Effects of measurement aggregation on predicting externalizing problems from preschool behaviour

Marcel Zentner1*, Milana Smolkina2 and Peter Venables3

1Department of Psychology, University of Innsbruck, Austria
2Institute of Psychiatry, King’s College, London, UK
3Department of Psychology, University of York, UK

In long-term studies of psychological development, the initial assessment of etiologically significant child behaviours is often carried out at a single point in time only. However, one-time assessments of behaviour are likely to possess limited reliability, leading to attenuated longitudinal correlation coefficient magnitudes. How much this bias might have affected behavioural continuity estimates in longitudinal research is presently unknown. Using a data set from the Mauritius Child Health Project, we particularize the attenuating effects of single-occasion behavioural assessments on consistency estimates of impulsive–aggressive behaviour over time. Specifically, two nursery teachers provided 15 consecutive weekly ratings of the aggressive behaviour of 99 four-year-old children. The same children were reassessed for the presence of externalizing behaviour problems at the ages of 8 and 10. There were substantial increases in both reliability and predictive correlation coefficient magnitudes when the preschool scores were aggregated across several weekly ratings. A further increase resulted after the two outcome assessments were combined into a composite score of school-age externalizing symptoms. A generalized procedure, developed from the correction for attenuation formula, is introduced to describe the relation of aggregation to predictive validity in longitudinal research.

The relationships between early-appearing behavioural characteristics and later personality traits and behaviour disorders are among the most important and controversial themes in developmental psychology and psychopathology. The general view is that, if examined over long period of time, predictive relationships tend to be small (Kagan & Zentner, 1996), even if these ‘small’ effects exhibit consistency across studies (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007; Zentner & Shiner, 2012). Often omitted from discussions of behavioural stability over time is the attenuation in stability estimates likely to arise from imperfect measurement. Nowhere is this omission more consequential than in the assessment of early childhood behaviours that are seen as risk factors for the development of behaviour problems. In most studies of life-course development, these behaviours have been assessed at one single time, typically in the context of a behavioural observation carried out during the first 3–4 years of life. This practice pervades various domains of etiological enquiry, including depression, substance abuse, and crime (e.g., Caspi, Moffitt, Newman, & Silva, 1996; Laucht, Becker, & Schmidt, 2006; Moffitt et al.,...

*Correspondence should be addressed to Marcel Zentner, Department of Psychology, University of Innsbruck, 6020 Innsbruck, Austria (email: marcel.zentner@uibk.ac.at).

DOI:10.1111/bjdp.12059
2011); executive functioning deficits (e.g., Friedman, Miyake, Robinson, & Hewitt, 2011); and attachment security (Pinquart, Feusser, & Ahnert, 2013).

Relying on single-occasion assessments is understandable due to issues of time and cost that will often preclude using repeated assessments of early childhood behaviour. However, insufficient behavioural sampling has its own costs: In most cases, it is likely to attenuate reliability and, consequently, predictive validity. In contrast, a set of multiple or repeated measurements will provide a more reliable assessment of behavioural attributes. Aggregate scores can be expected to be more valid than any of the individual measurements, because true score variance will cumulate, whereas the unique variance associated with each individual measurement will not cumulate. In addition, errors entailed in any single measurement are averaged out when aggregating across multiple measurements (e.g., Epstein, 1979; Haynes & O’Brien, 2000).

This psychometric notion, though widely known and accepted in theory, has seldom been empirically substantiated in child development studies (see Moscowitz & Schwarz, 1982; Rushton, Brainerd, & Pressley, 1983, for exceptions). Particularly unclear is the extent to which prediction of behaviour problems may be weakened by single-occasion assessments of early childhood behaviours. Closing this gap seems all the more important in a field in which conclusions must be drawn from a relatively small number of birth cohort studies, and thus, any bias is likely to substantially affect views about the role of early childhood behaviours in the development of personality and psychopathology.

The current research addresses this issue empirically by examining the attenuating effects of single-occasion assessments of early childhood aggression and impulsivity on the prediction of externalizing symptoms in school age. Disruptive behaviours were selected in this study because they are pervasive and a source of particular concern and cost to the public sector. Recent epidemiological figures suggest that about 6% of children and young people aged 5–16 suffer from clinically diagnosed conduct disorder in the United Kingdom, generating costs of about 5,500 GBP (over 3 years) per child to the public sector (Snell et al., 2013).

In this study, two nursery school teachers observed the aggressive–impulsive behaviours of 4-year-old children over a period of 15 weeks, providing 15 weekly ratings. At the ages of 8 and 10, the same children were reassessed for externalizing symptoms. From these data, it was possible to systematically compare longitudinal correlation coefficient magnitudes between single-occasion measurements versus multiple aggregated measurements of preschool aggressive behaviour and later externalizing behaviour problems.

**Methods**

**Participants**

Participants were 99 children (49 boys and 50 girls) from the sample of 1,795 children in the 1969 birth cohort in Mauritius, who had been selected for in-depth observations at nursery school. The ethnic make-up of this subset was as follows: 30.3% general population (Creoles), 38.4% Hindu, 23.2% Muslim, 7.1% Tamil, and 1% other. These children were selected to be educated at two nursery schools, which placed the children into a cognitively, nutritionally, and physically enriched environment (see Venables, 1978, for additional detail on sample composition and Raine, Mellingen, Liu, Venables, & Mednick, 2003, for details of the enrichment programme). Informed consent was obtained from the parents when the participants entered the nursery schools and
subsequently in their later school placements. The research activities were conducted in accordance with the principles outlined in the Declaration of Helsinki (World Medical Association, 1964).

**Procedure**

**Assessment at the age of 4**

The assessments began 7 months after the children had all been enrolled in nursery school and extended from June to December 1974. Two nursery school teachers provided weekly ratings of child behaviour over a period of 18 weeks. The number of children attending nursery school each week varied somewhat. Therefore, scores for aggression ranged between $N = 57$ and $N = 96$ for the first rater and between $N = 31$ and $N = 86$ for the second rater. To include the maximum number of observations in the analyses, we did not include Weeks 6, 10, and 18, when school attendance was low due to a religious holiday. Further details are given in the Data Analytic Strategy section. Of the two raters, one usually knew the children best, whereas the second rater was not as familiar with them and was not always present to provide an additional rating. The children were rated while playing with toys that had been placed in a particular room within the nursery schools. Children were observed for several behaviours that were each defined by one descriptive sentence. The two behaviours known to be predictors of externalizing behaviour were (1) *aggression*, defined as ‘child throws toys, tears things down, breaks toys, attacks other children, pushes, hits, takes things from others’ and (2) *impatience*, defined as ‘child is not able to wait to get something he/she wants, cannot wait his turn, cannot delay satisfaction’. The raters were asked to rate these two behavioural items on a 5-point scale ranging from 1 (never) to 5 (frequently) and to avoid any exchange of information with one another during and after the ratings.

**Assessment at the age of 8**

When the children were on average 8.25 years old ($SD = 0.94$), students from the Mauritian Institute of Education visited the primary schools in which the children had been placed after leaving the nursery schools and instructed the teachers to rate the children’s behaviour on the *Children’s Behaviour Questionnaire for Completion by Teachers Scale* (Rutter, 1967). The instrument used was designed to produce three factors that were termed ‘Hyperactive-Aggression’, ‘Worried-Fearful’, and ‘Physical/Speech difficulty’. However, a factor analysis of the data from the present project showed the presence of two factors (Venables *et al.*, 1983). In accordance with current terminology, these were labelled as externalizing and internalizing behaviour. The externalizing scale consisted of nine items ($\alpha = .74$) and included such items as ‘frequently fights with other children’, ‘often destroys own or others’ belongings’, or ‘is often disobedient’. Items were scored as 0 (‘does not apply’), 1 (‘applies sometimes’), and 2 (‘applies frequently’) (see Rutter, 1967).

**Assessment at the age of 10**

When the children were 10.25 years old ($SD = 0.69$), their parents completed the Achenbach Child Behavior Checklist (Achenbach & Edelbrock, 1978), with slightly modified wording for use in Mauritius. Only 89 items, which were equally applicable to
both sexes, were used (see Raine, Venables, & Mednick, 1997). There were three scales related to externalizing symptoms: Aggression (10 items), which includes items such as ‘gets into many fights’ and ‘cruel, bullying, and meanness to others’; Delinquency (10 items), including items such as ‘disobedient at home’ and ‘lies and cheats’; and Hyperactivity (nine items) with items such as ‘can’t concentrate’ and ‘impulsive and acts without thinking’. We created an overall externalizing scale with 29 items ($\alpha = .80$). Items were scored as 0 (‘not true’), 1 (‘somewhat or sometimes true’), and 2 (‘very true or often true’) (Achenbach & Edelbrock, 1978).

Results

Data analytic strategy

Aggregation
Aggregation was carried out by averaging across the weekly preschool observations. Thus, the Week 2 aggregate rating is the average of the ratings at Weeks 1 and 2; the Week 3 aggregate is the average of ratings at Weeks 1, 2, and 3; and so forth, up to the Week 15 aggregate, which is the average across all of the 15 weekly observations.

Reliability
In judgment studies, two types of reliability need to be distinguished. The correlation between the ratings made by two raters is the reliability of either rater, whereas the ‘effective reliability’ is the composite reliability of both raters (see Rosenthal & Rosnow, 2008, pp. 98–100). The latter can be calculated using Cronbach’s alpha or the Spearman-Brown prediction formula, also known as the Spearman-Brown prophecy formula. It is often used to predict increases in reliability as a function of intercorrelations between items, judges, or measurement occasions. In the current research, effective reliabilities were computed from the intercorrelations between weekly ratings using Cronbach’s alpha. As alpha cannot be computed for Week 1 alone, the reliability for Week 1 was estimated from the single measure intraclass correlation coefficient (ICC) computed over Weeks 1 and 2. It is worth noting that alpha as an indicator of consistency across repeated measurements has a different meaning from its more frequent use as an index of internal consistency, or the degree of interrelatedness, among questionnaire items. While a good internal consistency increases the likelihood of achieving consistency across repeated measurements, the latter cannot be extrapolated from the former.

Prediction from preschool ratings
The strength of associations between the preschool ratings and school-age externalizing behaviours was computed through the product–moment correlation coefficient ($r$).

Gains in reliability from aggregation over multiple measurement occasions
Figure 1a shows gains in alpha reliability from aggregation in aggression ratings across increasing weekly observations for both Rater 1 and Rater 2 separately. As noted earlier, the reliability for Week 1 was estimated from the single measure intraclass correlation
Figure 1. (a) Internal consistency reliability ($\alpha$) of ratings of preschoolers’ aggressive behaviour increases as more weekly observations are included in the aggregate score. Red line (circles): Rater 1. Blue line (triangles): Rater 2. (b) Increases in internal consistency reliability ($\alpha$) of the composite score of Raters 1 and 2 combined (filled circles, black) against the increases predicted from the Spearman-Brown prophecy (SB) formula (empty circles, light blue).
computed over Weeks 1 and 2, yielding values of .34 and .48 for Raters 1 and 2, respectively. As can be seen from the figure, the values increased steeply over the first weeks, reaching a maximum and near-perfect reliability of .97 (using Spearman-Brown reliabilities yielded virtually identical values). Because the ratings of the two teachers were sizably intercorrelated (average intercorrelation across the 15 weeks was $r = .48$), their ratings were averaged into a composite score for use in subsequent analyses. Figure 1b illustrates the actual gains in reliability from aggregation of this composite score as compared with theoretical gains in reliability that would be expected from the Spearman-Brown prophecy formula. The pattern for impatience was similar and can be obtained upon request.

**Gains in predictive power from aggregation over multiple measurement occasions**

The correlation between the Week 1 rating of aggression and externalizing symptoms at the ages of 8 and 10 was a mere $r = .07$ and $r = .10$, respectively ($ns$; see Figure 2, two lower lines). However, predictor-to-outcome correlation coefficients increased substantially by averaging across the multiple weekly preschool observations, reaching $r = .35$ ($p < .001$) and $r = .28$ ($p < .01$), respectively, for the assessments at the age of 8 and 10.

![Figure 2](image-url)
Aggregating across the two (standardized) outcome measures of externalizing behaviour at the age of 8 and 10 into a more broadly based composite score of school-age externalizing behaviour further strengthened the predictive correlation coefficient magnitude, which now reached an impressive $r = .41$ ($p < .001$) as can be seen from the top curve in Figure 2. Combining the outcome measures seemed justified by the significant correlation between both measures ($r = .34$, $p < .01$) and a satisfactory internal consistency of the composite scale ($\alpha = .79$). The results for the variable impatience were similar and are reported in the context of a different analysis below.

A potential alternative explanation to aggregation for the increases in reliability and validity reported in Figures 1 and 2 is that, with every additional assessment, there could also have been an increase in familiarity with the child's behavioural tendencies. When the study began, teachers had known the children already for several months (see Methods). This seems to reduce the probability that the steep increases in reliability and predictive validity occurring during the first weeks of observation could have been driven by increases in familiarity with the children's behavioural tendencies. Even so, it is possible that after the study began, teachers paid closer attention to the individual children's behavioural tendencies and might have become more accurate in their assessments as a result. If teachers became more accurate in their weekly assessment over time, the later assessments should exhibit stronger associations with the externalizing outcome, as compared with the early assessments.

There was no support for this expectation. Specifically, correlations between the final (or Week 15) rating of aggression and the externalizing outcomes were about as weak as the correlations between the first assessment and the externalizing outcomes ($r = .15$, ns, and $r = .14$, ns, with externalizing at the ages of 8 and 10, respectively). Using any of the ratings made during the final third of the study (Weeks 11–15) yielded similar findings. Thus, each of the later assessments seems to have entailed about as much error as the earlier ones. This interpretation is supported by the fact that, when we aggregated in reverse order, from Week 15 to Week 8 (the halfway point), there was a gradual rise in predictive association up to $r = .27$, $p = .02$, just as would be predicted from the aggregation principle. The results for impatience were similar and can be obtained from the authors upon request.

**Can gains in power from aggregation be estimated in advance?**

Issues of time, cost, and subject cooperation will often prevent researchers from carrying out numerous behavioural assessments. Therefore, knowing how much can be gained from each added assessment should be particularly helpful in conducting a cost-benefit analysis of using multiple rather than single-occasion measurements. Although the specific findings of this research are informative, a generalized procedure from which to estimate gains in power to be expected from each additional assessment would be preferable. This same procedure might then also be used to retrospectively estimate the extent of attenuation likely to have resulted from a single or a scant number of assessments.

A well-known means to remove the weakening effects of measurement error from a correlation coefficient is the correction for attenuation procedure, originally proposed by Spearman (Hunter & Schmidt, 2004; Muchinsky, 1996). Provided that the reliability of the correlated variables is known, one can estimate the correlation coefficient disattenuated from measurement error by means of the following formula:
\[ \rho = \frac{\operatorname{corr}(\beta, \theta)}{\sqrt{R_\beta R_\theta}} \]

where \( R_\beta \) and \( R_\theta \) denote the reliability of the two observed measures. Typically, the equation is used to correct the attenuation in the empirical correlation when the reliabilities of the variables are known; for example, if two variables are empirically correlated at \( r = .50 \) and their reliabilities are .60 and .70, the true correlation will be \( r = .77 \). The equation can also be rewritten to yield the empirical correlation from the true correlation and the reliabilities of the variables:

\[ \operatorname{corr}(\beta, \theta) = \rho \sqrt{R_\beta R_\theta} \]

Thus, when the true correlation is \( r = .77 \) and the reliabilities are .60 and .70, the expected empirical correlation ought to be \( r = .50 \) – representing an attenuation of 35% relative to the true correlation. Because in our study all quantities are known, it is possible to examine how well the rate of attenuation in the empirical correlation can be actually predicted from any reliability of the correlated variables. For instance, the observed correlations between preschool impatience and the school-age externalizing composite start at \( r = .20 \) (Week 1) and reach \( r = .34 \) (Week 9), when the alpha reliability of the preschool composite is at a near-perfect \( \alpha = .95 \) (see Figure 3, filled circles). As a

![Figure 3](image_url)

**Figure 3.** Increases in correlation magnitudes between preschool impatience and school-age externalizing behaviour, against predicted increases as derived from the correction for attenuation procedure (see main text).
correlation that is largely free from measurement error, the $r = .34$ figure may be seen as representing a value equivalent or close to the true correlation. From the attenuation equation as rewritten above, we can use the reliabilities at each weekly assessment (Figure 1) to predict how much of the true correlation will be reproduced depending on the number of weekly measurements, and compare the actual to the predicted values (Figure 3, empty circles).

As illustrated in Figure 3, the actual values can be predicted from this procedure with reasonable accuracy. The correlation between actual and predicted values for up to nine observations, when their respective maxima are reached, is $r = .91$. When elevation is taken into account by means of the ICC for absolute agreement, the resulting value drops to a still respectable ICC $= .79$ (the corresponding figures for aggression are $r = .88$ and ICC $= .66$). Differences between predicted and actual values may be due to limitations of Spearman’s correction for attenuation formula, such as its dependence on sample size and distribution of scores (e.g., Muchinsky, 1996; Zimmerman & Williams, 1997).

**Discussion**

In line with previous studies, the present research found preschool aggression and impatience to be linked with later externalizing problems (e.g., Caspi, Henry, McGee, Moffitt, & Silva, 1995). However, our results also indicate that single-occasion measures of aggressive or impulsive behaviours may underestimate their etiological or predictive significance by as much as 50%. This commands attention in the light of the widespread reliance on single-occasion behavioural measures of early childhood characteristics in long-term studies of psychological development.

As an example, consider the results of a recent meta-analysis about the continuity of attachment security. Continuity estimates of attachment security from infancy to late adolescence or later averaged a non-significant $r = .14$ (Pinquart et al., 2013). Most of the studies included in the analysis had however, used attachment measures obtained during a single observational session, many of which with either substandard or unknown test–retest reliability. From the empirical and psychometric analyses presented here, we would anticipate the true figure to be between $r = .20$ and $r = .30$, translating into a distinctly stronger case for the long-term stability of attachment security. Similar examples could be cited from the temperament literature, where effect–size correlations between behaviourally based measures of early childhood temperament and adolescent/adult outcomes in the domain of personality and psychopathology are typically small (e.g., Caspi et al., 2003; Kagan, Snidman, Kahn, & Towsley, 2007; see Zentner & Shiner, 2012, for an overview).

Although insufficient behavioural sampling has been identified as a source of attenuated validity correlations before, most studies focused on concurrent predictor-to-criterion correlations and found them to be two to five times larger after aggregation as compared with their pre-aggregation levels (e.g., Moscowitz & Schwarz, 1982; Rushton et al., 1983). The present study demonstrates that these figures generalize to the long-term prediction of disruptive disorders. This outcome is consistent with well-established statistical principles, and it is also plausible if one considers that consistently repeated behaviours are more likely to reflect a person’s characteristic disposition than those seen at any single point in time. This general principle may hold to a particular extent for young children whose behaviours are comparatively unstable (Roberts & DelVecchio, 2000).
Although the benefits of aggregation are manifest, carrying out multiple assessments may often be difficult. In considering practical issues, it is worth noting that one in-depth assessment also has its costs. In some cases, reallocating resources to conduct multiple but less onerous assessments might be a better strategy. The present study exemplifies such an ‘extensive’ approach as opposed to the more common ‘intensive’ approach. When adopting the former, the question of an optimal trade-off between costs and benefits of aggregation will inevitably arise. To address this question, we proposed a generalized procedure that may serve as a complementary tool for the analysis of predictive power associated with the frequency of behavioural measurement occasions.

Some caveats and limitations of the present study should be noted. First, findings need to be interpreted with caution, given the relatively small sample size. Unfortunately, it is difficult to find or create a data set from a much larger sample that has as many behavioural measurement occasions as the current one. Second, it would have been interesting to compare the present effects of aggregation across several weekly ratings with global or retrospective teacher or parent ratings of the same children’s preschool behaviour. Unfortunately, the latter were not available. It should be noted, however, that the current ratings were not behaviour frequency counts, but judgments of behavioural tendencies somewhat similar to those that informants would be asked to provide on questionnaires and rating scales.

Finally, the ratio of number of observations to reliability found in the current research does not necessarily generalize to other studies. This ratio will depend on the observational setting’s capacity to elicit a rate of the target behaviour from the child that is representative of his/her general base rate for that behaviour. The more representative this rate, the less pronounced the gains from aggregating across repeated observations will be. Still, even the most ingeniously designed and carefully conducted assessment procedure is vulnerable to transient error: That is, it will fail to capture a child’s natural disposition if she or he happens to behave uncharacteristically on the day of the assessment. This type of transient error, we should add, cannot be adequately quantified by means of single-administration indices of reliability such as internal consistency and requires period-sensitive measures of reliability instead. Not surprisingly, validity correlations in life-course research are more accurately predicted by test–retest than by single-administration reliability indices of the employed measures (McCrae, Kurtz, Yamagata, & Terracciano, 2011).

In conclusion, our data suggest that aggregation is a powerful determinant of behavioural stability estimates in long-term studies of psychological development. Researchers who wish to estimate the true stability or predictive significance of early childhood behaviours should therefore employ aggregation whenever possible. We hope that the present study can help researchers to strike a balance between assessment practicality and the precision needed to detect existing links between early childhood behavioural characteristics and later outcomes.

Acknowledgements

The authors would like to thank Prof Adrian Raine for his permission to use data from the Mauritius Child Health Project. The data collection was carried out with grant support from the Medical Research Council and the Danish International Aid agency and with support from the Mauritian Government.
References


Received 6 February 2014; revised version received 17 June 2014.