PSYCHOMETRIC PROPERTIES OF THE FRENCH VERSION OF THE BOXALL PROFILE

Jean-Yves Bégin, Luc Touchette, Caroline Couture, Cassandre Blais

Université du Québec à Trois-Rivières, Département de psychoéducation, 850 Avenue de Vimy, C.P. 32; Québec, Québec; G1S 0B7
Corresponding author: Jean-Yves Bégin, jean-yves.begin@uqtr.ca
Keywords: Boxall Profile, French version, psychometric properties, validity, ROC curves.

ABSTRACT

The Boxall Profile provides a framework for the structured observation of children in nurture groups. It is a detailed and rigorously trialled normative diagnostic instrument developed for teachers and teaching assistants to measure children's levels of emotional and behavioural functioning. Moreover, it highlights specific targets for intervention within a child's individual functioning. As of yet, the psychometric properties of the French version of Boxall Profile have not been extensively studied. In total 169 boys and 23 girls (N = 192) ranging from 6 to 13 years old (M = 9.24 years old, SD = 1.83) were assessed by their teacher using the French version of the Boxall Profile, the Strengths and Difficulties Questionnaire, and the Teacher Report Form. Internal consistency analysis (Cronbach's Alpha), correlation analysis, exploratory factor analysis, and area under the receiver operating characteristic curve (AUC) were performed. The results demonstrate good reliability of scales and sub-scales, the fit of the first level of factorial structure, and good concurrent validity. These results suggest the effectiveness of the French version of the Boxall Profile in properly identifying students facing difficulties. Possible solutions are discussed to improve the construct validity of the second and third tiers of the instrument.

INTRODUCTION

During an extensive longitudinal study (across 12 years 1998-2010) on child development (N = 2,120) in Quebec (Canada), it was found that almost half of the participating children had a high incidence of social emotional behaviour disorders symptoms (SEBD) during at least two of the eight data collection time intervals (from the 17th month through to the 10-year point). More specifically, around 25% exhibited a high level of internalising behaviours, 37% exhibited a high level of externalising behaviours, and 7% exhibited a high level of interpersonal difficulties (Riberdy et al, 2013). For children aged between 5 and 10 years, close to 19% of study participants received at least one formal neurodevelopmental disorder diagnosis during the study with the most prevalent being learning disorders (10%) and attention deficit/hyperactivity disorder (12%). All reported that social, emotional, behavioural, and learning difficulties were more prevalent in children from families with a history of sustained low socio-economic status (Riberdy et al, 2013).

Throughout Quebec primary schools, various special education programmes have been established to prevent SEBDs from worsening and to foster improvements. Inspired by Marjorie Boxall's UK-based nurture groups (NG), 'Kangaroo Class' (KC) is one such programme that has been operating in a considerable number of Quebec primary schools for over 10 years now. KC programmes cater to children who in their early years at school have not acquired the necessary social and/or behavioural maturity levels to attend mainstream classrooms and offer them an alternative setting which has been adapted to fit their developmental needs. Research has shown that NGs have a positive effect on the development of children in many areas and consistently shows that children who attend NGs make significant social and emotional gains after attending the groups (Sanders, 2007; Seth-Smith et al, 2010; Shaver & Mcclatchey, 2013). Furthermore, findings from a study conducted by Cooper and Whitebread (2007) simlarly indicate significant improvements for nurture group students in terms of social, emotional and behavioural development and noted that in schools providing NG programmes, students with SEBD who attend mainstream classrooms throughout the year

Accepted on **2 March 2020;** published on **23 October 2020**
Citation: Bégin, YV., Touchette, L., Couture, C., Blais, C. (2020) Psychometric properties of the French version of the Boxall Profile. *International Journal of Nurture in Education,* 6(1) 48–58. **Authors' note:** The authors certify that they have no existing or potential conflicts of interest.

had also significantly improved in terms of behaviour – more so than students both with and without SEBD attending schools that did not operate nurture group programmes. Quebec studies that sought to appraise the implementation and success of KC programmes reported similar findings. Research on the impact of KC programmes indicates that both KC teachers and the parents of KC students expressed high rates of satisfaction and had very positive perceptions of the level of SEBD gains, with over 80% reporting postive or very positive effects on the following criteria: students' behaviour in school, self-esteem, and general attitudes towards school and adults (Couture, 2009; Couture & Bégin 2010). Quantitatively, the same research data shows more noteworthy behavioural improvements in KC students on some particular criteria than for students attending other programme types (Couture & Bégin 2010; Couture & Lapalme 2007).

The original NG model advocates the use of an instrument named the Boxall Profile (BP; Bennathan & Boxall, 1998), a French version of which has since been translated for use within Quebec's KC programmes. This questionnaire was designed for teachers and teaching assistants to thoroughly assess a student's strengths and difficulties with an aim to design effective intervention plans for their specific and unique needs. The systematic use of the BP is scheduled across time intervals for each student to properly track student evolution and progress in order to make further adjustments to individual intervention plans as needed (Cooper & Whitebread, 2007).

Given the current lack of scientific validation for the translated French version of the BP, this study aimed to identify its psychometric qualities and limitations by applying a range of statistical analysis methods.

Boxall Profile

The BP comes in questionnaire form and is filled out by a school teacher or staff member who knows the student. Its creators support its use with children from ages 3 to 12, although it is only normalised for ages 3 to 8. Figure 1 shows how the BP is structurally broken down. Divided into two core sections, the first section is called Developmental Strands and deals with developmental factors underpinning the individual's ability to engage effectively in the learning process. This section is then divided into two parent scales: Organisation of Experience and Internalisation of Controls. Each of these scales comprises five subscales that reflect the child's level of engagement with the world as well as his or her level of personal development, and his or her awareness of others. Each subscale contains between two and five items each for a total of 34 items across the whole section.

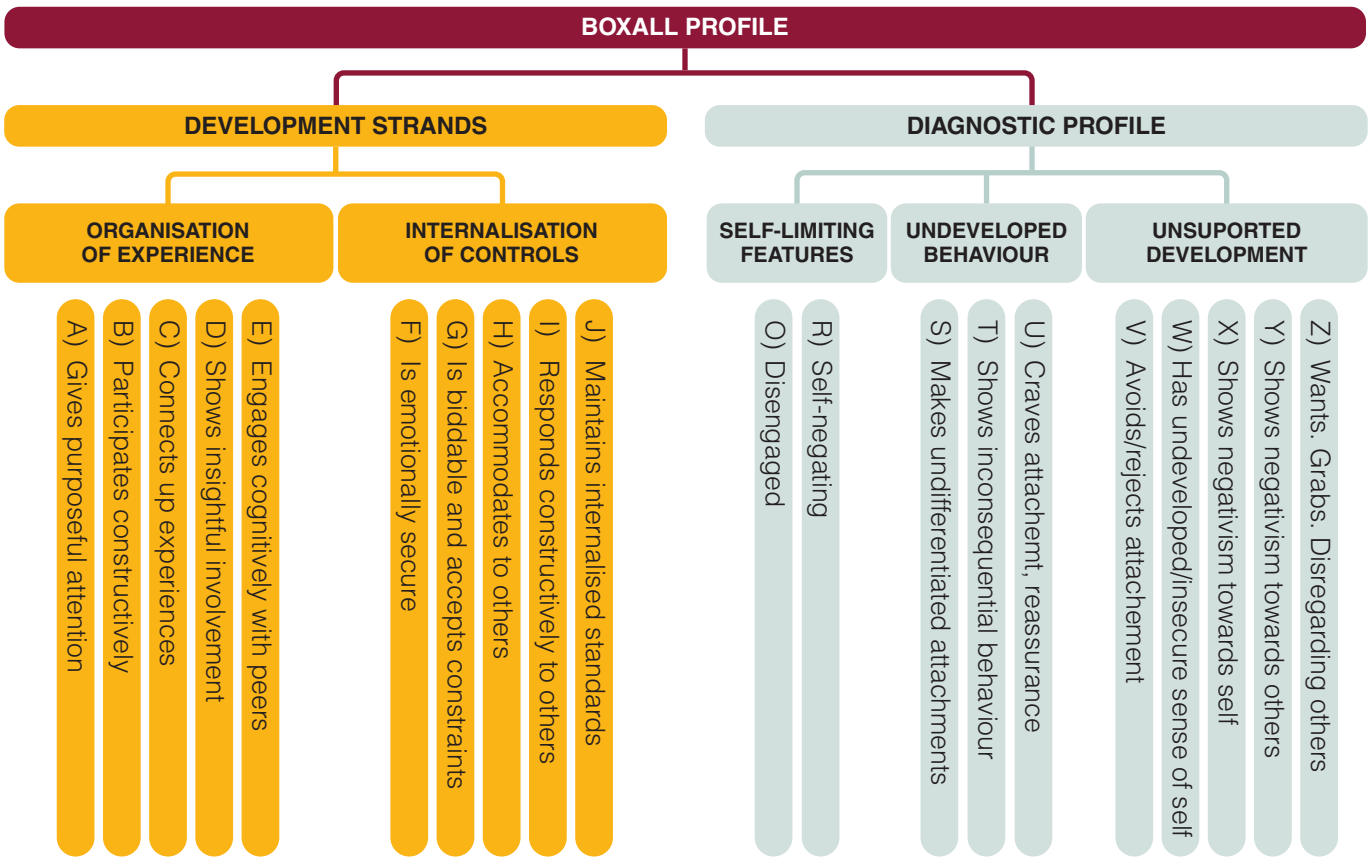

Figure 1: *Layout of the Boxall Profile structure. The first tier is comprised of two core sections. The second tier is broken down into five scales. The third tier involves 20 subscales that then break down to a total of 68 final items.*

The second section is called Diagnostic Profile and deals with any behavioural characteristics that may inhibit or interfere with the child's social and academic performance. This section is divided into three parent scales: Self-Limiting Features (with two subscales); Undeveloped Behaviour (three subscales); and Unsupported Development (five subscales). These subscales respectively reflect a) a lack of a normal thrust for growth, b) a lack of inner resources to relate to others and engage at an age-appropriate level, and c) a lack of early nurturing care. As with the first section, the second one comprises 34 items, albeit split up across 10 subscales. This study used the standardised version of the norms calculated in 1984 that evaluated 880 students in the United Kingdom.

Psychometric properties of the Boxall Profile

Bennathan and Boxall (1998) presented findings from a validation study conducted on the original instrument. Content validity was established by pooling the observations made by experts working with children in the context of NGs and mainstream classroom settings, as well as by one psychotherapist. The BP's items were defined as to represent children with developmental delays whose dysfunctional living context may contribute to emotional immaturity. Both the theoretical foundation of the BP and the NG intervention philosophy are rooted in attachment theory (Bennathan & Boxall, 2000). Construct validity was assessed using a sample of 880 children aged from 3 years 4 months through to 8 years of age. Specifically, 442 were in primary school NGs: 307 in mainstream primary classes, and 131 in mainstream nursery classes. The BP's sections, scales and subscales were initially created by grouping items using factor analysis. For both of the main sections, the subscales' clinical thresholds were derived from the scores of children whose average age was five, and who the teachers deemed to be developing typically and functioning well.

A study led by Couture, Cooper, and Royer (2011) assessed the concurrent validity of the BP using data previously collected by Cooper and Whitebread (2007). The sample consisted of 202 children and adolescents (70.3% boys, 29.7% girls) attending NGs at 25 schools spread across eight Local Education Authorities throughout the United Kingdom. Children were aged from 3 years 11 months to 14 years 3 months (M = 6.61, SD = 1.90), with 87.6% of the sample falling between 3 to 8 years of age, the range for which the BP has been normalised. The internal consistency of the five BP subscales was demonstrated using Cronbach's alpha, which varied between .24 and .87. To establish the BP's ability to differentiate children with difficulties from those without, the researchers split children into two comparison groups – normal range (N = 14) and abnormal range (N = 170) – based on their scores on the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997). Independent t-test results revealed that, using the SDQ as a baseline, four out of the five BP scales showed appropriate and significant differentiation of children in the normal and abnormal ranges. Only the Undeveloped Behaviour scale did not differentiate sufficiently both groups. To ascertain the level of construct convergence and divergence, Pearson's correlations were conducted between the six SDQ scales and the five BP scales. The results showed that all of the BP scales significantly correlate with at least three of the SDQ scales (r = -.45 to .58, p < .05) and all BP scales were significantly correlated with the SDQ total difficulties score (r = -.43 to .36, p < .001). In sum, findings indicated that both instruments measure reasonably comparable constructs in children with behavioural difficulties, even though each instrument has a different scoring approach.

All in all, the information reported by Bennathan and Boxall (1998) allows for the assessment of the BP's initial design process and content validity. On the other hand, the psychometric properties, along with the detail on the statistical analyses conducted, the results, and the conditions of data collection were either presented cursorily or altogether absent. Similarly, the process of validation and normalisation of clinical thresholds was only briefly presented, and the sample's characteristics, such as children's ages, were not representative of Quebec's primary school setting. Couture et al (2011) study addressed these limitations in part, but it also contained gaps, such as solely conducting a partial evaluation of the BP's psychometric properties and using unequal child comparison group sizes.

OBJECTIVES

The pertinence of this study rests primarily on the fact that there had not yet been any validation work done on the French version of the BP. Given that KCs have garnered significant interest in Quebec (Canada) and require the systematic use of the BP, this research aims to: (a) study the questionnaire's reliability and analyze the internal consistency of the scales and subscales; (b) study the construct convergence and divergence; (c) study the questionnaire structure using factor analysis, and; (d) study the concurrent validity as well as evaluate its diagnostic performance.

METHOD

Participants

Table 1 shows the demographic breakdown of the study's sample group which consisted of 169 boys and 23 girls (N = 192) between 6 and 13 years of age (M = 9.24 years of age, SD = 1.83). Participants were drawn from mainstream classrooms (N = 44), KC programmes (N = 94), and special-education type classes catering to children with behavioural disorders (N = 54).

Table 1: *Sample Group Demographic Breakdown (n = 192).*

	Mainstream classroom (*N*=44)	**Kangaroo class** (*N*=94)	**Special education type class** (*N*=54)
Boys	39	79	51
Girls	5	15	3
M age (*SD*)	8.13 (1.33)	9.35 (1.82)	9.85 (1.83)

Participants

Both the KC-programme children and the special-education-class children included in the sample were drawn from a three-year-long research project (2005 to 2008) financed by Quebec's *Ministère de l'Éducation du Loisir et du Sport du Québec* (MELS). The MELS-financed study sought to devise an adapted version of NG programmes to be more viable within the Quebec school setting, while concurrently assessing the effectiveness of KC programmes already underway in Quebec (Couture & Bégin, 2010; Couture & Lapalme, 2007). The procedure was approved by the education and social sciences ethics comittee at the Université de Sherbrooke. The research was conducted in six schools spanning five Quebec school districts. To accurately represent the population attending these specialised programmes, no specific inclusion or exclusion criteria were set. KC programme and special-education teachers were recruited on a voluntary basis and gave their free and prior informed consent to partake in the study. Each child's parent also formally consented. According to the MELS (2008), the average socioeconomic disadvantage index for the participating schools was situated in the 8th decile which positions the schools well below average disadvantage levels for Quebec. The average student body size per school was 383.3 (*SD* = 152.95, *MN* = 146, *max* = 612).

Children from mainstream classrooms were selected from a single Quebec school district out of three schools whose student body size and socioeconomic disadvantage index were somewhat consistent with the six schools participating in the MELS study. In all, nine teachers were chosen across the three schools and spanning a range of class levels (Grades 1-6). Each of the nine teachers then randomly selected five students from their class list to undergo a series of assessments, after first having eliminated any students they had known for less than two months or who may have shown signs of developmental delay, intellectual disability, other cognitive impairments or SEBD.

Measurement

Teachers administered three questionnaires per participant, those being the French version of the BP (Bennathan & Boxall, 1998), the French version of the *Teacher Report Form* (TRF; Achenbach & Rescorla, 2001) and the French version of the *Strengths and Difficulties Questionnaire* (SDQ; Goodman, 1997). The TRF and SDQ were completed by teachers as a point of comparison to establish a gold standard to assess the concurrent validity of the BP. These two questionnaires (TRF and SDQ) were chosen for their proven psychometric properties.

The French version of the Boxall Profile

The initial French version of the BP came into circulation in 2004 via a two-step translation process. The BP and accompanying teacher handbook were first translated (Bennathan & Boxall, 1998) by a bilingual Master's level (MSc) psychoeducator with several years work experience in the education field. The first translation was then submitted for review by a bilingual Master's level (MSc) psychologist and a bilingual doctoral-level (PhD) psychoeducator.

Teacher Report Form. Comprising 113 items, the TRF assesses behavioural difficulties via a range of specific subscales (Anxious/depressed, Withdrawn/depressed, Somatic complaints, Social problems, Thought problems, Attention problems, Rule breaking behavior, Aggressive behavior) as well as more general scales (internalizing/externalizing behavioural issues) culminating in a Total Score that reveals the overall severity of the behavioural difficulties experienced. Translated into multiple languages and used by over 80 different societies and cultural groups, this questionnaire has been scientifically considered psychometrically sound. Among others, the study carried out by Ivanova et al (2007) examined the use of the TRF in 20 societies, using confirmatory factor analysis to find that the items and constructs within the questionnaire had high cross-cultural likeness. As for reliability, Rescorla et al's (2007) study on the TRF's internal consistency when used across 21 countries reports Cronbach's alpha coefficients ranging from .64 to .96 (*M* = .82) across all scales. Average alpha coefficients for the three general scales (Total problems, Internalising problems and Externalising problems) sit respectively at .96, .82 and .92 (Rescorla et al, 2007).

Teacher Version of the Strengths and Difficulties Questionnaire

This questionnaire involves the assessment of 25 items across five scales: 1) Emotional symptoms, 2) Conduct problems, 3) Hyperactivity/inattention, 4) Peer relationship problems, and 5) Prosocial behaviour. The sum of all four *difficulties* scales (aforementioned scales 1 to 4 only) generates a Total Difficulties Score. The SDQ has been translated into over 60 languages. Moreover, it has been both extensively validated worldwide and put to extensive use in international epidemiological studies to assess childhood mental health. Shojaei et al (2008) stated that its psychometric properties have been evaluated in over 20 distinct studies. Furthermore, Capron et al (2007) reported

that the French version of the SDQ's scales (teacher version) showed good internal consistency. The Cronbach's alpha coefficients ranged from .64 to .87 for the scales overall. With respect to temporal stability, over a six-week interval the correlation coefficient for the total of the *difficulties* scales sat at .88 whereas correlations varied between .63 and .89 on its other scales (Capron et al, 2007).

Table 2: *Reliability of the Boxall Profile Scales and Subscales (N = 192).*

SCALES AND SUB-SCALES	CRONBACH'S ALPHAS
DEVELOPMENT STRANDS	
Organisation of experience	**.91**
a) Gives purposeful attention	.81
b) Participates constructively	.68
c) Connects up experiences	.76
d) Shows insightful involvement	.70
e) Engages cognitively with peers	.77
Internalisation of controls	**.89**
f) Is emotionally secure	.69
g) Is biddable and accepts constraints	.75
h) Accommodates others	.85
i) Responds constructively to others	.76
j) Maintains internalised standards	.62
DIAGNOSTIC PROFILE	
Self-limiting features	**.65**
q) Disengaged	.66
r) Self-negating	.75
Undeveloped behaviour	**.80**
s) Makes undifferentiated attachments	.62
t) Shows inconsequential behaviour	.82
u) Craves attachment, reassurance	.77
Unsupported development	**.90**
v) Avoids/rejects attachment	.71
w) Has undeveloped/insecure sense of self	.77
x) Shows negativism towards self	.73
y) Shows negativism towards others	.87
z) Wants, grabs, disregarding others	.78

Internal validity

Internal construct validity was assessed using cross-correlation matrices, first comparing the BP scales among each other and then subsequently comparing the scales to the subscales. Table 3 presents the BP interscale correlation results, showing the degree of convergence and divergence among the scale's constructs.

The two positive scales within the Developmental Strands – those being Organisation of Experience and Internalisation of Controls – showed a positive correlation with a coefficient of r = .83 (p < .001). Similarly, the Diagnostic Profile's negative scales – those being Self-Limiting Features, Undeveloped Behaviour and Unsupported Development – also indicated there was a positive correlation between the three of them. These correlation coefficients ranged from r = .75 to r = .84 (p < .001) suggesting that these scales measure essentially comparable constructs. In contrast, constructs assessed on the Diagnostic Profile scale indicated there was a negative correlation with those assessed on the Developmental Strands scale. These correlation coefficients ranged from r = -.43 to r = -.69 (p < .001) suggesting that the constructs differ.

To study relationships between the various scales and subscales across the two sections of the questionnaire, a second correlation matrix was conducted. All of the Developmental Strands subscales indicated that there was a positive correlation with this section's two scales, ranging from r = .59 to r = .94 (M = .80, p < .001). Correlation coefficients were higher between subscales when within their parent scale. In point of fact, coefficients between subscales 'a)' and 'e)' (see Table 2 for the complete titles of all subscales) and the Organisation of Experience scale ranged from r = .84

Table 3: *Interscale Correlation for the Boxall Profile (N = 192).*

	BOXALL PROFILE SCALES				
BOXALL PROFILE SCALES	**Organisation of experience**	**Internalisation of controls**	**Self-limiting features**	**Undeveloped behaviour**	**Unsupported development**
Organisation of experience	–	.83**	-.58**	-.50**	-.43**
Internalisation of controls		–	-.64**	-.68**	-.69**
Self-limiting features			–	.75**	.77**
Undeveloped behaviour				–	.84**
Unsupported development					–

*Note. **p < .01*

to $r = .92$ ($M = .88$, $p < .001$) while coefficients between subscales 'f)' and 'j)' and the Internalisation of Controls scale ranged from $r = .80$ to $r = .94$ ($M = .88$, $p < .001$). As expected, all Developmental Strands subscales indicated there was a negative correlation with the Diagnostic Profile scales, with correlation coefficients ranging from $r = -.20$ to $r = -.72$ ($M = -.51$, $p < .001$). All Diagnostic Profile subscales indicated there was a negative correlation to the Developmental Strands scales, with those correlations ranging from $r = -.31$ to $r = -.71$ ($M = .50$, $p < .001$). All Diagnostic Profile subscales indicated there was a positive correlation with the section's three scales, ranging from $r = .47$ to $r = .93$ ($M = .74$, $p < .001$). As noted previously for the Developmental Strands, correlation coefficients that linked the Diagnostic Profile subscales and their respective parent scales were also somewhat higher than out-of-scale. Indeed, the relationship between subscales "q)" and "r)" and the Self-Limiting Features parent scale were $r = .83$ and $r = .89$ ($p < .001$) respectively. Coefficients that linked subscales 's)' through 'u)' and their Undeveloped Behaviour parent scale ranged from $r = .80$ to $r = .93$ ($M = .86$, $p < .001$). Lastly, subscales 'v)' through 'z)' had relationships to their Unsupported Development parent scale showing correlation coefficients ranging from $r = .76$ to $r = .92$ ($M = .85$, $p < .001$).

External validity

Table 4 presents the correlation matrix comparing BP scales to SDQ scales. This helps us to consider the level of convergence or divergence among constructs measured by the BP in light of similar or opposing constructs measured by other proven, psychometrically validated instruments. Results showed a negative correlation between the two positive BP scales in the Developmental Strands section (Organisation of experience and Internalisation of controls) – which both measure child competency – and the negative SDQ scales which measure constructs related to behavioural problems. The coefficients ranged from $r = -.31$ to $r = -.70$ ($X = -.52$, $p < .001$). The negative relationship proved even stronger with the SDQ total difficulties score showing respectively $r = -.70$ ($p < .001$) and $r = -.77$ ($p < .001$). Beyond this, there was a positive correlation between the Developmental Strands' two positive scales and the SDQ's positive scale (both measuring social skills). Correlations with the SDQ's Prosocial Behaviour scale were $r = .63$ ($p < .001$) for the Organisation of Experience scale and $r = .71$ ($p < .001$) for the Internalisation of Controls scale. The inverse was true for the Diagnostic Profile's negative scales (Self-limiting features, Undeveloped behaviour and Unsupported development). In addition, the latter all showed negative correlations with the SDQ's Prosocial Behaviour scale.

In addition, there was a positive correlational relationship between the Diagnostic Profile's negative scales and the SDQ's negative scales which all measure behavioural challenges. These correlation coefficients ranged from $r = .46$ to $r = .70$ ($X = .56$, $p < .001$). As per the Developmental Strands scales, correlation coefficients between the Diagnostic Profile's negative scales and the SDQ's Total Difficulties Score were substantially higher. The positive correlations underscore the similarity of constructs and ranged from $r = .74$ to $r = .80$ ($X = .77$, $p < .001$).

Table 4: *Correlational relationships between the Boxall Profile scales and the Strengths and Difficulties questionnaire scales (N = 190)*

	STRENGTHS AND DIFFICULTIES QUESTIONNAIRE SCALES					
BOXALL PROFILE SCALES	**Emotional symptoms**	**Conduct problems**	**Hyperactivity inattention**	**Peer problems**	**Prosocial behaviour**	**Total difficulties score**
Organisation of experience	-.35**	-.48**	-.55**	-.56**	.63**	**-.70****
Internalisation of controls	-.31**	-.70**	-.67**	-.57**	-.71**	**-.77****
Self-limiting features	.60**	.50**	.54**	.49**	-.38**	**.74****
Undeveloped behaviour	.48**	.70**	.68**	.53**	-.41**	**.80****
Unsupported development	.46**	.70**	.53**	.49**	-.20**	**.76****

*Note. **p < .01*

Factor analysis

Subsequently, we conducted exploratory factor analysis (EFA) to explore relationships between the various constructs being measured by the BP, in light of the questionnaire's existing structural breakdown. EFA also helped in establishing whether it was possible to group certain constructs differently, potentially with constructs currently undevelopped. Principal component analysis was used as the factor extraction method. Axes were repositioned with the direct oblimin rotation technique and with an understanding that the extracted factors might be cross-correlated (Hair et al, 2010). The choice to rotate the axes is supported by the interscale correlations presented in the internal validity section of this study. To further explore the existing questionnaire's breakdown we extracted factors from the BP's 20 sub-scales. This was done to achieve stable estimates and correlations with a ratio of 10 participants per studied variable (Hair et al, 2010). The BP's EFA (KMO = .938; Bartlett's test of sphericity X^2 [190] = 3553.2, p < 0,001) pointed to the presence of two main factors; the first of which accounts for 57.8% (eigenvalue = 11.4) of the variance and the second of which accounts for 14.6% (eigenvalue = 2.8). Those two factors were negatively correlated (-.508). A second EFA conducted on the Developmental Strands' ten respective sub-scales (KMO = .930; Bartlett's test of sphericity X^2 [45] = 1826.9, p < 0,001) established the presence of a sole factor accounting for 71.3% (eigenvalue = 7.2) of the variance. Similarly, EFA conducted on the Diagnostic Profile (KMO = .916; Bartlett's test of sphericity X^2 [45] = 1440.2, p < 0,001) demonstrated the presence of a sole factor accounting for 65.5% (eigenvalue = 6.6) of the variance.

Concurrent validity and diagnostic performance

An area under a (AUC) Receiver Operating Characteristic (ROC) analysis (Hanley & McNeil, 1982) was performed for each of the BP's scales to study the instrument's concurrent validity. The ROC curve provided us with a graphic representation of the current relationship between the sensitivity and the specificity of the instrument to determine its diagnostic performance and predictive validity. Thus, the ROC curve analysis conducted on each of the BP scales was used to determine whether the questionnaire is in fact able to accurately identify children who present significant behavioural difficulties as well as those who do not. For comparison purposes, the predictive capabilities of each of the BP scales was measured against those of the SDQ using the method proposed by DeLong et al (1988).

Gold standard criteria were set in order to guide these analyses and to form two comparison groups using children's TRF Total Difficulties scale results. Children who scored above the 90th percentile on this scale received a positive diagnosis, accounting for age and sex. In the context of this study, the TRF's Total Difficulties scale had an internal consistency of alpha .87. In total, 99 children received a positive diagnosis for SEBD and 86 received a negative one. TRF data was incomplete for seven of the total sample of 192 children. The AUC for the various BP scales and SDQ scales are presented in Table 5.

The AUC estimates predictive ability, wherein an AUC of .50 represents an instrument's arbitrary predictive capability in accordance with the laws of probability (50% chance of a correct diagnosis). Therefore, as a scale approaches 1, its predictive validity is

Table 5: *Area under the Receiver Operating Characteristic Curves for the Boxall Profile and the Strengths and Difficulties questionnaire scales*

STRENGTHS AND DIFFICULTIES QUESTIONNAIRE		BOXALL PROFILE	
Scales	**Area**	**Scales**	**Area**
Emotional symptoms	.75*	Organisation of experience	.82*
Conduct problems	.87*	Internalisation of controls	.89*
Hyperactivity/ inattention	.85*	Self-limiting features	.90*
Peer relationship problems	.83*	Undeveloped behaviour	.96*
Prosocial behaviour	.76*	Unsupported development	.94*
Total difficulties score	.94*		

*Note. N = 185. Positive diagnosis: n = 99; Negative diagnosis: n = 86. *p < .001*

deemed better. For reference purposes, tests were distinguished as follows: zero contribution (AUC = .50); slightly informative (.50 < AUC < .70); moderately informative (.70 < AUC < .90), very informative (.90 < AUC < 1), and perfect (AUC = 1) (Delacour et al, 2005). For example, a test with an AUC of .80 means that subjects with pathology would be 80% more likely to receive positive test results as compared to their counterparts without pathology. The AUC for the SDQ scales ranged from .75 to .87 (X = .81, p < .001) while the area for the Total Difficulties scale alone was .94 (p < .001). The AUC for the Organisation of Experience scale was .82 (p < .001) while for the Internalisation of Controls scale it was .89 (p < .001). Lastly, the area for the various Diagnostic Profile scales ranged from .90 to .96 (X = .93, p < .001).

Comparative findings for the AUC of the two questionnaires' various scales are presented in Table 6. Scales that measured relatively similar constructs were compared. In comparing the AUC for the BP's two positive behavioural scales with the SDQ's positive Prosocial Behaviour scale, no significant difference was found between the latter and the BP's Organisation of Experience scale. Conversely, the BP's Internalisation of Controls scale predicted behavioural difficulties in children significantly better than the SDQ's Prosocial Behaviour scale X^2 (1, N = 185) = 22.02, p < .001.

By comparing the AUC of each questionnaire's negative scales, referring to behavioural difficulties, the BP's Diagnostic Profile scales were shown to have overall better predictive performance than the SDQ scales. At the same time, no significant difference was found between the BP's Diagnostic Profile scales and the SDQ's Total Difficulties score. The same phenomenon was observed between the BP's Self-Limiting Features scale (AUC = .90) X^2 (1, N = 185) = 4.68, p < .05, its Undeveloped Behaviour scale (AUC = .96) X^2 (1, N = 185) = 15.35, p < .001, and its Unsupported Development scale (AUC = .94) X^2(1, N = 185) = 10.90, p < .001, which had a significantly better predictive validity than the SDQ's Peer Relationship Problems scale (AUC = .83). Additionally, the BP's Undeveloped Behaviour scale (AUC = .90) X^2(1, N = 185) = 13.30, p < .001 and Unsupported Development scale also had a significantly better predictive performance than the SDQ's Conduct Problems scale (AUC = .87).

DISCUSSION

According to the results of all dataset analyses, the scales' and sub-scales' internal consistency shows as acceptably homogenous. Following De Vellis's (2017) proposed typology, the minimum acceptable homogeneity for Cronbach's alphas would fall between .65 and .70. Considering the entirety of the BP parent scales, solely the Self-Limiting Features scale, with a .65 Cronbach's alpha, might benefit from some rectifications. This finding is comparable to findings reported in Couture et al (2011) study wherein the Cronbach's alpha for this scale proved insufficient (.24). It is worth highlighting that this parent scale contains only two sub-scales, which might lead to a considerably lower Cronbach's alpha. Furthermore, it should not be forgotten that these same two sub-scales might not entirely align with the construct that the parent scale purports to measure (Nunnaly & Bernstein, 1994). All BP sub-scales met requirements, with the exception of two (Maintains internalised standards and Makes undifferentiated attachments) whose Cronbach's alpha sat just below the minimum threshold at .62 – and while this is low homogeneity it is not entirely insufficient (< .60; De Vellis, 2017).

Table 6: *Comparison of the area under the Receiver Operating Characteristic Curves for the Boxall Profile and Strengths and Difficulties questionnaire scales*

	BOXALL PROFILE SCALES				
SDQ SCALES	**Organisation of experience**	**Internalisation of controls**	**Self-limiting features**	**Undeveloped behaviour**	**Unsupported development**
Conduct problems	–	–	.73 (1)	13.30*** (1)	7.52** (1)
Peer relationship problems	–	–	4.68* (1)	15.35*** (1)	10.90*** (1)
Prosocial behaviour	3.24 (1)	22.02*** (1)	–	–	–
Total difficulties score	–	–	3.07 (1)	.85 (1)	.07 (1)

Note 1. These values represent the results of a chi-square X2 test with degrees of freedom indicated in parentheses. Chi-square test allows the analysis of the distribution of positive and negative diagnoses (gold standard) according to the scales compared between BP and SDQ. The comparison was carried out using the methods proposed by DeLong et al (1998).
Note 2. Positive diagnosis: n = 99, negative diagnosis: n = 86.
**p < .05; ** p < .01; ***p < .001*

As per the Self-limiting features scale, these two sub-scales have a low item count at two and three items respectively. Incidentally, De Vellis's (2017) proposed cut-off threshold is less stringent than the norms that other psychometric experts recommend. Point in fact, according to Nunnaly and Bernstein (1994), an acceptable threshold for Cronbach's alpha varies depending on how the test is being used. In their expert opinion, when an instrument is used in research endeavours its Cronbach's alpha should range from .70 to .90. In the context of this particular study, the Cronbach's alpha for an instrument used in clinical interventions should sit above .90 and ideally above .95 (Nunnaly & Bernstein, 1994). These norms are nonetheless contested by Streiner (2003) who suggests rather that Cronbach's alpha should not exceed .90 for instruments used in clinical contexts, to avoid any unnecessary item redundancy or across-item construct duplication. Among the BP scales, only two (Organisation of experience and Unsupported development) meet Nunnaly and Bernstein's (1994) requirements. However, in consideration of Streiner (2003), with the exception of the Self-limiting features scale, all BP scales have sufficient homogeneity for clinical-use purposes.

That the BP scales' internal consistency is found on the whole to be acceptable is no surprise, remembering that Cronbach's alphas are estimates based on item-to-item correlational averages within a given scale (Nunnally & Bernstein, 1994). In this respect, for the BP structure's first tier, EFA showed the presence of two distinct factors with a negative correlation. In the structure's second tier (parent scales), analyses emphasised a sole factor per section. In short, the two diametrically opposed factors identified using EFA in the first tier of the questionnaire structure are corroborated by the single factors noted in each of the two second-tier sections accounting respectively for 71.3% and 65.5% of the variance.

The divergence of the two above-mentioned factors can also be observed in the correlation matrices used to study the BP's internal and external construct validity. Correlation matrices indicate that the Developmental Strands scales correlate negatively with the Diagnostic Profile scales, confirming a sound level of construct divergence. Contrastingly, interscale correlations for each section show strong association and point to a good level of construct convergence. Furthermore, each of the BP sub-scales shows a strong correlation to its parent scale, further indicating a substantial level of convergence. Construct divergence on the BP sub-scales has been demonstrated by way of a negative correlation between its own opposing constructs, as well as when cross-correlated with the SDQ. This is evidenced by the fact that the Diagnostic Profile scales (measuring constructs associated with behavioural difficulties) show a negative correlation with SDQ Prosocial Behaviour scale (constructs measuring desirable skills). Similarly, comparable constructs measured by both questionnaires converge with positive correlations when comparing the Diagnostic Profile with the entirety of the SDQ's behavioural difficulties items, and when comparing the BP's Developmental Strands with the SDQ's Prosocial Behaviour items. All these results support Couture et al's (2011) study findings. Indeed, the two scales under the Developmental Strands section show a negative correlation with the three scales under the Diagnostic Profile and vice versa. Furthermore, the aforementioned correlations move in the same direction, while BP and SDQ cross-correlations were also the same. Contrary to Couture et al (2011) the correlational relationships in this present study, however, were overall found to be higher and more significant statistically.

The above evidence supports the presence of two distinct constructs that are diametrically opposed. We can therefore attest to the validity of the BP's structure insofar as the first tier is concerned (the two core sections of the questionnaire). The existence of this factor in the structure's first tier accounts for 57.8% of the variance and can likely be linked to its matching construct in behavioural difficulties. In considering the sample group closely, with 70% of the children in the study sample attending specialised classrooms (either KC or special-education type classes), this could very well explain the sound psychometric qualities associated with the Diagnostic Profile given that its creators had intended it specifically to assess behavioural challenges in children. This sample-group characteristic could certainly lead to biased results on the Diagnostic Profile analyses, particularly in light of a sole predominant factor being observed on the first tier of the BP structure in EFA. The aim being to assess students experiencing behavioural challenges, this likely led to a lower focus on concepts linked to the desirable competencies covered under the Developmental Strands. Notwithstanding this, findings demonstrate on the whole that the Diagnostic Profile contains better psychometric properties than the Developmental Strands, as much from a reliability standpoint as from a concurrent validity standpoint. The AUC for the Diagnostic Profile's three scales is greater than the Developmental Strands scales, and all comparative findings lead to the conclusion that these differences are statistically significant. We can conclude that the Diagnostic Profile's scales have a greater capacity to correctly discern which children are actually challenged in accordance with the gold standard (based on the TRF's clinical cut-off point). In addition, in comparing the BP's AUC with the SDQ's, the statistically insignificant differences suggest that the two questionnaires boast equivalent predictive validity insofar as their ability to distinguish between

children with behavioural challenges. That said, comparative analyses also show several BP scales (among the Diagnostic Profile's scales) as having better diagnostic performance than some of the SDQ scales. Nonetheless, both questionnaires may be used for different clinical and research purposes. Considering that the SDQ has the advantage of brevity (completion time) with a count of only 25 items (McCrory & Layte, 2012), it may be used as a routine outcome evaluation, in the context of systematic screening for mental health issues or in epidemiological research (Goodman, 1997; Johston & Gowers, 2005; Sosu and Schmidt, 2017). On the other hand the BP, which takes longer to complete, is mainly used for clinical contexts or for the purpose of designing individual educational plans in school (Bennathan & Boxall, 1998). We must underscore that any statements regarding the BP's predictive capabilities were somewhat skewed by our methodology in this study. First, the sample contained a low number of children falling in the normal range (n = 44). Second, the TRF's combined Total Difficulties scale which was adopted as our gold standard is conceptually in much closer alignment with the Diagnostic Profile than with the Developmental Strands section, potentially leading to a bias in consideration of the Diagnostic Profile's predictive capabilities.

CONCLUSION AND RECOMMENDATIONS

To summarise, the French BP has good predictive validity. In effect, the results herein confirm the BP's effectiveness in screening students with behavioural difficulties, which is consistent with those of Couture et al's (2011) study. There is undoubtedly some appeal to further exploring results that demonstrate the questionnaire's strong performance as a diagnostic tool. However, one must not lose sight of the instrument's primary function as a tool intended to aid teachers in a workplace setting. The BP was originally developed to provide a common language between teachers to discuss SEBD students, and with the ultimate goal of ensuring consistency in their interventions. These interventions were geared more toward item analysis and a deeper clinical interpretation of the BP's subscales, which does not entirely align with the type of statistical methods used in this study. That said, future endeavours to further investigate the validity of the French BP should lean more toward defining and conceptualising the sub-scale constructs to assess its usefulness as a clinical rather than diagnostic tool.

The BP's structural weaknesses that we observed via the factor analysis findings, could in part be explained by certain sample-group characteristics. However, other hypotheses could also account for this flaw. In essence, the translation and cultural adaptation of any questionnaire comes with limitations, especially if not done with proper attention and scientific rigour. While the BP French translation was done by bilingual professionals, many key transcultural adaptation steps were overlooked (Vallerand, 1989). It is important to recognise that the BP was originally developed in a particular cultural and linguistic context with its own specificities. That is to say, mental constructs might differ beyond a simple question of word-based semantics and terminology. In turn, a given user's interpretation and application of results could naturally be biased if the translated terms differ in meaning in another setting. Nevertheless, even with a rigorous translation process, there is still some risk of artificial correlation between word meanings from one language to another. Naturally, in consideration of the language barriers in play, any other-language versions produced cannot be considered better than the original (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014; Sarrazin, 2003). All things considered, to effectively mitigate any negative influence on the French BP's validity that might stem from any aforementioned socio-linguistic challenges, it would be advisable for the current French version of the questionnaire to be back-translated by a professional practitioner (Massourbe, 2002) and then subjected to a new round of factor analyses.

LIMITATIONS

This study comprises some non-negligible limitations. Indeed, the sample collected from 2005 to 2008 includes fewer girls than boys (29.7% girls). Furthermore, due to logistical or geographical constraints, with some teachers being made responsible for the selection process in the absence of researchers and despite the fact that teachers were instructed to recruit randomly, selection bias cannot be ruled out. Additionally, the BP norms were recently revised in 2017. Unfortunately, this study used the pre-2017 norms which is a considerable limitation as average scores used in this study might not reflect the typical range of skills currently expected of primary school students (Nurture Group Network Limited, 2017). Finally, the gold standard we relied on could prove questionable as it is based on the TRF rather than a clinical and a normative sample.

REFERENCES

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms and profiles*. Burlington, VT: University of Vermont.

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

Bennathan. M., & Boxall, M. (1998). *The Boxall Profile: Handbook for Teachers*. London, United Kingdom: The Nurture Group Network.

Bennathan, M., & Boxall, M. (2000). *Effective intervention in primary schools: Nurture Groups*. London, United Kingdom: David Fulton Publishers.

Capron, C., Thérond, C., & Duyme, M. (2007). Psychometric properties of the French version of the self-report and teacher strengths and difficulties questionnaire. *European Journal of Psychological Assessment*, 23(2), 79-88.

Cooper, P., & Whitebread, D. (2007). The effectiveness of nurture groups on student progress: Evidence from a national research study. *Emotional and Behavioural Difficulties*, 12, 171-190.

Couture, C. (2009). Kangaroo classes: An adaptation of nurture groups. In C. Cefai & P. Cooper (Eds.) *Promoting emotional education: Engaging children and young people with social emotional and behavioural difficulties* (pp.151–160). London, United Kingdom: Jessica Kingsley Publishers.

Couture, C., & Bégin, J. Y. (2010). Une adaptation des Nurture Groups au Québec : le cas des classes kangourou. In N. Trépanier & M. Paré (Eds.), *Des modèles de services pour favoriser l'intégration scolaire* (pp. 285-308). Québec, QC: Presse de l'Université du Québec.

Couture, C., Cooper, P., & Royer, É. (2011). A study of the concurrent validity of the Boxall Profile and the Goodman Strengths and Difficulties Questionnaire. *International Journal of Emotional Education*, 3(1), 20-29.

Couture, C., & Lapalme, M. (2007). Les retombées de la première année d'implantation des classes Kangourou au Québec. *Nouveaux cahiers de la recherche en éducation*, 10, 63-81.

Delacour, H., Servonnet, A., Perrot, A., Vigezzi, F. F., & Ramirez, J. M. (2005). La courbe ROC (receiver operating characteristic) : principes et principales applications en biologie clinique. *Annales de biologie clinique, 63*(2), 145-154.

DeLong, E. R., DeLong, D. M., & Clark-Pearson, D. L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: A non-parametric approach. *Biometrics, 44*(3), 837-845.

De Vellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oak, CA: Sage Publication.

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal for Child Psychology and Psychiatry, 38*(5), 581-586.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). Upper Saddle River, NJ: Pearson.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L. Almqvist, F., Bilenberg, N., et al. (2007). Testing the Teacher's Report Form syndromes in 20 societies. *School Psychology Review, 36*(3), 468-483.

Johnston, C., & Gowers, S. (2005). Routine Outcome MeasurementL A Survey of UK Child and Adolescent Mental Health Servives. *Child and Adolescent Mental Health, 10*(3), 133-139.

Massoubre, C., Lang, F., Burkard, J., Jullien, M., & Pellet, J. (2002). La traduction des questionnaires et des tests : techniques et problèmes. *Revue canadienne de psychiatrie, 47*(1), 61-67.

McCrory, C., & Layte, R. (2012). Testing competing models of the Strengths and Difficulties Questionnaire's (SDQ's) factor structure for the parent-informant instrument. *Personality and Individual Differences, 52*(8), 882-887.

Ministère de l'Éducation, du Loisir et du Sport (MELS). (2008). *Indice de défavorisation 2007-2008*. Québec, QC: Gouvernement du Québec.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed). New York, NY: McGraw-Hill.

Nurture Group Network Limited (2017). Addendum : The Boxall Profile handbook (revised). Retrieved from https://www.nurtureuk.org/sites/default/files/addendum_boxall_ profile_handbook_booklet.pdf

Rescorla, L. A., Achenbach, T. M., Ginzburg, S., Ivanova, M., Dumenci, L., Almqvist, F., et al. (2007). Consistency of teacher-reported problems for students in 21 countries. *School psychology review, 36*(1), 91-110.

Riberdy, H., Tétrault, K., & Desrosiers, T. (2013). *La santé physique et mentale des enfants: une étude des prévalences cumulatives*. Québec, QC: Institut de la Statistique du Québec.

Sanders, T. (2007). Helping children thrive at school: The effectiveness of nurture groups. *Educational Psychology in Practice*, 23(1), 45-61.

Sarrazin, G. (2003). *Normes de pratiques du testing en psychologie et en éducation*. Montréal, QC: Institut de recherches psychologiques.

Seth-Smith, F., Levi, N., Pratt, R., Fonagy, P., & Jaffey, D. (2010). Do nurture groups improve the social, emotional and behavioural functioning of at risk children? *Educational & Child Psychology, 27*(1) 21-34.

Shaver, I., & Mcclatchey, K. (2013). Assessing effectiveness of nuture groups in Northern Scotland. *British Journal of Learning Support, 28*(3), 97-102.

Shojaei, T., Wazana, A., Pitrou, I., & Kovess, V. (2009). The Strengths and Difficulties Questionnaire: Validation study in French school-aged children and cross-cultural comparisons. *Social Psychiatry and Psychiatric Epidemiology, 44*(9), 740-747.

Sosu, M., & Schmidt, P. (2017). Tracking Emotional and Behavioral Changes in Childhood: Does the Strength and Difficulties Questionnaire Measure the Same Constructs Across Time? *Journal of Psychoeducational Assessment, 35*(7), 643-656.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103.

Vallerand, R. J. (1989). Vers une méthodologie de validation trans-culturelle de questionnaires psychologiques: implications pour la recherche en langue française. *Psychologie canadienne, 30*(4), 662-680.