The Brief Problem Monitor-Parent form (BPM-P), a short version of the child behavior checklist: psychometric properties in Spanish 6- to 8-year-old children

Eva Penelo [1,2]

Núria de la Osa [1,3]

José Blas Navarro [1,2]

Josep Maria Domènech [1,2]

Lourdes Ezpeleta [1,3]

[1] Unitat d'Epidemiologia i de Diagnòstic en Psicopatologia del Desenvolupament

[2] Laboratori d'Estadística Aplicada, Departament de Psicobiologia i de Metodologia de les Ciències de la Salut, Universitat Autònoma de Barcelona (Spain)

[3] Departament de Psicologia Clínica i de la Salut, Universitat Autònoma de Barcelona (Spain)

Acknowledgements

Funding was from Spanish Ministry of Economy and Competitiveness grant PSI2012-32695 and PSI2015-63965-R (MINECO/FEDER). Thanks to the *Secretaria d'Universitats i Recerca, Departament d'Economia i Coneixement de la Generalitat de Catalunya* (2014 SGR 312). We would like to thank the participating schools and families.

Post-print version

Penelo, E., de la Osa, N., Navarro, J. B., Domènech, J. M. & Ezpeleta, L. (2017). The Brief Problem Monitor-Parent form (BPM-P), a short version of the child behavior checklist: psychometric properties in Spanish 6- to 8-year-old children. *Psychological Assessment*. doi: 10.1037/pas0000428

Abstract

We provide the first validation data on the Spanish version of the Brief Problem Monitor-Parent form (BPM-P), a recently developed abbreviated version of the 120-item Child Behavior Checklist for Ages 6 to 18 (CBCL/6-18) in young schoolchildren. Parents of a community sample of 521 children aged 6-8 answered the CBCL/6-18 yearly, and the 19 BPM-P items were examined; parents also provided different measures of psychopathology. Confirmatory factor analysis of the expected 3-factor model (Attention, Externalizing, and Internalizing) showed adequate fit (RMSEA ≤ .057), and measurement invariance across sex and age was observed. Internal consistency for the derived scores was satisfactory (ω ≥ .83). Concurrent validity with the equivalent scale scores of the original full CBCL/6-18 ($r \geq .84$) and convergent validity with parents' ratings of the Strengths and Difficulties Questionnaire scores ($r \geq .52$) were good. BPM-P scores at age 7 showed good predictive accuracy for discriminating the use of mental health services (OR ≥ 1.12), functional impairment ($B \leq -1.25$), and the presence of the corresponding disorders diagnosed with an independent clinical interview, both cross-sectionally at age 7 and longitudinally at age 8 (OR ≥ 1.24). The BPM-P provides reliable and valid scores as a very brief follow-up and screening tool for assessing behavioural and emotional problems in young schoolchildren.

Public Significance Statement

This study provides evidence on validity and reliability of the Brief Problem Monitor-Parent (BPM-P) scores in 6- to 8-year-old children. Given the shortness of the questionnaire, it can be used with psychometric guarantees as a brief follow-up and screening tool for monitoring children’s functioning and responses to interventions in young schoolchildren.

Keywords

Brief Problem Monitor; parents' ratings; reliability; Spanish version; validity

The Achenbach System of Empirically Based Assessment instruments (ASEBA; Achenbach, 2009) are one of the most widely used, comprehensive and standardised tools for the assessment and screening of behavioural, emotional, social, and thought problems and strengths (Achenbach, 2015). Three separate forms are available for the assessment of children and adolescents aged 6-18, which are completed by parents or caregivers, teachers, and 11- to 18-year-old youths: the Child Behavior Checklist for Ages 6-18 (CBCL/6-18), the Teacher's Report Form for Ages 6-18 (TRF), and the Youth Self-Report for Ages 11-18 (YSR), respectively (Achenbach & Rescorla, 2001).

Focusing on parents' ratings, the CBCL/6-18 has well-established psychometric properties in clinical, nonclinical, and cross-cultural populations (Ivanova et al., 2007; Rescorla et al., 2007, 2012). It provides 8 empirically based syndrome scale scores, plus a Total Problems score: Anxiety/Depression, Withdrawn/Depressed, and Somatic Complaints, which are combined in the Internalizing symptom scale; Aggressive Behaviour and Rule-Breaking Behaviour, which are combined in the Externalizing symptom scale; and Attention Problems, Thought Problems, and Social Problems. In addition, DSM-oriented scales have also been developed, based on judgements by 58 experts from 30 societies (Achenbach, 2013).

However, one of the limitations of CBCL/6-18, and also of YSR and TRF, is its length, since the three forms consist of more than 100 items each, which limits their use for repeated measurement to track change over time (Deighton et al., 2014). Given the convenience of frequent assessment in clinical research and care contexts, Chorpita et al. (2010) developed the Brief Problem Checklist (BPC) interviews for children and caregivers based on YSR and CBCL/6-18, respectively. In doing so, the aim was to have measures available that may be general enough to be suitable across a large number of cases and contexts, brief enough to avoid being an increasing burden with frequent use, and

psychometrically sound and correlated with lengthier established measures of psychopathology and functioning. Therefore, the BPC included 12 items selected by applying item response theory and factor analysis, 6 for Internalizing and 6 for Externalizing, which were intended to correspond to the targets of treatment for many children and adolescents (Chorpita et al., 2010). Both BPC versions for child informant and caregiver informant showed satisfactory psychometric properties: a moderately correlated internal structure of two factors after conducting exploratory factor analysis, good internal consistency and test-retest reliability at a 3-month period, convergent validity with the CBCL/6-18-DSM and YSR-DSM scales, concurrent and predictive validity with the corresponding syndrome scales on both CBCL/6-18 and YSR at the same time and 6 months later, respectively, and criterion validity with an independent diagnostic structured interview.

Based on this work, Achenbach, McConaughy, Ivanova, and Rescorla (2011) developed the Brief Problem Monitor for completion by parents (BPM-P), youths (BPM-Y), and teachers (BPM-T), the latter based on TRF. The BPM forms included a third Attention dimension, consisting of 6 items selected from the original 10 items common to the original CBCL/6-18 and TRF Attention problems syndrome by discriminant analysis. The BPM-T comprises 18 items (six each for Internalizing, Externalizing, and Attention) and the BPM-P and BPM-Y comprise 19 items, with both adding an item to the Externalizing dimension to assess disobedient behaviour at home separately from disobedient behaviour at school. These three BPM forms are applicable in children and adolescents with the same age-range as the respective original ASEBA forms, according to the original authors can be completed in 1 to 2 minutes, and have been proved to provide satisfactory internal consistency (Cronbach's $\alpha \geq .74$), test-retest reliability in an 8- to 16-day interval ($r \geq .77$), criterion-related validity (effect size via multiple regression analyses, as percentage of variance uniquely accounted for by differences between scores obtained by referred vs. non-referred children for mental health

services, ≥ 11% for BPM-Y and ≥ 25% for BPM-P), and low or moderate cross-informant agreement (*r* from .18 for Internalizing between BPM-T and BPM-Y to .42 for Externalizing between BPM-P and BPM-Y) (Achenbach et al., 2011).

To date, only three other published studies have evaluated the psychometric properties of some of the three BPM forms. The BPM-P specifically has been validated in an online recruited sample mainly from the Western United States with a mean child age of 11.5 years (Piper, Gray, Raber, & Birkett, 2014), a Norwegian version in a large sample from the general population with a mean child age of 10.6 years (Richter, 2015), and a German version in clinical and population-based samples with a mean child age of 11.5 and 12.3 years, respectively (Rodenacker, Plück, & Döpfner, 2015). Findings have shown acceptable internal consistency (α between .66 for Internalizing and .87 for Attention, and ≥ .83 for the Total problems score) and excellent concurrent validity with the corresponding scales of the CBCL/6-18 original long form (*r* between .83 for Internalizing and .97 for Attention, and ≥ .88 for the Total problems score)(Piper et al, 2014; Richter, 2015; Rodenacker et al., 2015). Evidence on criterion-related validity with psychiatric diagnoses based on information reported by caregivers through an online survey for the English version has also been reported, but no standardised effect sizes were provided and the authors only mentioned that BPM-P scores in the diagnosed group were at least twice those of the non-diagnosed group (Piper et al., 2014). The BPM-P scores of the German version adequately discriminated between clinical and non-clinical children as a whole (OR ≥ 1.25) and also separately when comparing children with and without an internalizing disorder and an externalizing or attention disorder (Rodenacker et al., 2015). The latter study also provided validity evidence based on internal structure, since confirmatory factor analysis (CFA) showed an adequate fit for a 17-item and 3-factor model after two items of the Attention dimension were removed (RMSEA ≤ .077).

No published study has been conducted to date with a Spanish version of BPM. Providing empirical evidence of psychometric properties in the particular setting in which the test is to be used, in middle childhood in our case, would meet the American Psychological Association's recommendations (AERA, APA, & NCME, 2014). Moreover, no study has examined the measurement invariance of BPM scores. Measurement invariance is necessary because only when it is supported can test scores be meaningfully compared across groups of responses or over time. Thus, measurement invariance is a precursor to any group or time/condition comparison, conducted for example between sex or in longitudinal designs to evaluate trends. Invariance across sex and age are particularly important for the scale scores in question, given that researchers often investigate sex and age differences in this type of scale.

The aim of the present study was to evaluate the psychometric properties of the Spanish version of the BPM-P in 6- to 8-year-old male and female schoolchildren. More specifically, we aimed to provide evidence based on: a) internal structure, by confirming the expected 3-factor structure and by analysing measurement invariance across sex and over age by means of CFA; b) internal consistency, also with CFA; c) concurrent and predictive validity with the CBCL/6-18; d) convergent and criterion-related validity with other psychopathology measures (dimensional and diagnostic); and e) relations to external variables such as use of services, functional impairment, age and sex.

Method

Participants

The data are part of a larger longitudinal project on behavioural problems in children followed from age 3 (Ezpeleta, de la Osa, & Domènech, 2014). We used a cross-sectional two-phase design which started with the inclusion of 2,283 children randomly selected from the census of grade P3 preschoolers (3-year-olds) in the area of Barcelona (Catalonia, Spain).

In the first phase of the study participated 1,341 families (58.7%), whose parents answered the Strengths and Difficulties Questionnaire parents' version (SDQ-P, see below) for screening purposes of their children. There were no differences regarding sex between participants and refusals.

In the second phase, in order to ensure the presence of children with behavioural problems, all children with a positive screening and around 30% of children with a negative screening (the number of children needed in the negative screening was calculated to guarantee statistical power for the subsequent analyses) randomly selected were invited to continue. Children with pervasive developmental disorders or intellectual disability were excluded. 10.6% of families invited declined to continue and the final sample in this second phase comprised 622 families. The mean age of children was 3.0 (*SD* = 0.16), 311 were girls (50.0%), 557 (89.5%) were Caucasian, and the family socioeconomic status (SES; Hollingshead, 1975) was as follows: 21.3% low, 44.9% middle, and 33.8% high. No statistically significant differences were found by sex ($p = .82$) or type of school ($p = .85$) between participants and those who refused to participate.

Given that CBCL/6-18 is applicable from age 6, this study includes data from the age 6 wave; thus, assessments at ages 6 (*N* = 510, 49.6% girls), 7 (*N* = 496, 50.7% girls), and 8 (*N* = 474, 50.3% girls) were used for the analysis. No statistically significant differences were observed by sex ($p \geq .464$), SES ($p \geq .133$), or SDQ-P scores at study entry ($p \geq .053$, after applying Finner's correction for multiple comparisons)(Finner, 1993) between participants remaining at each follow-up and those not retained at ages 6, 7 and 8. The CBCL/6-18 (see below) data were available for 481 children (244 girls and 237 boys) at age 6, 460 (234 girls and 226 boys) at age 7, and 426 (217 girls and 209 boys) at age 8.

Instruments

Child Behavior Checklist for Ages 6-18 (CBCL/6-18; Achenbach & Rescorla, 2001)/*Brief Problem Monitor-Parent form* for Ages 6-18 (BPM-P; Achenbach et al., 2011). The CBCL/6-18 comprises 120 items assessing behavioural and emotional problems that are answered on a 3-point Likert-type scale (0: *not true*, 1: *somewhat or sometimes true*, 2: *very true or often true*) by parents. Items are summed into 8 narrowband and 2 broadband syndrome scales, and a total score, with higher scores reflecting higher problem levels.

Nineteen items of the CBCL/6-18 make up the BPM-P for children aged 6-18: 6 items for Attention (all 6 from the CBCL/6-18 Attention Problems scale), 7 items for Externalizing (all 7 from the CBCL/6-18 Aggression scale), and 6 items for Internalizing (5 from the CBCL/6-18 Anxiety/Depression and 1 from the CBCL/6-18 Withdrawn/Depressed scales).

Diagnostic Interview of Children and Adolescents for Parents of Preschool Children and Young Children (DICA-PPYC; Ezpeleta, de la Osa, Granero, Domènech, & Reich, 2011). The DICA-PPYC is a semi-structured diagnostic interview that covers a wide range of categorical diagnoses for 3- to 7-year-old children following the Diagnostic and Statistical Manual of Mental Disorders criteria (DSM-5; APA, 2013). The diagnoses considered in the present study were: Attention-Deficit/Hyperactivity Disorder (ADHD), Oppositional Defiant Disorder (ODD), Conduct Disorder (CD), Major Depressive Disorder, and Anxiety Disorders (generalised anxiety disorder, separation anxiety disorder, social phobia, and specific phobia). Additional questions asked to parents examined service use (whether professional help had been sought, including consultation or treatment).

Strengths and Difficulties Questionnaire Parent version (SDQ-P$^{4\text{-}17}$; Goodman, 1997). The SDQ-P is a brief screening tool designed to asses children's psychopathology, and comprises 25 items with 3 response options (0: *not true*; 1: *somewhat true*; 2: *certainly true*) answered by parents. It has 5-item 5 subscales: conduct problems, hyperactivity/inattention, emotional symptoms, peer relationship problems, and prosocial behaviour. We also obtained

two broader dimensions, externalizing (sum of conduct and hyperactivity subscale items) and internalizing (sum of emotional and peer subscale items)(Goodman, Lamping, & Ploubidis, 2010), a total difficulties score (sum of the four problem-focused subscales), and the impact supplement (level of distress, social impairment, and burden that a problem of the child causes to others). Reverse items are recoded so that higher scores reflect more difficulties. We applied the official Spanish version (Ezpeleta, Granero, de la Osa, Penelo, & Domènech, 2013).

Children's Global Assessment Scale (CGAS; Shaffer et al., 1983). The CGAS assesses overall functional impairment in a numeric scale provided by the interviewer of the diagnostic interview. Ratings can range between 100 (*normal functioning*) and 1 (*maximum impairment*), with a cut-off point of 70 below which the functioning of the child is considered impaired (Ezpeleta, Granero, & de la Osa, 1999).

Procedure

The ethics committee of the authors' university approved the longitudinal study, which was fully explained to the heads of the 54 schools enrolled and the children's parents. Informed written consent from families recruited at the schools was obtained at the beginning of the project, when children were 3-year-old. Trained interviewers blind to the children's screening group (see detailed explanation in Ezpeleta et al., 2014) applied yearly the DICA-PPYC to parents at the school. Parents also completed the CBCL/6-18 yearly (at ages 6, 7, and 8); SDQ-P answered by parents and CGAS ratings provided by interviewers at age 7 were also considered.

Statistical Analyses

Analyses were performed with SPSS24 (IBM Corporation, 2016) and MPlus7.11

(Muthén & Muthén, 1998-2013). Given that in the second phase of the project the sampling for participants' selection was conditioned to the presence/absence of behaviour problems, analyses were conducted with weight procedures. The WEIGHT BY command of SPSS and the WEIGHT option of MPlus were used to identify the variable that contained sampling weight information (inverse proportion to the probability of being selected). In doing so, the distribution of behaviour problems was considered in statistical inferences.

Internal structure of the BPM-P items was analysed with CFA for categorical indicators with Weighted Least Squares Means and Variance (WLSMV) method of estimation and theta parameterization. The low endorsement for the *very true or often true* option (coded as 2) yielded to zero cells in contingency tables for several pairs of items; therefore, as suggested by Jöreskog (2002-2005), we reduced the number of categories, by collapsing the two highest categories into one category; then, for CFA all BPM-P items will be dichotomous (0: *not true* vs. 1 or 2: *somewhat* or *certainly true*). First, for each group of responses at ages 6, 7 and 8, a 3-factor model was evaluated within each subsample of girls and boys; then, a multi-group configural invariance model (equal form, which implies the same number of factors and the same items defining each construct) with all parameters freed to vary across sex was established. The following goodness-of-fit indices were used (Jackson, Gillaspy, & Purc-Stephenson, 2009): χ^2, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Tucker and Lewis Index (TLI). We followed the usual cut-off points (Brown, 2006): RMSEA < .08 and CFI and TLI > .90 would indicate acceptable fit, and RMSEA < .06 and CFI and TLI > .95 would indicate excellent fit. We considered RMSEA as the main index of model fit due that it has been shown to be less sensitive than other global fit measures to distribution and sample size and more sensitive to model misspecification (Hu & Bentler, 1998).

Second, measurement invariance across sex was analysed at each age. This procedure

involved the progressive comparison of nested models increasingly constrained (from least to most restrictive). The common sequence (Vandenberg & Lance, 2000) of tests used was as follows: equivalence of factor loadings (metric or weak measurement invariance), item thresholds (scalar or strong measurement invariance), item uniquenesses (strict measurement invariance), factor variances, factor covariances, and latent means (the three last conforming structural invariance). Weak (metric) measurement invariance entails that the constructs themselves have the same meaning across groups; if not, it does not make sense to evaluate how relationships involving the factor differ across groups (because the factor itself differs). Strong (metric plus scalar) measurement invariance enables the interpretation of differences on group responses based on latent means; if not, it does not make sense to evaluate if factor means differ across groups (because something else is causing those differences). Strict measurement invariance would ensure that the relationship between the factors and the observed item scores are the same across groups and may be a requirement for comparisons of observed scores (which do not control for measurement error). Lastly, and given measurement invariance, structural invariance evaluates if groups of responses differ in their distribution and/or means of the construct.

And third, longitudinal measurement invariance over age was analysed. This procedure involved a single-sample approach, with responses at ages 6, 7 and 8 considered as repeated measures. Thus, taking into account that ratings are non-independent, error covariances between analogous items over time were also freely estimated (Ferrando, 2000), in addition to factor covariances. For both types of invariance analysis, model identification for each step was established as detailed in Ezpeleta and Penelo (2015), using the fixed-factor method. The α level for scaled difference chi-square tests between nested models with the DIFFTEST option of Mplus for WLSMV method of estimation was set at .01 (e.g., Ezpeleta & Penelo, 2015; Ferrando, 2000). Additionally, a decrement in CFI grater than .010 and an

increment in RMSEA greater than .015 were indicative of worse fit and, hence, of non-equivalence, despite these guidelines have been reported for maximum likelihood estimation method (Chen, 2007; Cheung & Rensvold, 2002). If full invariance was not accomplished, modification indexes were examined in order to free parameters one at a time to be able to evaluate partial invariance. Internal consistency of the BPM-P scales was assessed with the omega coefficient (McDonald, 1999).

Next, relations between the derived direct BPM-P scores and external measures were analysed. The relationship of sex and age with BPM-P scores was analysed with two-factor mixed ANOVA considering a 2 (sex) × 3 (age: 6, 7, and 8) design; when the effect of age was statistically significant, trends over time were also evaluated. These analyses based on direct scores complemented those testing equivalence of latent mean with invariance analyses (6^{th} step), which had been performed with combined categories. Effect sizes were measured with Cohen's *d* and interpreted following the usual rules of thumb: a small effect for absolute values ranging between 0.20 and 0.50, medium between 0.50 and 0.80 and large above 0.80 (Cohen, 1992). Pearson's correlation coefficients (*r*) were used to evaluate the concurrent validity between BPM-P and the corresponding CBCL/6-18 scale scores and the convergent validity between BPM-P and SDQ-P scores at the same age. In order to investigate how the briefer measures would fare relative to the longer measures, we examined how well BPM-P scores at ages 6 and 7 would predict CBCL/6-18 scores at older ages (6-7, 7-8, and 6-8) and we compared this with how well CBCL/6-18 scores from ages 6 and 7 predicted CBCL/6-18 scores at the same older ages, as in Chorpita et al. (2010), with the test of difference between two dependent correlations (Steiger, 1980). The association between BPM-P scores and use of services (DICA-PPYC clinical interview) and functional impairment (CGAS score) was analysed with logistic and linear regression adjusted by sex, respectively. Lastly, the screening accuracy of the BPM-P scores at age 7 (screening tool) to identify the presence of

specific DSM-5 disorders according to the DICA-PPYC diagnostic interview (considered the reference measure) at age 7 (criterion-related validity) and at age 8 (predictive validity) was estimated through odds ratios (OR) and the area under the receiver operator curve (AUC) in logistic regression models adjusted by sex. AUCs below .65 were considered poor, between .65 and .75 moderate, and above .75 very good (Hosmer & Lemeshow, 2000). The DSM-5 disorders considered were ADHD, any externalizing disorder (ODD and/or CD), any internalizing disorder (depression, separation anxiety, generalised anxiety, specific phobia and/or social phobia), and any disorder (one or more of the aforementioned).

Results

Missing Data Analyses

Missing responses for the 19 BPM-P items were low (Graham, 2009): 0.48% at age 6, 0.14% at age 7, and 0.38% at age 8; only 20 participants (4.2%) at age 6, 11 (2.4%) at age 7, and 16 (3.8%) at age 8 exhibited missing values for one or more items. Hence, and given that the amount of missing data was not substantial, the WLSMV method of estimation with pairwise deletion was used for CFAs (Asparouhov & Muthén, 2010).

Internal Structure, Measurement Invariance and Internal Consistency

Table 1 displays the results of CFAs and measurement invariance analyses. Fit for baseline models for each sex (models #0a and #0b in Table 1) and for configural invariance across sex within responses at age 6, 7, and 8 (models A1, B1, and C1 in Table 1, respectively) was mostly satisfactory (RMSEA $\leq$ .071), except for boys at age 8 (RMSEA = .095). Thus, support for the 3-factor model solution was obtained. Full measurement and structural invariance across sex was found at age 7 (model B6 in Table 1), which indicates that all the parameters were equivalent across girls and boys. At age 6, and based on scaled

difference chi-square test but not on CFI and RMSEA changes, full strict invariance was not achieved, since two uniquenesses (items 21 and 71) were not equivalent (models A4 and A3+ in Table 1), but partial strict invariance can be assumed because more than 80% of parameters were found to be invariant (Dimitrov, 2010). Additionally, and also based only on scaled difference chi-square test, latent means for Attention were slightly higher in boys than girls (d = 0.382, p = .001; model A6+ in Table 1). At age 8, only the factor loading for item 97 ("threatens people") was found to be higher for girls than boys (model C2+ in Table 1) and one uniqueness (item 23) was also non-invariant across sex (models C4 and C3+ in Table 1), but the latter only based on scaled difference chi-square test, therefore partial weak and strict invariance was assumed (model C6 in Table 1). Table 1 (bottom) also presents fit for CFA baseline models at each age separately (models D0# in Table 1). Results for the repeated-measure invariance analysis over age showed that complete measurement and structural invariance was attained (model D6 in Table 1).

Fit for this final fully constrained model over age was satisfactory (RMSEA = .026). All standardised factor loadings showed values above .40 and were statistically significant (p < .001) and factor correlations were between .55 and .65 (p < .001), providing evidence for three inter-related but distinguishable factors (Figure 1). Regarding longitudinal agreement in the whole sample, factor correlation values between age 6, age 7 and age 8 ratings for analogous factor pairs ranged from .75 (Internalizing between ages 7 and 8) to .86 (Attention between ages 7 and 8). Factor correlation values between non-analogous factor pairs were lower (.32-.57).

Internal consistency was satisfactory, with omega values of .92 for Attention, .89 for Externalizing, .83 for Internalizing, and .93 for Total Problems scores.

Distribution of Raw Scores, Relation to Sex, and Trends over time

All the results from now on are based on direct scale scores derived from the sum of the corresponding original item ratings. Table 2 shows the weighted means and standard deviations for the BPM-P scale scores, effect sizes of mean differences across sex based on Cohen's *d* coefficient, and results for the 2 × 3 ANOVA. The interaction effect (sex × age) was not statistically significant for any of the BPM-P scale scores ($p \geq .125$). One statistical difference emerged by sex for Attention ($p = .022$), boys scoring slightly higher than girls. Differences over time were only observed for Externalizing scores ($p = .014$), following a negative linear trend ($p = .017$) plus a positive quadratic trend ($p = .011$), with a decrease from age 6 to 7 and a slight increase from age 7 to 8; and Internalizing scores showed a slightly positive linear trend over time ($p = .001$). However, sex differences were within the range of null or small effect sizes ($|d| \leq 0.31$) and null or very small regarding age ($|d|$ for repeated measures ≤ 0.23; detailed coefficients not shown).

Relation to Other Measures

In relation to concurrent validity with CBCL/6-18 at ages 6, 7 and 8, BPM-P scores were strongly correlated with the equivalent direct subscale scores of the larger CBCL/6-18 at the same ages ($r \geq .84$) (Table 3, top), taking into account that BPM-P Attention includes 6 of the CBCL/6-18 Attention Problems items, BPM-P Externalizing includes 7 of the CBCL/6-18 Aggressive behaviour items, and Internalizing includes 5 of the CBCL/6-18 Anxiety/Depression and only 1 of the CBCL/6-18 Withdrawn/Depressed items. Values between the remaining subscale scores were lower ($r \leq .64$) (Table 3, centre). The pattern of correlations was very similar across the three age groups. In addition, associations for Attention and Externalizing were slightly higher than for Internalizing scores.

Regarding convergent validity with SDQ-P, BPM-P scores at age 7 were moderately or highly associated with the theoretically most closely related 5-item SDQ-P subscales and

SDQ-P total difficulties at age 7 ($r \geq .52$), whereas values for divergent validity were lower ($r \leq .39$) (Table 3, bottom).

As can be seen in Table 4, predicted values using BPM-P scores at ages 6 and 7 performed very similarly to CBCL/6-18 at ages 6 and 7 in predicting CBCL/6-18 one year later (median *r* values: .69 *vs* .73; *p* values for all tests of differences between pairs of *r* values ≥ .051) with one exception (Internalizing between age 7 and 8, with CBCL/6-18 performing slightly better than BPM-P; *r* values: .66 *vs* .53; $p = .005$), and two years later (median *r* values: .66 *vs* .68; $p \geq .106$).

In total, 7.6% of the families (5.7% for girls and 9.4% for boys) reported use of services (yes/no, derived from the DICA-PPYC clinical interview) due to the child's psychological problems at age 7, while 20.8% (20.1% for girls and 21.6% for boys) showed functional impairment (CGAS score below cut-off of 70) associated with the child's symptoms at age 7. As shown in Table 5, all BPM-P scores were positively related to the use of services (OR ≥ 1.12, $p \leq .036$; the higher the BPM-P score, the higher the odds of consultation). Similarly, higher BPM-P scores were associated with lower CGAS scores ($B \leq -1.25$, $p < .001$), meaning greater impairment. In both cases, effects for BPM-P subscale scores were slightly higher than for the Total Problems score.

All BPM-P scale scores were associated with the corresponding DSM-5 diagnoses and the Total Problems score was associated with the presence of any of the disorders considered (OR ≥ 1.24, $p \leq .004$), showing that the instrument's screening accuracy was satisfactory cross-sectionally (questionnaire and interview at age 7; Table 6, top) and longitudinally (BPM-P at age 7 and clinical interview at age 8; Table 6, bottom). The discriminative ability was between high and very high for Attention, Externalizing and Total Problem scores, AUC ranging from .71 (Total Problems and any disorder one year later) to .93 (Attention and ADHD the same year) and Nagelkerke's R^2 between .24 and .54, moderate for Internalizing

cross-sectionally (AUC = .71, R^2 = .10), and poor but almost moderate for Internalizing longitudinally (AUC = .64, R^2 = .04).

Discussion

This is the first study to analyze the psychometric properties of the BPM-P with the Spanish version and in a specific sample of 6- to 8-year-old schoolchildren. The BPM-P items presented a satisfactory 3-factor structure (Attention, Externalizing, and Internalizing), as the German version had done, despite two of the Attention items not having been included in this case (Rodenacker et al., 2015). Moreover, nearly full measurement invariance was achieved across sex and over age and, to our knowledge, both approaches had not been published before with any of the BPM forms. Our findings show that all factor loading and item thresholds were equivalent, attending to both chi-square tests as well as CFI and RMSEA changes, with one exception: Item 97 ("threaten") loaded higher onto the Externalizing factor in girls than in boys aged 8 (.77 vs. .35), but this issue could be due to the very low endorsement of this question in the community sample studied. The fact that partial or almost full strict invariance based on chi-square test or complete strict invariance based on CFI and RMSEA changes can be assumed implies that comparisons of observed BPM-P scores are readily interpretable and differences found would reflect true differences in the latent construct. Longitudinal agreement in a 1-year interval between analogous factor pairs was also good regarding factor correlation values obtained with the longitudinal CFA in the whole sample; our findings are aligned with those obtained with the usual test-retest strategy in an 8- to 16-day interval (Achenbach et al., 2011). Internal consistency based also on the final fully constrained longitudinal CFA model was also satisfactory, with higher values for Attention and lower values for Internalizing subscale scores, as in all previous BPM validation studies (Achenbach et al., 2011; Piper et al., 2014; Richter, 2015; Rodenacker et al., 2015).

In relation to the direct mean scores, the ratings given to boys were slightly higher than those assigned to girls for Attention. However, taking into account the magnitude of the difference found (main effect of 0.50 points), we consider that the influence of child sex on Attention scores is almost negligible, effect sizes being very small, although they have been found to be slightly more frequent in children aged 6-11 than in adolescents aged 12-16 (Rescorla et al., 2007). Trends over time were also negligible, with mean differences ranging from 0.06 to 0.26 points, which is aligned with the fact that the same norms are usually applied for interpretation of CBCL/6-18 scores in children aged 6-11 (for Spanish norms see, for example, UED, 2013-2016). Moreover, results for comparisons of direct scores between sex and over time matched those found for equivalence of latent means through invariance analyses, showing low or null effects.

The high and significant correlations between the BPM-P scores and the corresponding full-length CBCL/6-18 scores at the same age provide evidence of concurrent validity, indicating that the short form is measuring substantively the same constructs as the CBCL/6-18. Values were slightly higher for Attention and lower for Internalizing, as in previous studies (Piper et al., 2014; Richter, 2015; Rodenacker et al., 2015), which may in part be explained in terms of higher and lower internal consistency, respectively, as seen before. However, as pointed out by Richter (2015), even though these findings are limited to the fact that analyses were based on the same data set based on responses to the long CBCL/6-18 version, high correlations have been also found between BPC interviews and CBCL/6-18 questionnaire scores (Chorpita et al., 2010).

In addition, the moderate to high and significant correlations between the BPM-P scores and external measures of similar or related constructs obtained with a well-established screening tool such as the SDQ-P provide evidence of convergent validity. As expected, the higher coefficients were obtained between BPM-P Attention and SDQ-P Hyperactivity and

also SDQ-P Externalizing, the latter comprising Hyperactivity and Conduct scores, followed by coefficients between BPM-P Total problems and SDQ Total difficulties, and next between BPM-P Externalizing and SDQ-P Conduct and between BPM-P Internalizing and SDQ-P Emotional. Values between Externalizing BPM-P and SDQ-P and between Internalizing BPM-P and SDQ-P scores were slightly lower, but still adequate (*r* around .55). Regarding Externalizing measures of both questionnaires, in the case of the BPM-P this only consists of seven items from the CBCL/6-18 Aggressive behaviour scale and none from the Rule-breaking behaviour scale, whereas for the SDQ-P it includes the five items from Hyperactivity and the five from Conduct, the latter asking about aggressive behaviour but also about rule-breaking such as lying or stealing. Therefore, and despite the coincidence in labels for both measures, they do not exactly assess the same construct, and consequently the expected value for the correlation between them would be lower than for more matching measures. The same applies to the Internalizing measures for BPM-P and SDQ-P, which are respectively made up of five items from Anxious/Depressed and one from Withdrawn/Depressed of the CBCL/6-18 and five items from Emotional but also five from Peer of the SDQ-P. Nevertheless, values for less or nearly non-related scale measures were lower, and despite the fact that the design cannot strictly be considered a multi-trait-multi-method approach, our results support the convergent and divergent validity of BPM-P with SDQ-P.

All BPM-P scores showed good criterion-related validity to differentiate between children with and without a diagnosis, both cross-sectionally as in previous studies (Achenbach et al., 2011; Piper at al., 2014; Rodenacker et al., 2015) and also longitudinally, specially subscales related to behaviour problems (Attention and Externalizing). The poorer ability of ASEBA Internalizing scores to discriminate children suffering an internalizing disorder has also been reported in preschoolers (de la Osa, Granero, Trepat, Domènech, &

Ezpeleta, 2016), younger schoolchildren (de Wolff, Vogels, & Reijneveld, 2014) and adolescents (Ferdinand, 2008). Conversely, BPM-P Internalizing scores proved to be the most useful for identifying cases with clinically impaired functioning and using professional services due to mental health problems.

Taken together, the results suggest that BPM-P scores are as reliable and valid as CBCL/6-18 ones, and can thus be considered as a trustworthy equivalent of the original larger version, when it is not feasible to administer a time-consuming tool, either the longer checklist or an extensive interview, or when a follow-up is required. One of the strengths of the present study may be the fact that, to date this is the first study that provides convergent evidence between BPM and SDQ scores, given that both ASEBA and SDQ child forms (CBCL/6-18 and SDQ-P) represent some of the most often internationally used standardised measures of child and adolescent emotional and behavioural problem symptoms in mental health (Richter, 2015), and both have been identified among the 11 existing broadband measures of mental health and wellbeing outcomes showing strong psychometric evidence, including sensitivity to change, and suitability for use in routine practice in child and adolescent mental health services, both having been translated into over 70-80 different languages (Deighton et al., 2014). In spite of the fact that we did not assess sensitivity to change in a strict sense, for example comparing an intervention group to a non-treated group both with frequent assessment over time, we found that BPM-P scores predicted CBCL/6-18 scores one or two years later as well as CBCL/6-18 scores themselves did, with the exception of the Internalizing problems dimension. Moreover, the *r* values we found after 1-2 years were similar or even higher (*Md* = .67; see Table 4) than those found by Chorpita et al. (2010) with the BPC-caregiver interview and CBCL/6-18 scores 6 months later (*Md* = .61).

However, some limitations should be taken into account on interpreting the present results. Since we studied a sample of the general population, two issues should be pointed out.

First, given the low endorsement for the *certainly true* option (coded as 2), BPM-P items were dichotomised for factor analyses. Therefore, measurement invariance testing equivalence of thresholds was based on combined categories (0 *vs*. 1 or 2) that do not represent the real scale (0, 1, or 2). And second, the discriminative power for BPM-P scores may have been affected because of the low frequency of psychopathology in community samples. Results should be generalised to the children population with caution, given that few families of low SES participated, leading to underestimated ratings of problem scores, as has been observed in mothers responding to the BPM-P (Piper et al., 2014) and also in parents answering the CBCL/6-18 (Raadal, Milgrom, Cauce, & Mancl, 1994). Thus, it would be highly desirable to complement data with reports from teachers and self-reports. For the former, and unlike other previous studies (Achenbach et al., 2011; Richter, 2015; Rodenacker et al., 2015), we were not able to apply, and subsequently validate, the TRF or BPM-T; and for the latter, given that we focused on 6- to 8-year-old children, it was not suitable to administer the YSR or BPM-Y in such young children, although some children younger than 11 may be able to complete it (Achenbach et al., 2011). Future follow-ups of our longitudinal study could fill this gap, combined with the pending matter of collecting data from teachers and also from parents of older children.

The availability of the BPM-P as a psychometrically adequate and short easy-to-use follow-up and screening tool equivalent across sex and over age is useful and important in early school years for identifying youths at high risk of behavioural or emotional problems; it can be also used for treatment monitoring and evaluation over time, in the same way as the earlier BPC interviews (Beidas et al., 2015; Chorpita et al., 2010). In doing so, many children would benefit from early detection and adequate intervention or prevention programmes, avoiding abuse of medical resources and services, chronic conditions, comorbidity and personal and family dysfunction throughout later childhood and adolescence.

References

Achenbach, T. M. (2009). *The Achenbach System of Empirically Based Assessment (ASEBA): Development, findings, theory, and applications*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Achenbach, T. M. (2013). *DSM-oriented guide for the Achenbach System of Empirically Based Assessment (ASEBA)*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Achenbach, T. M. (2015). Multicultural evidence-based assessment using the Achenbach System of Empirically Based Assessment (ASEBA) for ages 1½-90+. *Psychologia. Avances de la Disciplina*, *9*(2), 13-23. Available in: http://www.redalyc.org/articulo.oa?id=297241658001

Achenbach, T. M., McConaughy, S. H., Ivanovaa, M., & Rescorla, L. A. (2011). *Manual for the ASEBA Brief Problem Monitor ™ (BPM)*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families. Available in: http://www.aseba.org/ASEBA%20Brief%20Problem%20Monitor%20Manual.pdf

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

AERA (American Educational Research Association), APA (American Psychological Association) & NCME (National Council on Measurement in Education) (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

APA (American Psychiatric Association). (2013). *Diagnostic and Statistical Manual of mental disorders* (5th ed.). Washington, DC: Author.

Asparouhov, T. & Muthén, B. (2010). *Weighted least squares estimation with missing data*. Technical Report. Retrieved from

https://www.statmodel.com/download/GstrucMissingRevision.pdf

Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., … Mandell, D. S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice, 22*(1), 5-19. doi:10.1016/j.cbpra.2014.02.002

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (1st ed.). New York: Guilford.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. doi:10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255. doi:10.1207/S15328007SEM0902_5

Chorpita, B. F., Reise, S., Weisz, J. R., Grubbs, K., Becker, K. D., & Krull, J. L. (2010). Evaluation of the Brief Problem Checklist: Child and caregiver interviews to measure clinical progress. *Journal of Consulting and Clinical Psychology, 78*(4), 526-536. doi:10.1037/a0019602

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: A review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health, 8*(1), 14. doi:10.1186/1753-2000-8-14

de la Osa, N., Granero, R., Trepat, E., Domènech, J. M., & Ezpeleta, L. (2016). The discriminative capacity of CBCL/1½-5-DSM5 scales to identify disruptive and

internalizing disorders in preschool children. *European Child and Adolescent Psychiatry, 25*(1), 17-23. doi:10.1007/s00787-015-0694-4

de Wolff, M. S., Vogels, A. G. C., & Reijneveld, S. A. (2014). The empirical versus DSM-oriented approach of the Child Behavior Checklist. *European Journal of Psychological Assessment, 30*(1), 22-30. doi:10.1027/1015-5759/a000164

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*, 121-149. doi: 10.1177/0748175610373459

Ezpeleta, L., de la Osa, N., & Domènech, J. M. (2014). Prevalence of DSM-IV disorders, comorbidity and impairment in 3-year-old Spanish preschoolers. *Social Psychiatry and Psychiatric Epidemiology, 49*(1), 145-155. doi:10.1007/s00127-013-0683-1

Ezpeleta, L., de la Osa, N., Granero, R., Domènech, J. M., & Reich, W. (2011). The Diagnostic Interview of Children and Adolescents for Parents of Preschool and Young Children: Psychometric properties in the general population. *Psychiatry Research, 190*(1), 137-144. doi:10.1016/j.psychres.2011.04.034

Ezpeleta, L., Granero, R., & de la Osa, N. (1999). Evaluación del deterioro en niños y adolescentes a través de la Children's Global Assessment Scale (CGAS) [Assessment of impairment in children and adolescents with the Children's Global Assessment Scale (CGAS)]. *Revista de Psiquiatría Infanto-Juvenil, 1*, 18-26.

Ezpeleta, L., Granero, R., de la Osa, N., Penelo, E., & Domènech, J. M. (2013). Psychometric properties of the Strengths and Difficulties Questionnaire$^{3-4}$ in 3-year-old preschoolers. *Comprehensive Psychiatry, 54*(3), 282-291. doi:10.1016/j.comppsych.2012.07.009

Ezpeleta, L., & Penelo, E. (2015). Measurement invariance of Oppositional Defiant Disorder dimensions in 3-year-old preschoolers. *European Journal of Psychological Assessment, 31*(1), 45-53. doi:10.1027/1015-5759/a000205

Ferdinand, R. F. (2008). Validity of the CBCL/YSR DSM-IV scales Anxiety Problems and Affective Problems. *Journal of Anxiety Disorders, 22*(1), 126-134. doi:10.1016/j.janxdis.2007.01.008

Ferrando, P. J. (2000). Testing the equivalence among different item response formats in personality measurement: A structural equation modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(2), 271-286. doi:10.1207/S15328007SEM0702_7

Finner, H. (1993). On a monotocity problem in ste-down mulitple test procedures. *Journal of the American Statistical Association, 88*(423), 920-923.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*(5), 581-586. doi:10.1111/j.1469-7610.1997.tb01545.x

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British parents, teachers and children. *Journal of Abnormal Child Psychology, 38*(8), 1179-1191. doi:10.1007/s10802-010-9434-x

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. doi:10.1146/annurev.psych.58.110405.085530

Hollingshead, A. B. (1975). *Four-factor index of social status*. Unpublished manuscript, Department of Sociology, Yale University, New Haven.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Hu, L. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

doi:10.1037/1082-989X.3.4.424

IBM Corporation. (2016). IBM SPSS Statistics for Windows (Version 24) [Computer Software]. Armonk, NY: Author.

Ivanova, M. Y., Dobrean, A., Dopfner, M., Erol, N., Fombonne, E., Fonseca, A. C., … Chen, W. J. (2007). Testing the 8-syndrome structure of the child behavior checklist in 30 societies. *Journal of Clinical Child and Adolescent Psychology*, *36*(3), 405-417. doi:10.1080/15374410701444363

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6-23. doi:10.1037/a0014694

Jöreskog. K. G. (2002-2005). *Structural equation modeling with ordinal variables using LISREL*. Retrieved from http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf

McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale: Erlbaum.

Muthén. L. K. & Muthén, B. O. (1998-2013). *Mplus user's guide. Seventh edition*. Los Angeles: Muthén & Muthén.

Piper, B. J., Gray, H. M., Raber, J., & Birkett, M. A. (2014). Reliability and validity of Brief Problem Monitor, an abbreviated form of the Child Behavior Checklist. *Psychiatry and Clinical Neurosciences, 68*(10), 759-767. doi:10.1111/pcn.12188

Raadal, M., Milgrom, P., Cauce, A. M., & Mancl, L. (1994). Behavior problems in 5- to 11-year-old children from low-income families. *Journal of the American Academy of Child and Adolescent Psychiatry, 33*(7), 1017-1025. doi:10.1097/00004583-199409000-00013

Rescorla, L., Achenbach, T., Ivanova, M. Y., Dumenci, L., Almqvist, F., Bilenberg, N., … Verhulst, F. (2007). Behavioral and emotional problems reported by parents of children ages 6 to 16 in 31 societies. *Journal of Emotional and Behavioral Disorders, 15*(3), 130-142. doi:10.1177/10634266070150030101

Rescorla, L., Ivanova, M. Y., Achenbach, T. M., Begovac, I., Chahed, M., Drugli, M. B., … Zhang, E. Y. (2012). International epidemiology of child and adolescent psychopathology II: Integration and applications of dimensional findings from 44 societies. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*(12), 1273-1283.e8. doi:10.1016/j.jaac.2012.09.012

Richter, J. (2015). Preliminary evidence for good psychometric properties of the Norwegian version of the Brief Problems Monitor (BPM). *Nordic Journal of Psychiatry, 69*(3), 174-178. doi:10.3109/08039488.2014.951070

Rodenacker, K., Plück, J., & Döpfner, M. (2015). Fragebogen zum Problem-Monitoring für Eltern, Lehrer und Jugendliche–eine deutsche Fassung des Brief Problem Monitor (BPM). Konstruktion, Reliabilität und Validität. *Zeitschrift für Klinische Psychologie und Psychotherapie, 44*(3), 197-206. doi:10.1026/1616-3443/a000307

Shaffer, D., Gould, M. S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., & Aluwahlia, S. (1983). A children's global assessment scale (CGAS). *Archives of General Psychiatry, 40*(11), 1228-1231.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245-251. doi:10.1037/0033-2909.87.2.245

UED (Unitat d'Epidemiologia i de Diagnòstic en Psicopatologia del Desenvolupament). (2013-2016). *Baremos para CBCL/6-18 $_{2001}$. Población española*. Unpublished manuscript. www.ued.uab.es

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. doi:10.1177/109442810031002

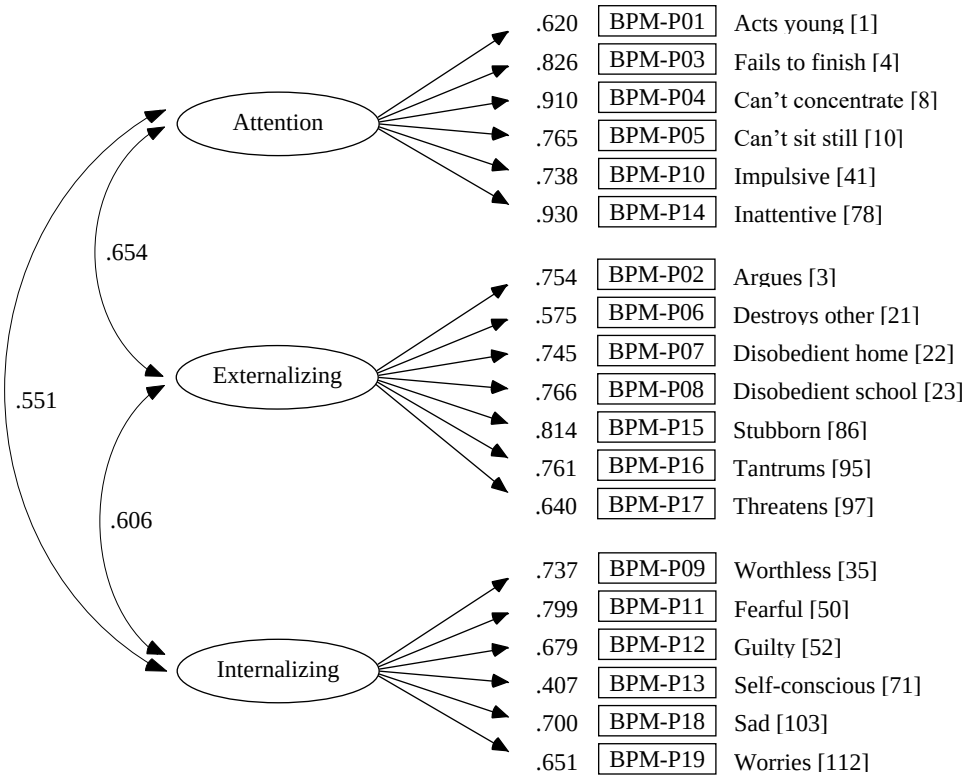

Figure 1: Standardised factor loadings and factor correlations for the final fully constrained solution of BPM-P items over age 6, 7, and 8 (Model D6 in Table 1). In brackets: original CBCL/6-18 item numeration

Table 1: Fit indices for measurement invariance analyses across sex within each age group (top) and repeated-measures measurement invariance analysis across age 6, 7, and 8 (bottom)

	Goodness-of-fit indices				Comparison				
Model fit and invariance	χ^2 (df)	CFI	TLI	RMSEA (90% CI)	Models	$\Delta\chi^2$ (Δdf)[a]	*p*	ΔCFI	ΔRMSEA
Across sex at age 6 (*N* = 481)									
A0a: females (*n* = 244)	194.6 (149)	.967	.962	.035 (.019; .071)					
A0b: males (*n* = 237)	259.1 (149)	.949	.941	.056 (.044; .067)					
A1: configural (equal form)	459.8 (298)	.954	.947	.048 (.039; .056)					
A2: A1 plus equal factor loadings (weak invariance)	481.3 (314)	.953	.948	.047 (.039; .055)	A2 *vs* A1	31.9 (16)	.010	−.001	−.001
A3: A2 plus equal thresholds (strong invariance)	505.2 (330)	.950	.948	.047 (.039; .055)	A3 *vs* A2	32.0 (16)	.010	−.003	.000
A4: A3 plus uniquenesses free (strict invariance)[b]	469.4 (311)	.955	.951	.046 (.037; .054)	A3 *vs* A4[b]	45.3 (19)	*.001*	−.005	.001
A3+: A2 except 2 uniquenesses unequal (partial)[b]	490.2 (328)	.954	.952	.045 (.037; .053)	A3+ *vs* A4[b]	29.8 (17)	.028	−.001	−.001
A5: A4 plus equal factor variances and covariances	475.5 (334)	.960	.959	.042 (.033; .050)	A5 *vs* A3+	5.00 (6)	.054	.006	−.003
A6:A5 plus equal factor means	494.4 (337)	.955	.955	.044 (.035; .052)	A6 *vs* A5	14.5 (3)	*.002*	−.005	.002
A6+: A6 except 1 factor mean unequal	**478.6 (336)**	**.960**	**.959**	**.042 (.033; .050)**	A6+ *vs* A5	3.84 (2)	.146	.005	−.002
Across sex at age 7 (*N* = 460)									
B0a: females (*n* = 234)	275.1 (149)	.888	.871	.060 (.049; .071)					
B0b: males (*n* = 226)	276.1 (149)	.920	.908	.061 (.050; .073)					
B1: configural (equal form)	551.2 (298)	.906	.893	.061 (.053; .069)					
B2: B1 plus equal factor loadings (weak invariance)	548.9 (314)	.913	.905	.057 (.049; .065)	B2 *vs* B1	20.3 (16)	.207	.007	−.004
B3: B2 plus equal thresholds (strong invariance)	569.4 (330)	.912	.908	.056 (.048; .064)	B3 *vs* B2	26.7 (16)	.045	−.001	−.001
B4: B3 plus uniquenesses free (strict invariance)[b]	565.6 (311)	.906	.896	.060 (.052; .067)	B3 *vs* B4[b]	21.8 (19)	.293	.006	−.004
B5: B4 plus equal factor variances and covariances	552.1 (336)	.920	.919	.053 (.045; .061)	B5 *vs* B3	4.24 (6)	.645	.008	−.003
B6:B5 plus equal factor means	**555.0 (339)**	**.920**	**.919**	**.053 (.045; .060)**	B6 *vs* B5	5.27 (3)	.153	.000	.000
Across sex at age 8 (*N* = 426)									
C0a: females (*n* = 217)	192.2 (149)	.952	.945	.037 (.019; .051)					
C0b: males (*n* = 209)	431.5 (149)	.848	.825	.095 (.085; .106)					
C1: configural (equal form)	622.4 (298)	.878	.860	.071 (.064; .079)					
C2: C1 plus equal factor loadings (weak invariance)	917.9 (314)	.773	.753	.095 (.088; .102)	C2 *vs* C1	185.1 (16)	*<.001*	*−.105*	*.024*
C2+: C2 except 1 factor loading unequal (partial)	619.5 (313)	.885	.874	.068 (.060; .076)	C2+ *vs* C1	30.5 (15)	.010	.007	−.003
C3: C2+ plus equal thresholds (strong invariance)	637.7 (329)	.884	.879	.066 (.059; .074)	C3 *vs* C2+	26.2 (16)	.052	−.001	−.002
C4: C3 plus uniquenesses free (strict invariance)[b]	628.1 (310)	.880	.868	.069 (.062; .077)	C3 *vs* C4[b]	39.9 (19)	*.003*	.004	−.003
C3+: C2 except 1 uniqueness unequal (partial)[b]	632.8 (328)	.885	.880	.066 (.037; .053)	C3+ *vs* C4[b]	34.8 (18)	.010	.005	−.003
C5: C4 plus equal factor variances and covariances	616.3 (334)	.894	.891	.063 (.055; .071)	C5 *vs* C3+	12.9 (6)	.045	.009	−.003
C6: C5 plus equal factor means	**619.1 (337)**	**.894**	**.892**	**.063 (.055; .070)**	C6 *vs* C5	6.17 (3)	.104	.000	.000
Across age, repeated measures (*N* = 521)									
D0c: age 6 (*n* = 481)	261.2 (149)	.968	.963	.040 (.032; .047)					
D0d: age 7 (*n* = 460)	335.7 (149)	.930	.920	.052 (.045; .060)					
D0e: age 8 (*n* = 426)	353.1 (149)	.919	.907	.057 (.049; .064)					
D1: configural (equal form)	2039.5 (1465)	.938	.932	.027 (.025; .030)					
D2: D1 plus equal factor loadings (weak invariance)	2050.6 (1497)	.940	.936	.027 (.024; .029)	D2 *vs* D1	53.4 (32)	.010	.002	.000
D3: D2 plus equal thresholds (strong invariance)	2084.0 (1529)	.940	.937	.026 (.023; .029)	D3 *vs* D2	38.1 (32)	.213	.000	−.001
D4: D3 plus uniquenesses free (strict invariance)[b]	2066.7 (1491)	.938	.933	.027 (.024; .030)	D3 *vs* D4[b]	59.2 (38)	.016	.002	−.001
D5: D4 plus equal factor variances and covariances	2084.8 (1541)	.941	.939	.026 (.023; .029)	D5 *vs* D3	13.6 (12)	.033	.001	.000
D6: D5 plus equal factor means	**2094.2 (1547)**	**.941**	**.939**	**.026 (.023; .029)**	D6 *vs* D5	16.2 (6)	.013	.000	.000

In bold: final model after invariance analysis; in italics: meaningful decrement in fit, based on $p < .01$, ΔCFI < −.010 or ΔRMSEA > .015

[a]$\Delta\chi^2$ based on DIFFTEST approach obtained from MPlus (scaled difference chi-square test for WLSMV method of estimation)

[b]Test for invariance of uniquenesses for categorical indicators proceeds backwards: uniquenesses are first freely estimated in the second group (Model #4), and are then compared to the model in which all uniquenesses are fixed at 1 in the second group so as to be in line with the first group (Model #3), because when a factor loading and an item threshold for a categorical factor indicator are free across groups, the uniqueness for the variable must be fixed at 1 for identification purposes (Muthén & Muthén, 1998-2013; e.g., http://www.lesahoffman.com/948/948_Lecture9_Invariance.pdf).

Table 2: Means, *SD* and ANOVA results for comparison of BPM-P direct scores between girls and boys at age 6, 7, and 8

		Age 6		Age 7		Age 8		ANOVA: *F* (*p*)		
BPM-P scale (min-max)	Sex	*M* (*SD*)	*d*	*M* (*SD*)	*d*	*M* (*SD*)	*d*	Sex × Age	Age	Sex
Attention (0-12)	Girls	2.24 (2.53)	0.31	2.36 (2.46)	0.22	2.24 (2.34)	0.29			
	Boys	3.08 (2.80)		2.95 (2.92)		3.02 (2.99)				
	Total	2.65 (2.70)		2.66 (2.71)		2.64 (2.71)		0.39 (.535)	0.35 (.554)	5.31 (.022)
Externalizing (0-14)	Girls	2.16 (2.07)	0.07	1.88 (1.97)	0.09	1.92 (2.19)	0.08			
	Boys	2.31 (2.12)		2.07 (2.10)		2.08 (2.03)				
	Total	2.24 (2.10)		1.98 (2.04)		2.00 (2.11)		1.14 (.287)	6.08 (.014)	0.01 (.938)
Internalizing (0-12)	Girls	0.96 (1.34)	0.14	1.18 (1.36)	−0.03	1.22 (1.38)	0.03			
	Boys	1.16 (1.51)		1.13 (1.54)		1.27 (1.70)				
	Total	1.06 (1.43)		1.16 (1.45)		1.25 (1.55)		2.37 (.125)	5.96 (.015)	0.02 (.878)
Total (0-38)	Girls	5.36 (4.74)	0.24	5.43 (4.56)	0.15	5.39 (4.70)	0.19			
	Boys	6.57 (5.10)		6.15 (5.10)		6.37 (5.38)				
	Total	5.95 (4.95)		5.79 (4.85)		5.88 (5.08)		0.61 (.437)	0.18 (.671)	1.73 (.189)

Note: *d* means Cohen's effect size

Table 3: Correlation coefficients between Brief Problem Monitor-Parent version (BPM-P) and Child Behavior Checklist (CBCL/6-18; top) and Strength and Difficulties Questionnaire-Parent version (SDQ-P; bottom) scores at the same age

	BPM-P Attention	BPM-P Externalizing	BPM-P Internalizing	BPM-P Total problems
Concurrent validity (at age 6 /age 7/age 8)				
CBCL/6-18 Attention problems [.83]	**.96/.97/.96**	.52/.44/.46	.35/.38/.45	.84/.84/.84
CBCL/6-18 Aggressive behaviour [.86]	.61/.54/.55	**.92/.93/.93**	.48/.43/.47	.86/.82/.83
CBCL/6-18 Anxious/Depressed [.72]	.42/.38/.45	.50/.47/.42	**.90/.88/.91**	.70/.67/.69
CBCL/6-18 Externalizing problems [.88]	.63/.57/.57	**.90/.91/.91**	.48/.43/.48	.86/.83/.83
CBCL/6-18 Internalizing problems [.80]	.40/.35/.45	.47/.43/.40	**.84/.84/.85**	.66/.63/.67
CBCL/6-18 Total problems [.92]	.73/.71/.74	.76/.73/.73	.67/.69/.71	**.91/.91/.91**
Relation to the remaining scale scores				
CBCL/6-18 Withdrawn/Depressed [.58]	.23/.20/.28	.32/.18/.26	.59/.59/.56	.43/.36/.43
CBCL/6-18 Somatic complaints [.54]	.26/.18/.25	.26/.28/.16	.43/.38/.34	.37/.33/.31
CBCL/6-18 Social problems [.65]	.52/.48/.53	.53/.47/.52	.58/.61/.57	.67/.64/.67
CBCL/6-18 Thought problems [.49]	.33/.32/.37	.36/.30/.29	.47/.45/.46	.47/.44/.46
CBCL/6-18 Rule-breaking behaviour [.57]	.55/.50/.49	.64/.64/.63	.37/.34/.39	.68/.65/.65
Convergent and divergent validity (at age 7)				
SDQ-P Hyperactivity [.81]	**.84**	.37	.24	.69
SDQ-P Conduct [.48]	.39	**.67**	.27	.58
SDQ-P Emotional [.63]	.27	.29	**.57**	.44
SDQ-P Peers [.51]	.37	.30	.39	.45
SDQ-P Prosocial [.65]	.14	.25	.12	.22
SDQ-P Impact [.75]	.43	.42	.35	**.52**
SDQ-P Externalizing [conduct + hyperactivity] [.77]	**.80**	**.54**	.29	.76
SDQ-P Internalizing [emotional + peer] [.70]	.37	.35	**.57**	.52
SDQ-P Total difficulties [.81]	.73	.54	.47	**.78**

In bold: Correlation coefficients between more expected related scale scores, assessing concurrent (top) and convergent (bottom) validity.

In brackets: Median for Cronbach's alpha values in the sample of the study.

Table 4: Correlation coefficients between Child Behavior Checklist (CBCL/6-18) scores as criteria at ages 7 (top) and 8 (bottom) and either Brief Problem Monitor-Parent version (BPM-P) and CBCL/6-18 data as predictors at younger ages

	Predictor at age 6			Predictor at age 7		
	BPM-P	CBCL/6-18	*p*	BPM-P	CBCL/6-18	*p*
Criterion at age 7						
CBCL/6-18 Attention problems	.69	.72	.410			
CBCL/6-18 Externalizing problems	.67	.72	.140			
CBCL/6-18 Internalizing problems	.56	.63	.134			
CBCL/6-18 Total problems	.70	.73	.276			
Criterion at age 8						
CBCL/6-18 Attention problems	.66	.68	.674	.77	.79	.385
CBCL/6-18 Externalizing problems	.65	.69	.433	.74	.77	.431
CBCL/6-18 Internalizing problems	.49	.58	.106	.53	.66	.005
CBCL/6-18 Total problems	.68	.70	.566	.73	.78	.051

Note: *p* values based on test of difference between two dependent correlations

Table 5: Association between BPM-P scores and the use of services and functional impairment at age 7 (n = 460)

Criterion	BPM-P scale score	OR (95% CI)	p
Use of services (*yes*)	Attention	1.22 (1.09; 1.37)	.001
	Externalizing	1.17 (1.01; 1.36)	.036
	Internalizing	1.32 (1.09; 1.60)	.004
	Total problems	1.12 (1.05; 1.19)	< .001
Criterion	BPM-P scale score	*B* (95% CI)	p
Impairment (CGAS scores)	Attention	−1.99 (−2.27; −1.71)	< .001
	Externalizing	−2.38 (−2.76; −2.00)	<. 001
	Internalizing	−2.36 (−2.94; −1.78)	< .001
	Total problems	−1.25 (−1.39; −1.10)	< .001

Table 6: Predictive accuracy of BPM-P scores at age 7 for DSM-5 disorders at age 7 (n = 460) and age 8 (n = 426)

	Criterion	Prevalence [a]	Predictor at age 7	OR (95% CI) [b]	AUC (95% CI) [b]
DSM-5 disorder at age 7	ADHD	9.8	BPM-P Attention	2.21 (1.82; 2.69) ***	.93 (.90; .96)
	ODD-CD	6.7	BPM-P Externalizing	2.01 (1.65: 2.46) ***	.87 (.82; .93)
	Depression-anxiety-phobia	7.7	BPM-P Internalizing	1.54 (1.28; 1.86) ***	.71 (.62; .80)
	Any disorder	20.1	BPM-P Total problems	1.25 (1.19; 1.33) ***	.78 (.73; .83)
DSM-5 disorder at age 8	ADHD	10.9	BPM-P Attention	1.79 (1.55; 2.06) ***	.88 (.82; .93)
	ODD-CD	6.1	BPM-P Externalizing	1.95 (1.60; 2.38) ***	.90 (.85; .94)
	Depression-anxiety-phobia	9.6	BPM-P Internalizing	1.30 (1.09; 1.56) ***	.64 (.55; .73)
	Any disorder	22.2	BPM-P Total problems	1.24 (1.17; 1.30) ***	.77 (.72; .82)

* p <.05; ** p < .01; *** p < .005

[a] Weighted prevalence for the DSM-5 disorder/s (%)

[b] OR (Odds Ratio) and AUC (Area Under ROC Curve) values obtained through logistic regression adjusted by sex.

ADHD: attention-deficit/hyperactivity disorder; ODD: oppositional defiant disorder; CD: conduct disorder; Any disorder: ADHD, ODD, CD, depression, and/or anxiety/phobia