

Article Title:

The reliability and validity of the Pupil Behaviour Questionnaire: a child classroom
behaviour assessment tool

Corresponding Author:

Mr Matt Allwood, Department of Psychiatry, University of Oxford, Warneford Hospital,
Oxford, OX3 7JX

Email: matthew.allwood@psych.ox.ac.uk

Co-Authors:

Ms Kate Allen, Department of Child Health, University of Exeter Medical School, Room
2.05, South Cloisters, St Luke's Campus, Exeter, EX1 2LU

Ms Anna Price, Department of Child Health, University of Exeter Medical School, Room
2.05, South Cloisters, St Luke's Campus, Exeter, EX1 2LU

Dr Rachel Hayes, Department of Child Health, University of Exeter Medical School, Room
2.05, South Cloisters, St Luke's Campus, Exeter, EX1 2LU

Ms Vanessa Edwards, Department of Child Health, University of Exeter Medical School,
Room 2.05, South Cloisters, St Luke's Campus, Exeter, EX1 2LU

Miss Susan Ball, NIHR CLAHRC South West Peninsula (PenCLAHRC), University of
Exeter Medical School, Room 2.05, South Cloisters, St Luke's Campus, Exeter, EX1 2LU

Dr Obioha C Ukoumunne, NIHR CLAHRC South West Peninsula (PenCLAHRC),
University of Exeter Medical School, Room 2.05, South Cloisters, St Luke's Campus, Exeter,
EX1 2LU

Professor Tamsin Ford, Institute of Health Research, University of Exeter Medical School,
Room 2.05, South Cloisters, St Luke's Campus, Exeter, EX1 2LU

Word Count: 5066

Abstract

Background: Disruptive pupil behaviour is a major concern for teachers in the classroom. There is a need for brief valid measures that can assess the impact of interventions in practice and in research as current behaviour assessment tools are either too lengthy and/or do not focus on classroom-based behaviours.

Method: This paper assesses the reliability and validity of the teacher-completed Pupil Behaviour Questionnaire (PBQ), by comparing it to the already extensively validated teacher-completed Strengths and Difficulties Questionnaire (SDQ). Participants included children participating in a universal school-based trial (Supporting Teachers And childRen in Schools; STARS; N=2074 children aged 4-9 from 80 schools) and vulnerable children who were taking part in a study exploring the impact of exclusion from school (N=41 aged 4-12 from 22 schools).

Results: The exploratory factor analysis results (first factor accounts for 80.8% of the variation in the items) and the high Cronbach's alpha value of 0.85 indicate that the PBQ consists of one substantive factor / dimension. Strong correlations between the total PBQ score and the conduct sub-scale (Spearman's correlation coefficient (r_s) = 0.67) and total difficulties score ($r_s=0.59$) of the SDQ indicate convergent validity.

Conclusions: This study suggests that the PBQ is a reliable measure, and provides some evidence of validity. Further work is needed to test the PBQ in an older, more diverse population and to measure sensitivity to change.

Keywords

Behaviour; School children; Questionnaire; Validity; Reliability

Key Practitioner Message

- Many useful measures are available to assess children's mental health and behaviour, but most are relatively long and few focus on classroom-based behaviours.
- The Pupil Behaviour Questionnaire (PBQ) is a quick and easy measure of low level disruptive behaviour in the classroom that is reliable and valid for younger children in primary schools.
- Because of its brevity, the PBQ may be particularly useful for teachers and schools wanting to assess the behaviour of groups of children.
- The PBQ is freely available to download (<http://medicine.exeter.ac.uk/research/healthresearch/childhealth/child-mental-health/pbq/>) provided the person downloading it is not charging for its use.

Introduction

Disruptive childhood behaviour predicts poor educational, occupational, health, relationship and social outcomes, with associated increased costs to society (Scott, 2007). Complaints about disruptive pupil behaviour are common and are reported to be a major contributor to stress and burnout among teachers (Brouwers & Tomic, 2000; Greene et al., 2002). In a recent survey, 42% of teachers reported that stress impacted on their ability to teach and 76% said that their stress levels had repercussions on their health (Teachers Assurance, 2013). Only 77% of teachers described their pupils' behaviour as good or very good in a recent survey completed by teachers on pupil behaviour (Weaving & Aston, 2013); while the chief Inspector of Ofsted, Sir Michael Wilshaw, has expressed concern that "teachers were too often ignoring low-level disruption leading to a culture of casual acceptance of misbehaviour that is holding schools back" (Sky News, 2013). Improvements in behaviour are now a major policy focus (Ofsted, 2015).

There are many varying definitions of problematic behaviour in schools. Most of the discussion revolves around the perception of what is described as disruptive, a concept that includes both the frequency and severity of the behaviour (Education Standards Analysis and Research Division, 2012). Cameron (1998) has proposed grouping disruptive behaviour into five categories; behaviour that is aggressive, that challenges authority, or is disruptive either physically, socially or to the child themselves. This is in contrast to Watkins and Wagner (2000) who describe low level disruption and 'talking out of turn' as the most frequent constellation of problematic behaviours that teachers report to be the trickiest to deal with. Valid and reliable measures of such pupil behaviour could be one way of clarifying the impact of different disruptive behaviours in schools.

It is important to quantify children's behaviour for various purposes, which include: tracking a child's progress over time, comparing whole class behaviour across a school, assessing behaviour at whole school level regionally or nationally, or to evaluate a behaviour management tool or intervention. Questionnaires provide a systematic approach that allows comparison across individuals, groups and times providing that the measure is reliable, valid and sensitive to the types of changes to be expected from common interventions. One of the first systematic attempts to measure children's classroom behaviour was made by Wickman in the early 1920s (Haggerty, 1925; Wickman, 1928) but with the questionnaire including 26 items it was deemed too time consuming for teachers to complete about every child in their class (Rutter, 1967).

In 1988 the Department of Education and Science commissioned an enquiry into discipline in schools in response to concerns about the difficulties facing teachers in England and Wales - the Elton Report (Department of Education and Science, 1989). This enquiry consulted with 3,500 teachers who reported on their experiences in and around their school during the previous week. In 2009 the Scottish government's study on the behaviour in

schools in Scotland developed the survey used in the Elton Report, which included a teacher-reported measure of behaviour that was 26 items long and covered low level indiscipline, disengagement and serious indiscipline/violence (The University of Edinburgh, 2009).

Arguably the best teacher report measure of classroom behaviour is the revised Sutter-Eyberg questionnaire, which comprises 38 items that are rated on both Intensity and Problem scales and is validated for children aged 2-16 years (Eyberg & Pincus, 1999). While the Sutter-Eyberg offers detail and depth of assessment for children with complex difficulties, the PBQ offers brevity for children with less severe difficulties or when teachers are required to complete measures on groups of children.

In order that a questionnaire is feasible for completion by school teachers on every child in their class it needs to be short and easy to complete (Slade et al., 1999). The aim of the current study was to assess the reliability and validity of a shorter version of the 26 items developed from the Elton Report, focussing only on the low level classroom disruption. The resulting Pupil Behaviour Questionnaire (PBQ) was developed for use in the Supporting Teachers And childRen in Schools (STARS) (Ford et al., 2012) trial in consultation with teachers and educational psychologists. We assessed the validity of the PBQ by comparing it to the already extensively validated Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997) within a sample of children from mainstream schools. A second sample of children at risk of permanent exclusion was used to assess whether the PBQ could discriminate between a vulnerable sample and a community sample of children.

Methods

Study samples/participants

This study was a secondary analysis of data from two studies: the STARS trial (Ford et al., 2012), which evaluated the effectiveness of a teacher classroom management course

and the SKIP (Supporting Kids, avoidIng Problems) study that explored mental health, attainment and ability among children at risk of permanent exclusion from school (Parker et al., 2016). Both studies had ethical permission from the University of Exeter Medical School Research Ethics Committee. The STARS sample comprised the baseline data of a large cluster randomised controlled trial of teacher classroom management, which included 80 primary schools across Devon, Plymouth and Torbay (Figure 1). Included children were in Reception to Year 4 (aged 4 to 9 years). Data were collected from three overlapping cohorts of schools at the beginning of the academic year in September 2012, 2013 and 2014.

The SKIP study (Parker et al., 2016) was a case-control study that assessed the mental health of children at risk of or who had experienced exclusion from primary school and the first year of secondary school. Children were identified by an educational or mental health practitioner and recruited from the Southwest of England between September 2011 and July 2013. Here we only present data from the cases, or those at risk of exclusion.

** INSERT FIGURE 1 HERE**

Measures

To be feasible for completion by primary school teachers on every child in their class, the PBQ was limited to six items from the questionnaire used in the Elton Report (Department of Education and Science, 1989) and, in consultation with teachers and educational psychologists, focused on the types of low level disruption reported to be common and most difficult to manage in primary schools (Watkins & Wagner, 2000). The 6 items were; 1) talking out of turn, 2) interrupting other pupils, 3) making unnecessary noise, 4) verbal abuse towards other pupils, 5) physical aggression towards other pupils, and 6) cheeky or rude remarks to the teacher. Each of the six items addresses a negative behaviour, with responses on a 3 point scale, where 0, 1 and 2 correspond to the child *never*,

occasionally and *frequently* exhibiting that behaviour, respectively. The total score for the PBQ is calculated as the sum of the six items and therefore has a possible range between 0 and 12, with higher scores indicating more disruptive behaviour. The PBQ is freely available to download (<http://medicine.exeter.ac.uk/research/healthresearch/childhealth/child-mental-health/pbq/>) provided the person downloading it is not charging for its use.

The Strengths and Difficulties Questionnaire (SDQ) (Goodman, 1997) is a well-validated measure of childhood mental health and behaviour; as such, we would expect to see a strong association between individual children's scores on the SDQ and PBQ. The SDQ is a 25-item questionnaire with responses to each item on a three point Likert scale; *not true* (0), *somewhat true* (1) and *certainly true* (2) for difficulties or reversed for strengths so that a high score indicates difficulty except for the pro-social scale where the opposite applies (Goodman, 1997). Of the 25 items, 10 are thought of as strengths and 15 are thought of as difficulties. The questionnaire is split into five sub-scales: *emotional*, *conduct*, *hyperactivity*, *peer problem* and *pro-social*. The total score for each of these sub-scales ranges from 0 to 10, and is calculated by adding the scores of the items for that sub-scale. The sub-scale scores, excluding the pro-social sub-scale, are summed to form the total difficulties score which ranges from 0 to 40. The SDQ was completed by both teachers and parents in both studies.

The reliability and validity of the SDQ have been extensively tested. Goodman (2001) has shown that the internal consistency is satisfactory (Cronbach's alpha (α)=0.73) and after 4 to 6 months the retest stability of teacher ratings were satisfactory (mean 0.62) in a national sample of school aged children (Goodman, 2001).

Statistical analyses

Participant demographic characteristics, PBQ and SDQ total and sub-scale scores were summarised using means and standard deviations for continuous variables and numbers

and percentages for categorical variables. Analyses were carried out using Stata 13.1, R (3.2.0) and Mplus Version 6.

Factor Analysis

An exploratory factor analysis was carried out on the 6 PBQ items to identify themes (factors) that underlie the measure (Bryman & Cramer, 2005) using Mplus (Muthen & Muthen, 2007). As the response set was comprised of only three levels the items were treated as ordinal categorical in the analysis, as opposed to continuous. Mean- and variance-adjusted ('robust') weighted least-square parameter estimates of the factor loadings were obtained based on analysis of polychoric correlations between the PBQ items. The analysis allowed for clustering within schools and we report within-cluster factor loadings.

Internal consistency

The internal consistency of the PBQ was assessed using Cronbach's alpha, with values of 0.70 to 0.90 considered to demonstrate good reliability, whilst bearing in mind that a Cronbach's alpha of greater than 0.90 could suggest that some items were redundant (Streiner, 2003; Tavakol & Dennick, 2011).

Discriminant validity

Using a random effects linear regression model, the mean PBQ total score in the community sample (STARS) was compared to the vulnerable children from the SKIP sample. This method takes account of the correlation between the responses related to children from the same class. We predicted that the children with difficulties in coping with school would have higher levels of disruptive behaviour due to the threatened or actual breakdown of their school placement.

Construct validity

Construct validity was assessed by examining convergent and divergent validity. A measure will be more closely related to another measure of a similar construct than a measure that assesses a different construct. We reported the Spearman's correlation coefficient for the association between the PBQ with the different subscales of the SDQ as completed by both teachers and parents. We expected there to be stronger correlations between teacher SDQ and PBQ than between parent SDQ and PBQ since the same respondent (the class teacher) completed both the teacher SDQ and PBQ measure.

Convergent validity was quantified by the correlation between the PBQ total score and the behaviour subscale, hyperactivity subscale and the total difficulties score from the SDQ. These SDQ subscales measure disruptive behaviours that would be expected to positively correlate with the total score of the PBQ. *Divergent validity* was measured by comparing the total score from the PBQ with the pro-social subscale from the SDQ. This SDQ subscale measures positive behaviours which would be expected to negatively correlate with the PBQ focussing on disruptive behaviour.

Gender and school effects

Random effects linear regression models were fitted to examine the relationships between total PBQ score (outcome) and each of gender and school year (potential predictors). It would be expected that boys would score higher as they tend to exhibit more hyperactive and antisocial behaviour than girls (Sylva et al., 2012). We expected that low level disruptive behaviour would be more common in younger children, which would be reflected in a higher total PBQ (Ford et al., 2007).

Attainment

Using the STARS sample, we estimated the correlation between the total PBQ score and each of the literacy and numeracy attainment scores. Given the strong association

between learning difficulties and disruptive behaviour, we expected a negative correlation between the PBQ and attainment (Ford et al., 2007).

Results

Study participants

The STARS study comprised 2,074 children in 80 different schools, taught by 80 different teachers who completed the PBQ and SDQ questionnaires on 2,074 of these children. Of these 2,074 children, 1,463 (70.5%) also had a parent completed SDQ, as shown in Table 1. The SKIP study provided a sample of 41 children aged 5-12 years from 22 schools in Devon. Of these 41 children, 30 (73%) had a PBQ completed by their current class teacher and all 41 had a parent-completed SDQ, as shown in Table 1 and Figure 1.

The distribution of children in the STARS sample was balanced in terms of gender and age. In contrast, the SKIP children, who were at risk of exclusion, were, as we had predicted, mainly boys (95%) and tended to be older. The STARS sample had more complete data from teachers who were the primary recruiting source, while parents provided the most data in the SKIP study since they were the primary contact and contact with teachers was only made if the parent and child agreed. Item-level data for the PBQ shows scores were generally low for the STARS study and high for the SKIP study across all items, as shown in Figure 2.

**** INSERT TABLE 1 AND FIGURE 2 HERE ****

In the factor analysis, the eigenvalue of the first factor was 4.845 (accounting for 80.8% of the total variation in the PBQ item scores) and those of the remaining five factors were all less than 0.5. This indicates that there is only one salient factor or substantive dimension that underlies the PBQ item scores and the items measure a single construct. The

loadings for the first factor from the one factor solution indicate strong correlations between each of the PBQ items and the factor: 0.908 for talking out of turn, 0.952 for interrupting other pupils, 0.883 for making unnecessary noises, 0.899 for verbal abuse towards peers, 0.874 for physical aggression towards peers, and 0.828 for rudeness towards teachers. The PBQ demonstrated excellent internal consistency both overall (Cronbach's alpha=0.85) and for most subgroups defined by gender and age (Table 2).

**** INSERT TABLE 2 HERE ****

As expected, the SKIP sample demonstrated much higher levels of disruptive behaviour than the STARS sample; on average, total PBQ was 7 points higher in the SKIP sample (95% CI: 6.1 to 7.9, $p < 0.0001$). Moderate to strong positive correlations were found between the PBQ and each of the teacher-rated SDQ conduct, hyperactivity and total difficulties scores ($p < 0.0001$ for all), which demonstrates convergent validity, as shown in Table 3. Weak correlations were found between the PBQ and the teacher-rated SDQ emotional symptoms subscale and peer problems subscale and a moderate negative correlation was found with the pro-social subscale ($p < 0.0001$ for all), demonstrating divergent validity, as shown in Table 3. As anticipated, weaker correlations were found between the PBQ and parent-completed SDQ, than between the PBQ and teacher-completed SDQ (Table 3).

**** INSERT TABLE 3 HERE ****

Relationships of PBQ score with gender and school year

On average, total PBQ was estimated to be 1.3 points lower among females than males (95% CI for reduction: 1.1 to 1.5, $p < 0.0001$). There was little evidence of a relationship between total PBQ and school year ($p = 0.9$).

Attainment and the PBQ

There were small negative correlations found between the PBQ and both literacy (-0.20; 95% CI: -0.25 to -0.16; $p < 0.0001$) and numeracy (-0.13; 95% CI: -0.18 to -0.09; $p < 0.0001$) attainment scores.

Discussion

This is the first study to test the psychometric properties of this novel measure of low level disruption for primary school age pupils. Our initial findings suggest that it is both reliable and valid. The PBQ measures a single construct with excellent internal consistency and is correlated highly with both teacher- and parent-completed SDQ scores in the expected directions. It is widely accepted that a scale with internal consistency estimates between 0.7 and 0.9 indicates that a test is a reliable measure (McCrae et al., 2011). It is encouraging that the children at risk of permanent exclusion from the SKIP sample scored significantly higher than those from the community sample, but also that they were not consistently scoring at the highest level on the PBQ. This suggests that there was not a ceiling effect with this scale, even among children whose behaviour was so challenging they were at risk of exclusion from mainstream school.

As expected, on average, girls had a lower average total PBQ score than boys, but no relationship was found between age and total score among our primary school sample. Differences might emerge if tested on a secondary school sample. The association found between disruptive behaviour reported on the PBQ and lower attainment scores replicates previous work (Ford et al., 2007).

As teachers completed both the PBQ and the teacher SDQ it is not surprising that there was a stronger correlation between the PBQ and the teacher-completed SDQ than the PBQ and the parent-completed SDQ. A similar pattern has been seen in other studies with the

teacher-completed SDQ correlating more strongly with the Child Behaviour Checklist (CBCL) than the parent-completed SDQ, and relatively low inter-rater correlation is a commonly reported phenomenon in child mental health research (Goodman & Scott, 1999; Stone et al., 2010). Children may function differently in different environments and each informant will have their own skills and biases that will influence reporting (Collishaw et al., 2009), but the PBQ is designed to measure classroom behaviour and so scores will be more salient to behaviour at school than at home.

Given our findings have demonstrated robust psychometric properties in this short-item measure, the PBQ has the potential to become a widely used method for quantifying the behaviour of primary school children in the classroom. The PBQ focuses on behaviours that commonly disrupt teaching.

Our findings are particularly pertinent currently given the Government's recommendations for a greater emphasis on early interventions for children (Department of Health, 2015) and the drive towards evidence-based interventions in schools and health care (Education Endowment Foundation, 2015). A lack of effective support in a child's early years has been identified as a potential cause of many of the costly and damaging social problems in society (Scott et al., 2001).

Many academics accept the use of Cronbach's alpha to assess internal consistency as an acceptable measure of reliability but internal consistency does not necessarily equate to a measure that reliably reproduces the same results (McCrae et al., 2011). Future research is needed to assess the test-retest reliability of the PBQ as well as longitudinal studies assessing the stability of scores over time; epidemiological studies show that without intervention, disruptive behaviour is a highly persistent trait (Costello & Maughan, 2015). In addition, to be useful as an outcome measure in practice, we need to test the PBQ's sensitivity to change

when evaluating evidence-based interventions. The appropriateness of the measure to children aged 10 and over and with severe behavioural difficulties also remains to be assessed.

This analysis benefitted from a comparison measure that was validated and frequently used in both practice and research and from the fact that the STARS sample offered access to a large unselected primary school sample. It is not, however, without limitations. As with all secondary data analyses, we were restricted by the constraints of the data that we had access to. We lacked data on children over the age of 9. The PBQ could be useful and feasible for older children, particularly those coming up to transition to secondary education or in the earlier years of secondary school, but this requires empirical testing. The longer Scottish questionnaire from which the PBQ was developed was aimed at both primary and secondary school aged children, but has never been validated (The University of Edinburgh, 2009). Our findings suggest that it might be worth further testing of the PBQ with samples of secondary school aged children.

The STARS trial was based solely in the South West Peninsula of England (Ford et al., 2012), which has a lower than average ethnic diversity, compared to the rest of England and Wales (2.5% vs 14%) (Strategic Intelligence, 2012). The findings, therefore, may not be generalizable outside the South West of England. This limitation can be addressed by comparing results from children from a wider ethnic background.

Further refinement of the measure could involve a qualitative study exploring the reasons why teachers made the responses they did and how these may be different for each teacher completing the measure.

In summary, our initial findings suggest that the PBQ is a promising, psychometrically valid measure of low level behavioural disturbance within primary school

classrooms and could be of use in relation to the assessment of individual, class and school level difficulties. Future work needs to focus on testing the PBQ in different populations of children and on assessing how sensitive it is over time.

Acknowledgements

Funding

This project was funded by the National Institute for Health Research Public Health Research Programme (project number 10/3006/07) and the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula.

The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the NIHR Public Health Research Programme, NIHR, NHS or the Department of Health.

Conflict of Interest

The author(s) have declared that they have no competing or potential conflicts of interest

Contributors

The manuscript was prepared by MA with assistance from all co-authors. Data analysis were carried out by MA, SB, and OU. TF oversaw all aspects of the research project.

References

- Brouwers, A., & Tomic, W. (2000). A longitudinal study of teacher burnout and perceived self-efficacy in classroom management. *Teaching and Teacher Education, 16*(2), 239-253.
- Bryman, A., & Cramer, D. (2005). *Quantitative Data Analysis with SPSS 12 and 13: A Guide for Social Scientists*. Hove: Routledge.

- Cameron, M. (1998). School discipline in the United Kingdom: promoting classroom behaviour which encourages effective teaching and learning. *School Psychology Review*, 27(1), 33-44.
- Collishaw, S., Goodman, R., Ford, T., Rabe-Hesketh, S., & Pickles, A. (2009). How far are associations between child, family and community factors and child psychopathology informant-specific and informant-general? *Journal of Child Psychology and Psychiatry*, 50(5), 571-580.
- Costello, E. J., & Maughan, B. (2015). Annual Research Review: Optimal outcomes of child and adolescent mental illness. *Journal of Child Psychology and Psychiatry*, 56(3), 324-341.
- Department of Education and Science. (1989). *Discipline in Schools Report of the Committee of Enquiry chaired by Lord Elton*. London: HMSO.
- Department of Health. (2015). *Future in mind: Promoting, protecting and improving our children and young people's mental health and wellbeing*. London: Department of Health.
- Education Endowment Foundation. (2015). *Annual report 2014/15*. London: Education Endowment Foundation.
- Education Standards Analysis and Research Division, Department for Education. (2012). *Pupil Behaviour in Schools in England*. London: Department for Education.
- Eyberg, S., & Pincus, D. (1999). *Eyberg Child Behavior Inventory & Sutter-Eyberg Student Behavior Inventory-Revised: Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Ford, T., Collishaw, S., Meltzer, H., & Goodman, R. (2007). A prospective study of childhood psychopathology: independent predictors of change over three years. *Soc Psychiatry Psychiatr Epidemiol*, 42(12), 953-961.

- Ford, T., Edwards, V., Sharkey, S., Ukoumunne, O. C., Byford, S., Norwich, B., et al. (2012). Supporting teachers and children in schools: the effectiveness and cost-effectiveness of the Incredible Years teacher classroom management programme in primary school children: a cluster randomised controlled trial, with parallel economic and process evaluations. *BMC Public Health*, *12*, 719.
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry*, *38*(5), 581-586.
- Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337-1345.
- Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is Small Beautiful? *Journal of Abnormal Child Psychology*, *27*(1), 17-24.
- Greene, R. W., Beszterczey, S. K., Katzenstein, T., Park, K., & Goring, J. (2002). Are Students with ADHD More Stressful to Teach?: Patterns of Teacher Stress in an Elementary School Sample. *Journal of Emotional and Behavioural Disorders*, *10*(2), 79-89.
- Haggerty, M. E. (1925). The Incidence of Undesirable Behavior in Public-School Children. *Journal of Educational Research*, *12*(2), 102-122.
- McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28-50.
- Muthen, L. K., & Muthen, B. O. (2007). Mplus User's Guide. Los Angeles: Muthen & Muthen.

- Ofsted. (2015). *The Annual Report of Her Majesty's Chief Inspector of Education, Children's Services and Skills 2014/15*. London: Ofsted.
- Parker, C., Paget, A., Ford, T., & Gwernan-Jones, R. (2016). 'he was excluded for the kind of behaviour that we thought he needed support with...' A qualitative analysis of the experiences and perspectives of parents whose children have been excluded from school. *Emotional and Behavioural Difficulties*, 21(1), 133-151.
- Rutter, M. (1967). A children's behaviour questionnaire for completion by teachers: preliminary findings. *Journal of Child Psychology and Psychiatry*, 8(1), 1-11.
- Scott, S. (2007). Conduct disorders in children. *BMJ*, 334(7595), 646-646.
- Scott, S., Knapp, M., Henderson, J., & Maughan, B. (2001). Financial cost of social exclusion: follow up study of antisocial children into adulthood. *BMJ*, 323(7306), 191.
- Sky News. (2013). Ofsted: Schools Ignoring Bad Behaviour. Retrieved 4th March 2014, from <http://news.sky.com/story/1180591/ofsted-schools-ignoring-bad-behaviour>
- Slade, M., Thornicroft, G., & Glover, G. (1999). The feasibility of routine outcome measures in mental health. *Social Psychiatry and Psychiatric Epidemiology*, 34(5), 243-249.
- Stone, L. L., Otten, R., Engels, R. C., Vermulst, A. A., & Janssens, J. M. (2010). Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4- to 12-year-olds: a review. *Clinical Child and Family Psychology Review*, 13(3), 254-274.
- Strategic Intelligence. (2012). *2011 Census Profile: Devon County*. Exeter: Devon County Council.
- Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99-103.

- Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2012). *Final Report of the Key Stage 3 Phase: Influences on Students' Development from age 11 – 14*. London: Department for Education.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55.
- Teachers Assurance. (2013). Teachers' stress levels affecting performance. Retrieved 14th March 2014, from <https://www.teachersassurance.co.uk/money-news/teachers-stress-levels-affecting-performance>
- The University of Edinburgh. (2009). *Behaviour in Scottish Schools 2009*. Edinburgh: Scottish Government.
- Watkins, C., & Wagner, P. (2000). *Improving School Behaviour*. London: SAGE Publications.
- Weaving, H., & Aston, H. (2013). *Teacher Voice Omnibus May 2013 Survey: pupil behaviour*. London: NFER.
- Wickman, E. K. (1928). *Children's behavior and teachers' attitudes*. Oxford: Commonwealth Fund, Division of Publications.

Table 1: Participant characteristics and response rates for the PBQ and SDQ by sample.

	Community sample	Exclusion sample
	STARS	SKIP
DEMOGRAPHICS	N = 2,074	N = 41

Male, n (%)	1,101 (53)	39 (95)
Age in years, mean (SD)	6.3 (1.3)	8.6 (2.1)
Year group, n (%) ¹		
Reception	270 (13)	1 (3)
Year 1	368 (18)	6 (15)
Year 2	410 (20)	4 (10)
Year 3	608 (29)	5 (13)
Year 4	418 (20)	5 (13)
Year 5	-	10 (25)
Year 6	-	5 (13)
Year 7	-	4 (10)
TEACHER COMPLETED QUESTIONNAIRES - PBQ		
	N = 2,074	N = 30
PBQ, mean (SD)	1.9 (2.4)	8.97 (2.4)
TEACHER COMPLETED QUESTIONNAIRES - SDQ		
	N = 2,074	N = 32
SDQ total difficulties score, mean (SD)	6.7 (5.9)	22.3 (5.1)
SDQ conduct sub scale score, mean (SD)	0.8 (1.5)	5.4 (1.8)
SDQ prosocial score, mean (SD)	7.5 (2.4)	3.9 (1.9)
PARENT COMPLETED QUESTIONNAIRES		
	N = 1,462	N = 41
SDQ total difficulties score, mean (SD)	7.0 (5.6)	22.3 (6.9)
SDQ conduct sub scale score, mean (SD)	1.5 (1.6)	5.2 (2.3)
SDQ prosocial score, mean (SD)	8.4 (1.7)	6.0 (2.1)

¹ data only available for 40 children in the SKIP sample

Table 2: Cronbach's alpha (α) for the PBQ in the STARS sample by gender and age subgroups.

School year	Male		Female	
	N	α	N	α
Reception	152	0.81	118	0.77

Year 1	184	0.85	184	0.81
Year 2	219	0.87	191	0.65
Year 3	319	0.86	289	0.79
Year 4	227	0.87	191	0.81

Table 3: Spearman's correlation coefficient (r_s) between the PBQ and SDQ in the STARS sample

Variables compared	r_s	95% CI	p-value
<i>Correlations between PBQ and teacher completed SDQ</i>			
Total PBQ vs teacher rated SDQ total	0.59	0.56 to 0.62	<0.0001
Total PBQ vs teacher rated conduct subscale	0.67	0.65 to 0.70	<0.0001
Total PBQ vs teacher rated pro-social subscale	-0.53	-0.56 to -0.50	<0.0001
Total PBQ vs teacher rated hyperactivity subscale	0.72	0.70 to 0.75	<0.0001
Total PBQ vs teacher rated peer problems subscale	0.19	0.14 to 0.23	<0.0001
Total PBQ vs teacher rated emotional symptoms subscale	0.01	-0.03 to 0.05	<0.0001
<i>Correlations between PBQ and parent completed SDQ</i>			
Total PBQ vs parent rated SDQ total	0.25	0.20 to 0.30	<0.0001
Total PBQ vs parent rated conduct subscale	0.28	0.23 to 0.32	<0.0001
Total PBQ vs parent rated pro-social subscale	-0.24	-0.29 to -0.19	<0.0001
Total PBQ vs parent rated hyperactivity subscale	0.33	0.28 to 0.38	<0.0001
Total PBQ vs parent rated peer problems subscale	0.12	0.06 to 0.17	<0.0001
Total PBQ vs parent rated emotional symptoms subscale	0.00	-0.05 to 0.06	0.86