | *20.0%* | | *48.2%* |
| Villa Alemana | 197 | High | 20.0% | 74.7% | |
| *Sixth Region* | *313* | | *20.0%* | | *49.4%* |
| Rancagua | 313 | High | 20.0% | 78.8% | |
| *Santiago Region* | *488* | | *47,4%* | | *44.1%* |
| Providencia | 130 | High | 20.0% | 42.4% | |
| Cerro Navia | 39 | Low | 87.7% | 87.7% | |
| Pedro Aguirre Cerda | 49 | Low | 90.4% | 84.0% | |
| Quilicura | 37 | Low | 86.2% | 66.8% | |
| San Ramon | 30 | Low | 84.3% | 88.3% | |
| Conchalí | 9 | Middle | 66.3% | 78.0% | |
| Puente Alto | 194 | Middle | 32.8% | 76.5% | |

Elaborated from data from JUNAEB (2018) and MINEDUC (2017)

analyse reliability, we used the psych package (Revelle, 2018). We used confirmatory factor analysis (CFA) to replicate the factor structure of the questionnaire. All these analyses were performed using the lavaan package (Rosseel, 2012, 2014). We studied how items could be configured using the exploratory factor analysis (EFA) implementation of the psych package (Revelle, 2018). In the factor analyses, we sought factor loadings higher than .30; also, in the reliability analyses, we sought corrected item-test correlations higher than .30 (Field et al., 2012). Finally, in confirmatory factor analyses, we tried to ensure that fit indexes were satisfactory according to the recommendations set out by Hu and Bentler (1999).

## Results

### Short Form (94 items)

Descriptive statistics are shown in the Supplementary Online Resources (OR) in Table OR1

#### Validity

We found insufficient validity evidence for the CBQ-sf. Both the original 15-factor structure and an exploratory 7-factor structure failed to reach adequate fit. We ran a confirmatory factor analysis to test the 15-factor structure originally proposed for the short form. Using the WLSMV estimator, the most suitable for Likert data (Brown, 2015), the model did not converge; so we used Maximum Likelihood with the Satorra-Bentler correction. Even though a solution was found, it should be interpreted with caution, given that the covariance matrix of latent variables was not definitely positive. The model ($\chi 2$ (4172, $N$ = 998) = 12921.99, $p$ < .001) showed unsatisfactory fit (CFI = .58, TLI = .56), even though residual based indexes were appropriate (RMSEA = .046, SRMR = .080).

Given that the original model did not fit as expected, we ran an exploratory factor analysis of the 94 items to tap into the item level factor structure. We assessed matrix factorizability, obtaining adequate evidence from the KMO sampling statistic (KMO = .86), and a significant Bartlett's sphericity test ($\chi 2$ (4371, N = 998) = 28073.16, $p$ < .001). Based on parallel analysis, we extracted 15 factors from the polychoric correlation matrix using the unweighted least squares method and Promax rotation. The 15 factors explained 67% of the variance.

We dropped 7 factors that comprised less than three items and tried an 8-factor solution that explained 39% of the variance of the 71 items that remained. However, several items had loadings smaller than .30 and several items cross loaded on more than one factor.

Items with problematic loadings were removed one by one, until a new solution was reached. This resulted in a model that explained 39% of the variance of the 57 items that remained. The factor loadings and intercorrelations are presented in Table OR2.

This seven-factor model was confirmed in a CFA analysis using weighted least squares with mean and variance correction (Brown, 2015). Even after the elimination of item 33 due to its lack of discriminant validity, the results ($\chi 2$ (1463, N = 998) = 6108.55, p <.001) still showed unsatisfactory fit (CFI = .81, TLI = .80), even though residual based indexes were appropriate (RMSEA = .059, SRMR = .072). A pair of scales (Joy-Hapiness and Sensory Sensibility) showed a lack of discriminant validity, defined as an average variance extracted (AVE) less than the squared correlation between the latent variables (Fornell & Larcker, 1981).

### Reliability

We analysed Cronbach's α for the 15 original subscales of the questionnaire. Alphas ranged from .30 (Approach-Approximation) to .74 (Anger-Frustration). Only two of the scales had α higher than .70, while 5 were in the .60 to .70 range. Eight scales had α lower than .60. Item level statistics are available in Table OR3 of the Supplemental Online Resources and Fig. 2A.

Regarding the reliability of our seven-factor solution, ordinal Cronbach's α ranged from .65 to .87, which are all acceptable values, with all but one being greater than .70. Figure 2B shows the standard Cronbach's α and corrected item-test correlations, and detailed item-level statistics are available in Table OR4 in the Supplemental Online Resources.

## Very Short Form (36 items)

### Validity

The theoretical structure of the VSF was not replicated. Using the WLS-MV estimator, we obtained a model ($\chi 2$ (591, N = 998) = 10095.68, $p < .001$) with unsatisfactory fit (CFI = .56, TLI = .53), and almost acceptable residual indexes (RMSEA = .085, SRMR = .091). Factor loadings ranged from .04 to .57 for surgency (with two loadings so low as to be statistically insignificant), .09 to .68 for Negative Affect, and .22 to .63 for Effortful Control.

Given these unsatisfactory results, further models were tried, with problematic items being removed. Gradually, a better, albeit still not fully satisfactory model emerged after the elimination of items with low loadings (absolute value < .30) and items with discriminant validity issues (i.e. items that had theoretically ungrounded cross-loadings suggested by modification indices). This model ($\chi 2$ (225, N = 998) = 1584.40, $p < .001$) had 23 items remaining and achieved less than ideal fit (CFI = .82, TLI = .80), but residual based fit measures were more adequate (RMSEA = .078, SRMR = .074). Note that two pairs of items were allowed to covary, however, item error covariance was limited to within factor item pairs. All pairs of latent variables exhibited discriminant validity, with squared correlations between them being smaller than the average variance extracted from each latent factor. In this model, most of the eliminated items belonged to the surgency scale (which only retained 4 items).

We explored the factor structure of the original items comprising the surgency scale due to the fact that most items were removed from that factor. Most of the removed items were related to energy levels (e.g. He is always full of energy, even at night), whereas the retained items were related to sociability (e.g. He seems comfortable with anyone). We ran an exploratory factor analysis with unweighted least squared and Promax rotation. The two-factor solution explained a total of 26% of variance (evenly distributed as 13% in both factors). The details of this exploratory analysis of the surgency scale are shown in Table OR5 in the Online Supplemental Resources.

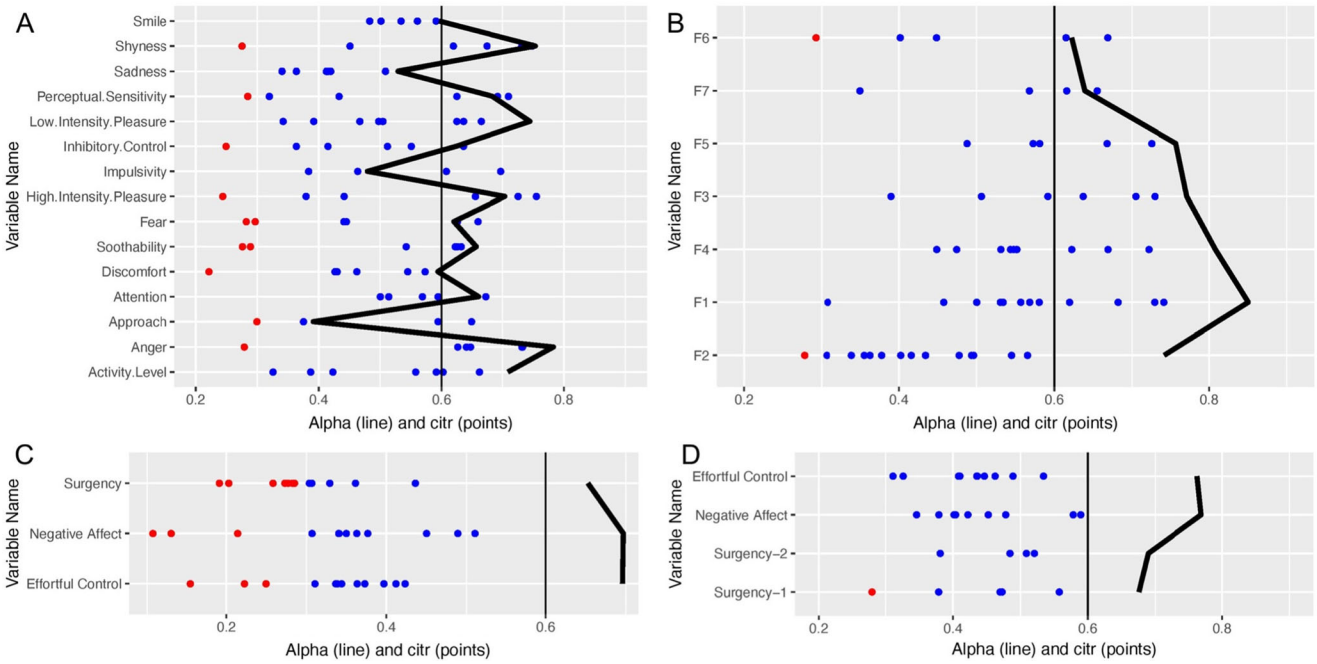


**Fig. 2** Corrected item-test correlations are plotted in blue if higher than .30 and in red if lower. The thick black line shows Cronbach's α for each subscale. The vertical line shows a cut-off at α = .60. Subfigures **A**, **B**, **C**, and **D** present statistics for original and exploratory versions of the 94-item instrument, as well as original and exploratory versions of the 36-item instrument

Therefore, we reworked our confirmatory model to include both facets present in the surgency items. We tried a first-order and second-order model, however the latter produced negative variances and was therefore discarded. The four-factor model, using the WLS-MV estimator ($\chi 2$ (342, N = 998) = 2311.92, $p < .001$), yielded unsatisfactory fit (CFI = .77, TLI = .75) and almost acceptable residual indexes (RMSEA = .076, SRMR = .077). Factor loadings ranged from .35 to .68 for Surgency 1 (S1), .42 to .71 for Surgency 2 (S2), .37 to .72 for Negative Affectivity, and from .37 to .68 for Effortful Control. Figure 3. shows the resulting model.

### Reliability

Regarding the reliability of the CBQ-vsf, Surgency showed acceptable reliability ($\alpha = .65$), with corrected item-test correlations between .24 and .44. Negative affect showed adequate levels of reliability ($\alpha = .70$), with corrected item-test correlations ranging from .14 to .61). Finally, reliability for effortful control was also adequate ($\alpha = .70$) with corrected item-test correlations ranging from .19 to .52. Figure 2C shows $\alpha$ and corrected item-test correlations for the 3 subscales. We provide item-level statistics in the Supplemental Online Materials in Table OR6.

Regarding the reliability of our model, ordinal $\alpha$ were .68 for Surgency 1, .69 for Surgency 2, .77 for Negative Affect, and .76 for Effortful Control. Figure 2D shows the standard Cronbach's $\alpha$ and corrected item-test correlations, the item-level statistics are available in the Supporting Online Resources in Table OR7.

## Discussion

This study is the first to be conducted in Chile in order to assess the evidence of validity and reliability of the Spanish-language version of the CBQ-sf and of the CBQ-vsf in a socioeconomically diverse sample of 998 children aged 3 – 7 years. To this end, we analysed reliability (using Cronbach's $\alpha$ coefficient of internal consistency) and validity information (with evidence derived from the internal structure of the questionnaire with confirmatory and exploratory analyses). For the short form, results did not support the proposed structure of 3 second-order factors and 15 first-order factors. In contrast, after the elimination of several items, our analyses yielded a 7-factor structure. The reliability of these factors was adequate compared to the unacceptable reliability of the 15 original factors. As for the very short form, our results reflect the existence of 4 factors: Negative Affect, Inhibitory Control, Surgency 1 and Surgency 2). Unlike other validation proposals (e.g., de la Osa et al., 2014) that replicate the three-factor structure, in our sample indicated the need to separate the Surgency factor into two different factors (Surgency 1 and Surgency 2) due to the contents of the items associated with each. Allan et al. (2013) also suggested a modification to the factor structure of the CBQ-vsf to include a shyness scale, parallel to our sociability scale, and independent from the general surgency scale (matching our energy level scale).

There are significant differences in the socioeconomic makeup of the samples in which the CBQ has been studied. Prior analyses have been done in richer countries, like Spain (de la Osa et al., 2014), the Netherlands (Sleddens et al., 2011), Italy, Canada and the U.S. (Kotelnikova et al., 2016; Allan et al., 2013). There are significant differences in educational attainment and socioeconomic status between Latin American countries and North America and Europe. In this sense, the present study presents a contribution in being the first effort in identifying how the CBQ might perform in lower-income countries. The only notable exception is Iran, where Najarpourian et al. (2017) investigated the CBQ. However, since the article is not available in English or Spanish, we are unable to compare our findings.

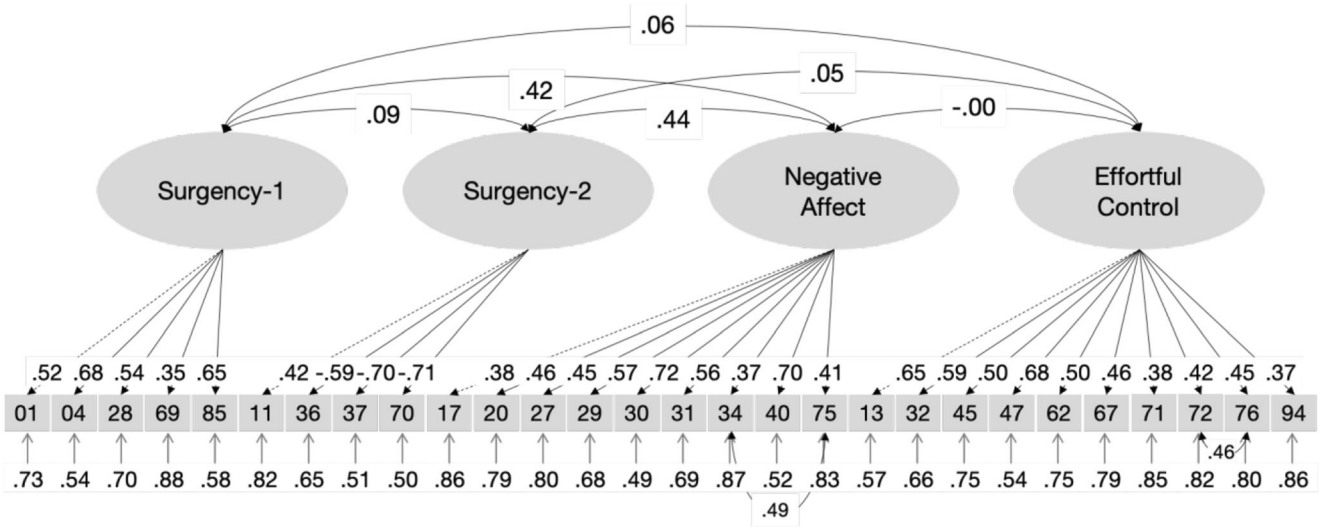


**Fig. 3** Confirmatory factor model for the CBQ-vsf

A second point relates to the fact that to date, no item level analyses of the CBQ-sf have been published. This fundamental difference in analytic approach might help explain our differing results. All the studies that have replicated the structure of three second-order factors (i.e. SU, AN, CI), have used first-order factors (e.g. Sadness, Positive Anticipation, Attentional Focusing) directly, that is, inputting the scores of the fifteen subscales into the analysis. This is a limitation of these prior studies because the psychometric properties of these subscales have not yet been analysed, and authors assume that they are valid and reliable enough. In the present study, we analysed two levels simultaneously and no convergence was found in the confirmatory factor analysis model. As a second choice, we attempted to replicate the 15-factor structure. The factor structure of the first level has been assumed as a given in prior research, and in our sample, we failed to find enough evidence for this.

Finally, the published studies often draw from different item pools, further complicating the comparisons between this study and previous research. For example, Kotelnikova et al. (2016) analysed the full instrument (195 items), which makes it difficult to compare the 4-factor solution that they propose with the results of this study. Even so, their Sensation-Seeking and Low Inhibitory Control-Disinhibition dimensions resemble the fifth factor of this analysis (Impulsivity). The dimensions Low Negative Affect and Smiling/Laughter are not clearly replicated in these results. Rather, our study revealed separate factors for negative affect (fear, sadness), while Smiling/Laughter resemble the Joy factor.

Why did we fail to replicate the theoretically expected factor structure for the SBQ-sf and CBQ-vsf? The first possibility that comes to mind are cultural differences. Latin American culture is highly likely to have upbringing patterns that differ from North American, Asian or European countries. A good approach to how cultural differences in child-rearing can interfere in a child's temperament is presented by Putnam et al. (2019). The authors mentioned that collectivist cultures (e.g, Latin-american countries) tend to be more relationally orientated–focusing more on helping others and obeying adults. On the other hand, individualistic cultures (e.g., Northern European countries) promote autonomy –focusing on independence and self-esteem. These cultural differences in terms of what behaviors are promoted and valued can translate into differences in parenting. These differences in turn, may have direct repercussions on the answers provided by families due to the value that some behaviours have, and how those behaviours are regulated according to social expectations. Therefore, these differences may, at least in part, explain why directly replicating the factor structure of the CBQ was impossible.

The fact that the psychometric qualities of exploratory versions of the test also failed to attain adequate indexes of validity and reliability somewhat weakens the cultural argument. If the results obtained were a result of robust cultural differences, the exploratory versions would be more likely to reach more appropriate validity and reliability indexes. Therefore, it is likely that it is not robust cultural differences in the *structure* of temperament what explain our results, but lower -level aspects of measurement, wording, and respondent's characteristics.

Secondly, our results may have been affected by some features of the questionnaire. The length of the test and the relevance and complexity of the items (especially those worded negatively) might have fatigued some responders. This issue could be reduced by preference of the CBQ-vsf.

In this study, several negatively worded items were eliminated (10 items) because the translation of items containing negation might be harder to interpret in a culturally diverse sample. The translation into Spanish may have introduced a language-related limitation because these items tend to be harder to understand and answer. It would be interesting to analyse how these items behave when worded directly, not inversely.

As previously noted, and based on the above, the CBQ-vsf appears to be a better choice than the SF in the sample studied. Although it was necessary to remove several items, the confirmatory analysis presented in this study represents a contribution with respect to the factor structure of the questionnaire. The Negative Affectivity and Effortful Control dimensions do not greatly differ from the original model; however, the Surgency/Extraversion dimension forms two different facets. Surgency 1 relates to children's energy level as reported by their parents; while Surgency 2 is closely related to sociability expressed by children. Putnam and Rothbart (2006) note that the Surgency/Extraversion factor comprises subscales such as Activity Level, which can be linked to the Surgency 1 factor in our model, and the Shyness subscale, which is similar to the Surgency 2 factor. Although it is possible to generate a single factor with items of this type in these authors' study, the psychometric analysis conducted with this sample suggests that they be divided into two dimensions. Undoubtedly, the proposal made in this study opens up new avenues of research for authors seeking to explore the factor structure of the CBQ.

Some limitations detected may have affected our results. First, the sample is quite heterogeneous, representing different levels of socioeconomic status (and educational status by proxy). A second limitation concerns the length of the questionnaire, since it is possible for the CBQ-SF to generate some degree of fatigue. Longer questionnaires can suffer from response biases like the acquiescence effect[2]. Finally, we focused only on the factor structure and the internal consistency of the scale. Evidence of temporal stability (i.e., test-retest

[2] In our sample, participants tended to pick the more positive response options (see table S1), and this tendency to respond using positive response options increased as the questionnaire went on ($r$ = .12).

reliability), and concurrent and discriminant validity (i.e., relations with other scales) remain the focus of future research.

In conclusion, our results suggest that the CBQ-sf should be interpreted with caution in the Chilean population. The shorter and more parsimonious structure of the CBQ-vsf appears preferable, and our results suggest the distinction between energy levels and sociability be made. Future research can further establish the cultural robustness of the separation between the energy and sociability aspects of the surgency construct.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s12144-021-01871-9.

**Acknowledgments** We are especially grateful to Dr. Helena Montenegro Maggio for her positive attitude and her assistance in the analysis and interpretation of the results obtained.

**Authors' Contributions** **Carolina Caffarena Barcenilla** participated in the conceptualization, data curation, formal analysis, project administration, and writing of the original draft and final review and editing.

**Benjamin Lira Luttges** participated in the conceptualization, data curation, formal analysis, investigation, methodology, visualization, and writing of the original draft and final review and editing.

**Cristian A. Rojas-Barahona** participated in the conceptualization, investigation, and supervision of the research paper.

**Anna Lucía Campos** participated in providing data for the study.

**Funding** This study was supported by the Chilean government's National Fund for Scientific and Technological Development-ANID (project FONDECYT number 1210989).

This study was supported by the Programa de Investigación Asociativa (PIA) en Ciencias Cognitivas, Research Center on Cognitive Sciences (CICC), Faculty of Psychology. Universidad de Talca, Campus Lircay, Chile.

**Data Availability** The datasets generated during and/or analysed during the current study are available from the corresponding author upon reasonable request.

## Declarations

**Ethical Approval** The study protocol was approved by the ethics committee of the Pontificia Universidad Católica de Chile.

**Informed Consent** Informed consent was given by all participants in order to get their permission for this study.

**Conflict of Interest** On behalf of all the authors, the corresponding author states that there is no conflict of interest.

## References

Allan, N. P., Lonigan, C. J., & Wilson, S. B. (2013). Psychometric evaluation of the Children's Behavior Questionnaire-Very Short Form in preschool children using parent and teacher report. *Early Childhood Research Quarterly, 28*(2), 302–313. https://doi.org/10.1016/j.ecresq.2012.07.009.

Brown, T. A. (2015). *Methodology in the social sciences.Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Carranza, J. A., González-Salinas, C., & Ato, E. (2013). A longitudinal study of temperament continuity through IBQ, TBAQ and CBQ. *Infant Behavior and Development, 36*(4), 749–761. https://doi.org/10.1016/j.infbeh.2013.08.002.

de la Osa, N., Granero, R., Penelo, E., Domènech, J. M., & Ezpeleta, L. (2014). The short and very short forms of the Children's Behavior Questionnaire in a community sample of preschoolers. *Assessment, 21*(4), 463–476. https://doi.org/10.1177/1073191113508809.

Ellis, L., & Rothbart, M. (2001). Revision of the Early Adolescent Temperament Questionnaire. *Poster presented at the 2001 Biennal Meeting of the Society for Research in Child Development*. https://doi.org/10.1037/t07624-000

Evans, D. E., & Rothbart, M. K. (2007). Developing a model for adult temperament. *Journal of Research in Personality, 41*(4), 868–888. https://doi.org/10.1016/j.jrp.2006.11.002.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. Sage Publications.

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.2307/3151312.

Frohn, S. (2017). *An Evaluation and Revision of the Children's Behavior Questionnaire Effortful Control Scales*. [Doctoral dissertation, University of Nebraska - Lincoln]. Retrieved March 12, 2021, from https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1299&=&context=cehsdiss&=&seiredir=1&referer=https%253A%252F%252Fscholar.google.com%252Fscholar%253Fhl%253Den%2526as_sdt%253D0%25252C10%2526q%253Dfrohn%252B2017%252Bcbq%2526btnG%253D#search=%22frohn%202017%20cbq%22

Gagne, J. R., Prater, J. C., Abramson, L., Mankuta, D., & Knafo-Noam, A. (2015). An Israeli study of family expectations of future child temperament. *Family Science, 6*(1), 356–361. https://doi.org/10.1080/19424620.2015.1076494.

Gartstein, M. A., & Rothbart, M. K. (2003). Studying infant temperament via the Revised Infant Behavior Questionnaire. *Infant Behavior and Development, 26*(1), 64–86. https://doi.org/10.1016/S0163-6383(02)00169-8.

Gouze, K. R., Lavigne, J. V., Hopkins, J., Bryant, F. B., & Lebailly, S. A. (2012). The relationship between temperamental negative affect, effortful control, and sensory regulation: A new look. *Infant Mental Health Journal, 33*(6), 620–632. https://doi.org/10.1002/imhj.21363.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Kotelnikova, Y., Olino, T. N., Klein, N. D., Kryski, K., & Hayden, P. E. (2016). Higher- and lower-order factor analyses of the Children's Behavior Questionnaire in early and middle childhood. *Psychological assessment, 28*(1), 92–108. https://doi.org/10.1037/pas0000153.

Najarpourian, S., Abdolvahab, S. S., & Neda, A. (2017). Psychometric properties of the very short form of the Children's Behavior Questionnaire (Cbq): Investigation of temperament At 3 to 7 years. *Journal of Child Mental Health, 4*(300246), 165–175.

Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment, 87*(1), 102–112. https://doi.org/10.1207/s15327752jpa8701_09.

Putnam, S. P., Gartstein, M. A., & Rothbart, M. K. (2006). Measurement of fine-grained aspects of toddler temperament: The Early Childhood Behavior Questionnaire. *Infant behavior and development, 29*(3), 386–401.

Putnam, S. P., Jacobs, J., Gartstein, M. A., & Rothbart, M. K. (2010). Development and assessment of short and very short forms of the Early Childhood Behavior Questionnaire. Poster presented at International Conference on Infant Studies, Baltimore, MD. - Google Search. (n.d.).

Putnam, S. P., Helbig, A. L., Gartstein, M. A., Rothbart, M. K., & Leerkes, E. (2014). Development and assessment of short and very short forms of the Infant Behavior Questionnaire–Revised. *Journal of personality assessment, 96*(4), 445–458.

Putnam, S; Gartstein, M, Broos, H; Casalin, S; Lecannelier, F. (2019). Cross-cultural differences in socialization goals and parental ethnotheories. In. Gartstein, M. & Putman, S., *Toddlers, parents, and culture: Findings from the joint effort toddler temperament consortium.* (pp.59–67). Routledge.

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved March 12, 2021, from https://www.r-project.org/

Revelle, W. (2018). psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois. Retrieved March 12, 2021, from https://cran.r-project.org/package=psych

Rosseel, Y. (2012). {lavaan}: An {R} Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved March 12, 2021, from http://www.jstatsoft.org/v48/i02/

Rosseel, Y. (2014). Structural Equation Modeling with lavaan. In *Using R for personality research* (pp. 1–127). Bertinoro: Ghent University.

Rothbart, M. K. (1981). Measurement of temperament in infancy. *Child Development, 52*(2), 569–578. https://doi.org/10.2307/1129176.

Rothbart, M. K. (2007). Temperament, development, and personality. *Current Directions in Psychological Science, 16*(4), 207–212.

Rothbart, Mary K., & Bates, J. E. (2007). Temperament. En *Handbook of Child Psychology*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470147658.chpsy0303

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at Three to seven years: The Children's Behavior Questionnaire. *Child Development, 72*(5), 1394–1408. https://doi.org/10.1111/1467-8624.00355.

Simonds, J., & Rothbart, M. K. (2004). The Temperament in Middle Childhood Questionnaire (TMCQ): A computerized self-report instrument for ages 7–10. *Poster Sess Present Occas Temperament Conf Athens, GA*.

Sleddens, E. F. C., Kremers, S. P. J., Candel, M. J. J. M., De Vries, N. N. K., & Thijs, C. (2011). Validating the Children's Behavior Questionnaire in Dutch children: Psychometric properties and a cross-cultural comparison of factor structures. *Psychological Assessment, 23*(2), 417–426. https://doi.org/10.1037/a0022111.

Slobodskaya, H., Kozlova, E., Han, S., Gartstein, M., & Putnam, S. (2019). Cross-cultural differences temperament. In M. Gartstein & S. Putman (Eds.), *Toddlers, parents, and culture: Findings from the joint effort toddler temperament consortium* (pp. 29–37). Routledge.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Multivariate imputation by chained equations. *Journal Of Statistical Software, 45*(3), 1–67. https://doi.org/10.1177/0962280206074463.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686.