THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Validation of a brief self-report measure of adolescent bullying perpetration and victimisation

OPEN ACCESS

Validation of a brief self-report measure of adolescent bullying perpetration and victimisation: the Zurich Brief Bullying Scales (ZBBS)

Abstract

Although a wide range of measures of bullying have been developed, there remains a need for brief psychometrically supported measures for use in contexts in which there are constraints on the number of items that can be administered. We thus evaluated the reliability and validity of scores from a 10-item self-report measure of bullying victimisation and perpetration in adolescents: the Zurich Brief Bullying Scales (ZBBS). The measure covers social exclusion, property destruction, verbal and physical aggression, and sexual bullying in both traditional and cyber- forms. We evaluated factorial validity, internal consistency, developmental invariance, gender invariance, and convergent and divergent validity of the measure. Our sample was the normative longitudinal Zurich Project on Social Development from Childhood to Adulthood (z-proso) sample (n=1304). The study involved the administration of ZBBS to participants at ages 11,13,15 and 17. Strengths and weaknesses of the measure and recommendations for utilising and improving the measure were identified. Overall, results suggest that the items provide a reasonable general but brief measure of bullying victimisation and perpetration that can be used across early to late adolescence and in both males and females.

Bullying can be defined as harmful and repeated actions designed to cause fear, distress, or harm (Olweus, 1993) and is differentiated from peer aggression by the presence of an imbalance of power between the bully and victim. Bullying affects a substantial proportion of young people, with an estimated prevalence of around 36% for perpetration and 35% for victimisation (Modecki, Minchin, Harbaugh, Guerra, & Runions, 2014). As well as short-term distress, the long-term harms of victimisation can include mental and physical health problems, poorer educational attainment, employment problems, and difficulties forming and maintaining social and romantic relationships in adulthood (see Arsenault, 2018, for a review). On the perpetrator side, those who bully during their youth are at increased risk of serious anti-social behaviour and delinquency later in life (Ttofi, Farrington, Lösel, & Loeber, 2011). Efforts to illuminate the causes of bullying and best strategies to prevent it are thus a high priority within child development and related fields. The scale of research interest in bullying was indicated by a recent umbrella review by Zych et al. (2015) who retrieved and summarised 66 meta-analyses and systematic reviews on the topic.

The success of efforts to understand bullying and its prevention depend on the availability of measures that yield valid and reliable scores for both perpetration and victimisation. Psychometrically supported measures are, for example, crucial for ensuring that any positive (or negative) effects of anti-bullying interventions are accurately captured (e.g. Chalamandris & Piette, 2015). However, in the context of this considerable interest in bullying research and recognition of its practical importance, only one recent review has been devoted to documenting and comparing available instruments for measuring bullying. Vivolo-Kantor, Martell, Holland, and Westby (2014) presented an overview of the psychometric properties and contents of available multi-item measures of bullying. They systematically reviewed bullying measures administered to participants aged between 12 and

20 in the English language, developed or revised between 1985 and 2012, and for which psychometric data were available. Forty-one measures were included in their review.

Vivolo-Kantor et al. (2014) identified a number of strengths of available bullying measures. The psychometric properties of the reviewed measures were generally favourable. For example, the 90% of measures that reported internal consistency, showed high average reliability for victimisation (mean Cronbach's alpha = .84; SD=0.07) and perpetration scores (mean Cronbach's alpha= .82; SD=0.07). Validation studies published since have also generally reported strong psychometric properties for bullying questionnaires (e.g., Ahmed et al., 2018; Shaw, Dooley, Cross, Zubrick, & Waters, 2013).

The review authors also noted that almost all measures included both victimisation and perpetration items. This is important because it is not uncommon for empirical studies to show patterns of results that diverge, dependent on whether victimisation or perpetration is the focal outcome (e.g., Ttofi & Farrington, 2011). Further, a small but important minority of children can be classified as both victims and perpetrators of bullying, often labelled 'bully-victims'. These individuals tend to show higher levels of maladjustment than either pure bullies or pure victims (Haynie et al., 2001), as well as different patterns and higher levels of perpetration and victimisation than the corresponding 'pure' groups (e.g. Yang & Salmivalli, 2013). Distinguishing this group from other groups involved in bullying, of course, requires that perpetration and victimisation are assessed in a reliable, valid and parallel manner.

The review by Vivolo-Kantor et al., (2014) also noted that many measures now move beyond traditional conceptions of bullying as direct physical or verbal acts. For example, many now also include indirect forms of bullying such as social exclusion (included in 56% of measures reviewed), spreading rumours (37%), cyberbullying (24.4%) and sexual harassment (20%). These are understood to be important manifestations of bullying.

However, the review also highlighted important gaps, namely the need for measures that include both traditionally and newly recognised forms of bullying, including cyber- and sexual bullying; for measures that are brief enough to be administered in time and resource-constrained contexts; and for measures explicitly validated for use across males and females and across different stages of adolescence.

Only a minority of measures in the review by Vivolo-Kantor et al. (2014) included cyberbullying. Cyberbullying - which may include online behaviours such as flaming, harassment, cyberstalking, denigration, impersonation, outing, or trickery – has a victimisation and perpetration prevalence of around 15% (Cantone et al., 2014; Modecki et al., 2014). Cyberbullying conceptually fits with, and is moderately correlated with traditional forms of bullying (r = .47 and .41 for perpetration and victimisation respectively; Modecki et al., 2014) and has similar impacts on the victim (e.g. Cassidy, Faucher, & Jackson, 2013). Although specific measures of cyberbullying are available, there is arguably a strong rationale for including cyberbullying in general bullying measures so that they can be measured on a comparable scale (e.g. see Thomas, Connor & Scott, 2015 for a discussion),

Even fewer of the measures reviewed by Vivolo-Kantor et al. (2014) included a measure of sexual bullying. There is some debate as to whether sexual bullying, which may also be labelled 'sexual harassment', should be included in measures of bullying (e.g. Miller et al., 2013). One argument against this is that it risks losing the important gender dimension to sexual bullying (e.g. Gruber & Fineran, 2008) and lumps legal behaviours (bullying) with illegal (sexual harassment). However, as we argue below, gender is relevant for other forms of bullying as well, not only sexual bullying. Further, irrespective of their legal status, sexual bullying and other forms of bullying are similarly harmful to the victim (e.g., Arsenault, 2018; Gruber & Finegan, 2008). The value of including sexual bullying in general measures of bullying is at least partly an empirical question: if sexual bullying strongly co-occurs with

other forms of bullying, this speaks in favour of its inclusion in general bullying measures. Some studies suggest that those who are the subject of verbal and physical bullying are at greater risk of sexual bullying (e.g. Gruber & Fineran, 2008). However, from a psychometric perspective, it is useful to know whether sexual bullying loads on a common factor with other forms of bullying. This would provide evidence in favour of including sexual and non-sexual bullying items in the same scales.

Further, although the review by Vivolo-Kantor et al. (2014) identified some brief measures of bullying (~10 items or less), the average number of items was 27.4, suggesting that most available measures are considerably longer. The availability of brief but psychometrically robust measures is essential for contexts in which there are constraints on the administration of longer tools. Brief measures are, for example, valuable in cohort studies in which a broad range of constructs must be measured, or in intervention studies with regular monitoring, in which there are concerns about participant burden and attrition. They may be especially valuable in newly emerging intensive longitudinal designs in which measures may be completed once a day or more. In educational contexts, brief measures may be useful as initial screens; with follow-up with comprehensive assessment indicated for respondents who show evidence of victimisation or perpetration.

Only the minority of brief measures of bullying discussed in the review by Vivolo-Kantor et al. (2014) or published since, provided a measure of general bullying likely to be suitable for these purposes (e.g., Gottheil & Dubow, 2001; Shaw et al., 2013). Others either focussed on specific forms of bullying (e.g., Eisenberg, Neumark-Sztainer, & Perry, 2003; Gable, Ludlow, McCoach, & Kite, 2011; Poteat & Espelage, 2005), or used items where the concept of bullying was not clearly differentiated from peer aggression (Bosworth, Espelage & Simon, 1999; Orpinas, & Frankowski, 2001). These measures are likely to be valuable in

the contexts for which they were designed; however, they are unlikely to be optimal as brief general measures of bullying.

Finally, although many validation studies of scores from measures of bullying address some fundamental psychometric properties such as reliability, criterion validity, convergent validity, and face validity, there has been a lack of attention paid to gender and developmental invariance. Gender invariance refers to equivalent psychometric functioning of a measure across males and females, such that the observed distribution of scores given latent bullying levels are independent of gender (e.g., Millsap, 2012). It is an important assumption when making comparisons across males and females. When there is a lack of gender invariance, levels of bullying could, for example, be underestimated for one gender as compared to the other, or the construct captured by the measure could vary in meaning altogether across gender. Given that males and females show different levels of bullying (e.g., Mitsopoulo & Giovazolias, 2015) and different patterns of aggressive behaviour more generally (e.g., Archer, 2004) there is a strong possibility that gender invariance could be violated for measures of bullying. Shaw et al. (2013) tested this idea, examining gender invariance in the 10-item scale: The Forms of Bully Victimisation (FBS-V) and Perpetration (FBS-P) scales. They found that within a multi-group confirmatory factor analysis model, items measuring relational and physical bullying varied in their factor loadings and thresholds across males and females. Results were in line with the idea that males are relatively more likely to use physical forms, and females to use relational forms of bullying (e.g., see Winter & McKenzie, 2017). The magnitude and direction of the effects were such that failing to model the lack of gender invariance would have led to a relative underestimation of victimisation in males and a relative under-estimation of perpetration in females. The results thus highlight both the value of testing gender invariance to understand the different

manifestations of bullying across males and females and the importance of doing to so avoid bias in empirical studies of bullying.

There has been equally scant attention paid to the developmental invariance of bullying measures. Following the same logic as gender invariance, developmental invariance implies that a measure functions equivalently irrespective of developmental stage, such as across early and late adolescence (e.g., Murray, Obsuth, Eisner, & Ribeaud, 2017). Developmental invariance is particularly important in the context of longitudinal studies which seek to understand bullying trajectories. At present, evidence on age differences in bullying are equivocal (e.g. see Zych, Ortega-Ruiz, & Del Rey, 2015 for a review); however, without ensuring the comparability of measures over development, conclusions about developmental trends are difficult to draw. Rosen, Beron, and Underwood (2013) examined the developmental invariance of a measure of peer victimisation across grades 7,8,9 and 10 and concluded that the measure was developmentally invariant. This allowed them to compare mean levels of overt and social aggression over development, finding no developmental change in the former but an increase followed by a decrease in the latter.

It should be noted that in the context of gender and developmental invariance, it is not necessary for every item to be identical nor invariant. Rather, as long as there is a small core of invariant items, the latent bullying variables could usually be put on a comparable metric (e.g. Van de Schoot, Lugtig, & Hox, 2012). However, it is important to identify which items are and are not invariant in order that this can be appropriately modelled.

In sum, there remains a need for brief psychometrically validated measures of bullying that include cyber-, sexual, direct and indirect forms of bullying, and that can be administered to both male and female respondents of a range of ages. We thus report on the validation of a 10-item measure of bullying victimisation and perpetration: the Zurich Brief

Bullying Scales (ZBBS). We evaluate factorial validity, gender and developmental invariance, and convergent and divergent validity of the ZBBS scores in a large normative sample of youth aged 11 at baseline and assessed again at age 13, 15 and 17. Convergent and divergent validity is assessed via the evaluation of the nomological network of ZBBS scores (Cronbach & Meehl, 1955). A nomological network can be defined as a set of theoretical relations among constructs and correlational analyses can be used to determine whether the observed scores from a measure have a pattern of relations with scores from other measures that match the predictions of the relevant nomological network. For example, it can be expected based on a large body of past research that a valid measure of bullying perpetration will show positive relations with measures of aggression, violence and substance use (e.g.,Ttofi, Farrington & Lösel, 2012; Ttofi, Farrington, Lösel, Crago, & Theodorakis, 2016) but a negative relation with empathy and helping (e.g., Gini, Albiero & Altoè, 2007). On the other hand, a valid measure of victimisation should show positive associations with measures of internalising (e.g., Ttofi, Farrington, Lösel & Loeber, 2011a). We hypothesised that the new ZBBS measure could provide a good solution to the need for brief, psychometrically supported measures of bullying victimisation and perpetration that reflect the diversity of bullying behaviours recognised as potentially important in contemporary bullying research.

Method

Ethics

Given the minimally intrusive nature of the study design, questions and interventions, as well as the focus on social science research questions, the relevant Ethics Committee of the Canton of Zurich issued, based on the Swiss Human Research Act, a "declaration of no objection" for the Zurich Project on Social Development from Childhood to Adulthood (z-proso) project, from which data for the current study is derived. It states that the project falls

outside the remit of the Ethics Committee of the Canton of Zurich, and furthermore declared z-proso as ethically unproblematic.

Participants

Participants were from the Zurich project on Social Development from Childhood to Adulthood (z-proso); a longitudinal cohort study of child and adolescent development with a particular focus on criminal and aggressive behaviours. The first wave of the study was in 2004 when the children were aged 7 and in the first grade. Children were invited to take part based on belonging to one of 56 schools in Zurich, selected based on a stratified random sampling procedure, with stratification by school size and location. The current study concerns the data from the measurement waves when the participants were (median) aged 11, 13, 15 and 17, corresponding to the 4th through to 7th main data collection waves. These are the waves at which self-reported bullying perpetration and victimisation data were collected.

The numbers of participants providing bullying data at each wave is provided in Table S1 of Supplementary Materials. To contextualise these, the initial target sample was 1675 and the number of participants providing data at any of the 7 currently completed waves was 1572. A comprehensive analysis of non-response and drop-out is reported in Eisner, Murray, Eisner, and Ribeaud (2018). In brief, there is little evidence that responders differ systematically from non-responders, the main exception being that youth whose parents do not speak German as their first language are slightly under-represented. Thus, z-proso can be considered approximately representative of the underlying same-aged population. Comprehensive details of z-proso more broadly; its recruitment, assessment, measures and previous findings, can be found in previous publications (e.g., Eisner et al., 2018) and at the study website (http://www.jacobscenter.uzh.ch/en/research/zproso/aboutus.html).

Measures

Zurich Brief Bullying Scales (ZBBS)

Rather than shortening a specific existing scale to meet the need for brief measures of bullying, a new amalgamated scale, the ZBBS, was adapted from multiple sources. This was in order or ensure that the main forms of bullying identified as important in contemporary research were covered, while also addressing the gaps in existing measures identified in the Introduction and other issues such as the fact that not all bullying occurs within the school context. Items were drawn and adapted from several sources. These include a previous series of surveys conducted in German schools by the Criminological Research Institute of Lower Saxony (KFN) (Wetzels, Enzmann, Mecklenburg, & Pfeiffer, 2001), the earlier waves of z-proso, which used a picture-based questionnaire capturing similar behaviours to those in the self-report items (Alsaker, 2012) and the Zurich Youth Study (Eisner, Manzoni & Ribeaud, 2000). The KFN survey used an adaptation of the scales developed by Olweus (1996). This was then used as the basis for the instrument in the Zurich Youth Survey (Eisner, Manzoni, & Ribeaud, 2000). Separately, a picture-based bullying scale developed by Alsaker for use in younger children was administered at the age 8 wave of z-proso. At the age 11 wave of z-proso, the Zurich Youth Survey and age 8 wave of z-proso item sources were combined into a 4-item paper and pencil self-report questionnaire (where the picture-based items were adapted to written form). From the age 13 wave on an item on sexual bullying from the Zurich Youth Survey was added by the z-proso team to reflect the emerging importance of this behaviour in this age group. Adaptations in the ZBBS version as compared to the original administrations of the items included lengthening the recall period to the previous 12 months, broadening the contexts referred to from the school to any context, amending the introductory text, and increasing the number of response options from 5 to 6.

The items were selected and administered based on recommendations developed in previous studies. First, although teacher reports were also collected, self-reports were the

main focus over other informant reports (e.g., peer or teacher reports) because there are several aspects of bullying behaviour that are difficult to ascertain from an observer's perspective (Furlong, Sharkey, Felix, Tanigawa, & Green, 2010). These include the intentionality of harm, the power imbalance, and behaviours which are deliberately covert to avoid sanctions. Second, the term 'plagen' in German was used, which does not translate directly to using the term 'bullying' but has a similar effect as it implies a power imbalance. In translating and adapting the measure for use in a context in which Swiss German is spoken there was no unequivocal one-to-one translation for the word 'bullying' used in the introductory text, but two main choices: 'Mobbing' and 'Plagen'. The latter was chosen as the former is a rather technical neologism, used by professionals rather than in everyday speech among adolescents. Thus, it was suggested that the former could be misunderstood by younger adolescents and those with low linguistic skills. Its use could also promote under-reporting as it connotes an act of a higher level of seriousness than implied by the term 'plagen'. It was thus judged that 'plagen' provided the best conceptual equivalent to 'bullying' in English. While some have raised concerns that using the term 'bullying' (or an equivalent term) could lead to under-reporting (e.g., Kert, Codding, Tyron & Shiyko, 2010), a study by Ybarra, Boyd, Korchmaros, and Oppenheim (2012) demonstrated that using the term minimised the mis-classification of students.

However, a definition of bullying was *not* provided over and above the behavioural exemplars encoded in the items. This is because our goal was to keep the measure brief and previous studies have suggested that the addition of a definition makes no difference to reported prevalence rates (Huang & Cornell, 2015; Ybarra et al., 2012). Third, both perpetration and victimisation were measured and in a parallel format to ensure their comparability. Victimisation was presented first, in an effort to increase the reporting of perpetration. Finally, the items were selected to cover the range of forms of bullying that

have been delineated and are considered important forms of perpetration and victimisation: verbal, physical, indirect, relational, cyber (specifically, through the internet), and sexual. The questionnaire (in English and German as administered to participants) is provided in Appendix I of Supplementary Materials.

Bullying victimisation was measured using 5 items. A brief explanation introduced the items: 'This part is about **bullying [Plagen]**. Adolescents can be quite mean to each other sometimes. How about you? In the last year, i.e. **since [date of previous wave]**, have you been bullied by other adolescents? *This could be, for example, at school, on the way to school, when out in the evening, at home, or on the internet.'* In the measurement wave at age 11, 'kids' was used instead of 'adolescents'. In addition, at age 17 'in the workplace' was added as another example of a context in which the bullying might take place. The five items referred to being purposely ignored or excluded; laughed at, mocked or insulted; hit, bitten, kicked or having hair pulled; having possessions stolen, broken or hidden; and being sexually harassed (hit on, groped). The sexual bullying was not included at age 11. Responses are provided on a 6-point scale measuring the frequency with which each form of victimization occurred. Response options were:1 = *never*, 2 = *1 to 2-times*, 3 = *3 to 10-times,* 4 = *about once a month,* 5 = *about once a week,* and 6 = *(almost) every day*.

Bullying perpetration was measured by 5 items which immediately followed the victimization items. The wording used was the same as for victimization but here the actions are referred to as being perpetrated against other youth. As for the bullying victimization items, the item on sexual bullying was not included in the measurement wave at age 11. The item response formats were identical to those for bullying victimization.

Measures to construct a nomological net.

To assess the convergent and divergent validity of the scores from the bullying measures we included a number of additional measures from the age 17 wave of z-proso. *Proactive aggression*, *reactive aggression, prosociality and internalising* were measured using the self-reported Social Behaviour Questionnaire (SBQ; Tremblay et al., 1991). Proactive aggression was measured using the mean of 4 items (scaring, bossing, humiliating, and threatening others to get one's way). Reactive aggression was measured using the mean of 4 items (responding aggressively when teased, insulted, having something taken away, and when not getting something they wanted). Prosociality was measured using the mean of 10 items referring to a range of helping and empathic behaviours. Internalising was measured using the mean of 9 items measuring anxiety and depression. All responses to SBQ items were on a five -point Likert-type scale ranging from *Never* to *Very Often*. The validity and reliability of the SBQ scores in the current sample have been supported in previous studies (Murray, Obsuth, et al., 2017; Murray, Eisner, & Ribeaud, 2017), building on psychometric evaluations in other samples (e.g., Tremblay et al., 1991). Murray, Eisner & Ribeaud (2017) found that the teacher-reported SBQ scores showed adequate reliability for a wide range of phenotypic values. Murray, Obsuth et al. (2017) evaluated the developmental invariance of the self-reported SBQ scores, finding that all scales used in the current analysis showed at least metric invariance over ages 11,13,15 and 17. Murray, Eisner, Obsuth & Ribeaud (2017a) provided evidence for the factorial validity of the SBQ scores, with the vast majority of items loading on the intended factors. Omega reliabilities in the current sample for the SBQ subscale scores were for proactive aggression =.74, reactive aggression = .66, prosociality = .87, and internalising=.85.

Substance use was measured using 4 items capturing alcohol use (with separate items for 'beer-like' versus 'spirit-like' drinks), cigarette smoking, and cannabis use. Items were introduced with: 'Listed below are some drugs, intoxicants and other substances. Have you

ever taken any of them and if yes, how many times in the last 12 months (i.e. since

[DATE])?'. Participants indicated the frequency of the use of each substance over the

previous 12 months on a 6-point scale from *never* to *daily*, specifically: ('never', 'once', '2 to

5 times', '6 to 12-times (monthly)', '13 to 52 times (weekly)' and '53 to 365 times (daily)'.

As they are used as single items it is not possible to compute internal consistency for the

substance use scores. However, past research suggests that they show expected patterns of

association with other construct scores (e.g., Murray, Eisner, Obsuth & Ribeaud, 2017b).

Criminal violence was measured as the mean of 4 items referring to engagement in

criminal acts: weapon carrying, extortion, robbery over the previous 12 months. Responses to

these items were on a binary *yes* versus *no* scale. Omega reliability (based on tetrachoric

correlations) for these scores was .91.

All measures were administered in paper and pencil format in German; the official

language of the study location. They were part of a larger questionnaire measuring constructs

related to antisocial and prosocial behavior and other dimensions of psychosocial functioning.

Statistical Procedure

Factorial validity and reliability.

We assessed the unidimensionality of the victimisation and bullying item scores at

each measurement wave in males and females separately by fitting single factor confirmatory

factor analysis (CFA) models. Well-fitting models by conventional fit criteria (CFI>.95,

TLI>.95, RMSEA<.05; with RMSEA<.08 suggesting 'acceptable fit'; Hu & Bentler, 1999;

Schermelleh-Engel et al., 2003) were taken as evidence for sufficient unidimensionality.

Modification indices (MIs) and expected parameter changes (EPCs) for poor-fitting models

were examined and modifications made whenever justifiable based on theory. Single factor

models were fit using lavaan within R statistical software and invariance models were fit in

Mplus (Muthén & Muthén, 2010) using robust maximum likelihood estimation (MLR). As our items had 6 response options this method was preferred over using a categorical estimator such as WLSMV. MLR provides better treatment of missing data (WLSMV uses pairwise deletion) and is generally more appropriate when there are more than 6 categories (Rhemtulla, Brosseau-Laird & Savalei, 2012). Reliability at each wave for males and females was estimated from the final models using the omega coefficient.

Gender and developmental invariance.

Gender and developmental invariance were tested using the final models from the previous stage as the basis for the configural model. We began by fitting the configural model where other than the minimal constraints necessary for identification, parameters were free to vary across time and gender. Specifically, the mean and variance of the female victimisation (or perpetration) factor at age 11 were fixed to 0 and 1 respectively and the loading and intercept of the first item fixed equal across both gender and time. All other latent factor means and variances were freely estimated. Residual covariances between the same itemmeasured at different time points were included in the model.

If the configural model fit well by conventional criteria, we added metric constraints (equal loadings), and then scalar constraints (equal intercepts) constraints over time and gender. Metric and then scalar invariance constraints across gender and development were added simultaneously. Metric invariance was judged to hold if the addition of metric constraints resulted in a CFI decrease of no more than 0.010, an RMSEA increase of no more than 0.015, and an SRMR increase of no more than 0.030 (Chen, 2007). Scalar invariance was judged to hold if the addition of scalar constraints resulted in a CFI increase of no more than 0.010, an RMSEA increase of no more than 0.015, and an SRMR increase of no more than 0.010 (Chen, 2007). Missing data in all of the above-described analyses were handled

using full information maximum likelihood estimation (FIML) which provides unbiased parameter estimates provided that data are missing at random (MAR).

Nomological network.

Correlation analysis was used to assess the relations between bullying and victimisation scores and other construct scores to which they could be assumed to be differentially related. To aid interpretation of the pattern of associations, correlations between variables were visualised as a network using the qgraph function in R Statistical Software (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012; R Core Team, 2018).We hypothesised that perpetration scores would be positively correlated with proactive aggression, reactive aggression, criminal violence, and substance use scores and negatively correlated with prosociality scores (e.g. Gini, Albiero, Benelli, & Altoè, 2007; Ttofi, Farrington, Lösel & Loeber, 2011b; Ttofi, Farrington, Lösel, Crago, & Theodorakis, 2016). We hypothesised that bullying victimisation scores would be positively correlated with internalising scores (e.g. Arseneault, 2018). Missing data in these analyses were handled using pairwise deletion, which provides unbiased parameter estimates only when data are missing completely at random (MCAR).

Results

Descriptive Statistics

Descriptive statistics are provided in Table S1 of Supplementary Materials. For victimisation, data were skewed towards low levels; however, respondents did utilise the full range of the scale. For perpetration, the skewing towards low levels was more marked. The full range of the response scale was utilised for most items at most waves; however, the highest response option was not selected by any participant for items 3 and 4 at age 17.

Factorial validity and reliability

Model fits for the single factor CFAs at each time point for males and females suggested that the victimisation models mostly fit well; however, some models had RMSEA and/or TLI values outside the range generally accepted to represent good fit. We considered whether parameter estimates and MIs/EPCs pointed to any areas of local mis-fit. These highlighted that the first two items in the scale were more correlated than the others and pointed to the inclusion of a residual covariance between these items. Given that these items both measure social forms of aggression, we judged that there was a theoretical basis for including these residual covariances. Model fits for the modified models including these residual covariances are provided in Table S2 of Supplementary Materials.

With only a couple of exceptions, the perpetration models initially fit poorly. Examining parameter estimates and MIs/EPCs suggested that again the first two items in the scale were more strongly correlated with one another than the remaining items. Fit was improved by including a residual covariance between these items in all models. Following this, all models had fits that could be considered good by conventional criteria, except the TLI value for the female age 17 perpetration model. Model fits for the modified models are provided in Table S2 of Supplementary Materials.

Omega coefficients are also provided in Table S2 of Supplementary Materials. For males these were generally >.70 with the exception of perpetration at age 17. For females, however, omega values were all <.70 and as low as .39 for perpetration at age 17. The items thus appeared to show higher internal consistency for males than for females.

Gender and developmental invariance

Victimisation.

For victimisation, the multi-group longitudinal model with no cross-group nor cross-time constraints beyond the necessary minimum for identification fit well (CFI = .956,

RMSEA = .040, SRMR = .041). The addition of metric invariance constraints across both group and time resulted in a substantial worsening of fit (CFI = .925, RMSEA = .049, SRMR = .065). Guided by MIs and EPCs, loading constraints were removed on item 3 at age 11, 13, and 17 to achieve partial metric invariance (CFI = .948, RMSEA = .041, SRMR = .050). Adding scalar invariance constraints to this model resulted in a substantial worsening of fit (CFI = .891, RMSEA = .057, SRMR = .069). Partial scalar invariance was achieved by removing the constraints on the intercept of items 1,4 and 5 at age 17, item 5 at age 13 and 15, and item 1 at age 15 (CFI = .939, RMSEA = .044, SRMR = .051).

Parameter estimates from the final model are provided in Table 1. In terms of gender differences, loadings tended to be higher in males than in females. For example, item 3 (physical bullying) had a loading of 0.38 in females across time but a loading in males of 0.76 and 0.60 at ages 11 and 13. When there were gender differences in intercepts, sometimes the intercept was larger for males and sometimes for females. For item 3, for example, male intercepts tended to be larger; however, for item 1 (social exclusion) female intercepts tended to be larger. In terms of developmental differences, the main change was in item 3 for males. Its loading tended to decrease over time along with its intercept.

Perpetration.

The configural model for perpetration fit well (CFI = .954, RMSEA = .039, SRMR = .047); however, the addition of metric constraints resulted in a substantial deterioration in fit (CFI = .927, RMSEA = .047, SRMR = .062). Releasing loading constraints on items 3 and 4 at age 11, item 5 at age 15, and items 2,3 and 5 at age 17 allowed partial metric invariance to be achieved (CFI = .944, RMSEA = .042, SRMR = .055). Adding scalar constraints resulted in a worsening of fit (CFI = .929, RMSEA = .046, SRMR = .058) with partial scalar

invariance achieved after the release of constraints on the intercepts of item 2 at age 11 and item 3 at age 13 (CFI = .935, RMSEA = .044, SRMR = .057).

Parameter estimates from the final model are provided in Table 1. In terms of gender differences, the loadings for item 3 (physical bullying) – as for victimisation - tended to be larger in males. There were, however, several examples of higher loadings in females: for item 4 (breaking/stealing possessions) at age 11, item 5 (sexual harassment) at age 15 and item 2 (mocking) at age 17. Where there were gender differences in intercepts, males tended to show a higher intercept. In terms of developmental differences, the loading of item 3 decreased between age 11 and 17 in both males and females. The intercept of this item also decreased over time in both genders.

Nomological network

The correlation matrix of bullying victimisation and perpetration and hypothesised correlate scores is provided in Table S3 of Supplementary Materials and visualised in Figure 1. Thicker lines (edges) represent stronger relations between variables (nodes). Bullying perpetration scores were significantly associated with all variable scores except internalising (negatively with prosociality). As hypothesised, bullying victimisation scores were significantly associated with internalising scores. They were also significantly associated with other predictor scores but the correlation was weaker than their correlation with bullying perpetration.

Discussion

In this study, we evaluated a brief 10-item measure of general bullying, the ZBBS, designed to cover the range of bullying behaviours that are recognised as potentially important to assess in adolescence. We found that the internal consistency of the scores was sometimes lower than ideal, but convergent and divergent validity were supported. There

were some violations of gender and developmental invariance, revealing how bullying differs in manifestation between males and females and across different stages of adolescence. Specifically, females are relatively more likely to perpetrate and be victimised via social exclusion and males via physical aggression. Males also appear more likely to admit perpetrating bullying. Finally, physical aggression becomes a less relevant form of bullying at later stages of adolescent development.

Internal consistency values varied across perpetration and victimisation, and across gender and age. In many cases, internal consistency fell below levels conventionally considered acceptable (0.70). The brevity of the scales (only 5 items each) together with the fact that items were selected to sample a range of different bullying behaviours and that we assumed continuous item distributions in our modelling are the most likely explanation for the sometimes-low internal consistencies. The set of items selected represented a trade-off between content validity and internal consistency, as higher internal consistency could have been achieved by selecting 5 similar items but at a cost to ensuring that the range of bullying behaviours highlighted as important in previous research was covered. Further, we explicitly took steps to avoid the inflation of reliability by including residual covariances between the two items that were somewhat similar, which means our analyses are likely to yield lower reliability estimates than studies that have not taken this precaution.

That said, the sexual bullying item consistently showed low loadings in both males and females and for both perpetration and victimisation, making it difficult to justify its inclusion. This supports the (debated) view that sexual bullying, though correlated with other forms of bullying, represents a distinct phenomenon. McMaster, Connolly, Pepler, and Craig (2002), for example, describe a two-factor model of sexual bullying. Here sexual bullying can represent either crude attempts to show developmentally appropriate sexual interest or it can represent an expression of hostility. While the former is presumed to be most commonly

directed towards the opposite gender, the latter is presumed to be typically directed towards a victim of the same gender and potentially involves homophobic content. The first form of sexual bullying arguably serves a different function from traditional bullying and would, therefore, not necessarily belong in a general bullying scale. The second form is conceptually more closely related to other bullying behaviours. We believe the current ZBBS sexual bullying item does not clearly differentiate the latter form from the former. We thus recommend that future iterations of the scale either refine this item or consider replacing it with an additional non-sexual bullying behaviour item.

Internal consistency tended to be lower in females, suggesting that the exploration of female-specific items to improve reliability for females may be merited. Items need not be identical across males and females to facilitate cross-gender comparisons provided that a small core of items can be shown to be gender invariant and provide an anchor through which gender-specific variations can be put on a common metric.

Our gender invariance analyses provide some suggestions as to which items could be selected to capture bullying in females. These suggested that for the same 'level' or 'severity' of bullying, females were relatively more likely to use and experience social exclusion while males were relatively more likely to use and experience physical aggression. This is similar to a previous study that found, for the same level of bullying severity, relational bullying and physical bullying were more common in females and males respectively (Shaw et al., 2013). Thus, to achieve measures that are optimally calibrated to detect bullying in females, social aggression items could be prioritised for administration over physical aggression items.

Our invariance analysis also identified changes over development, in particular, a decrease in the loading of physical aggression over time. This suggests that physical

aggression becomes a poorer marker of overall bullying severity over time. In this case the loadings remained strong enough to justify the inclusion of the item even in measures administered to older adolescents. However, these and the other violations of measurement invariance identified necessitate the use of a partial invariance measurement model when using the items across genders and developmental stages. Partial measurement invariance refers to the situation where some items are invariant but others are not. There is some debate around how many items need to be invariant for a valid comparison; probably because in practice the amount of invariance that is tolerable depends on the purpose of the analysis as well as the magnitude of the violations in the context of the magnitude of the group differences in latent variables. At a minimum, two items per latent factor need be invariant for valid comparisons, ideally more. Further, the level of invariance required depends on whether there is a need to have latent variances and covariances on a common scale across gender or development (requiring metric invariance), or latent means (requiring scalar invariance) (e.g., see Meredith & Teresi, 2006). Although much attention has been paid to the psychometric validation of measures of bullying (e.g., Vivolo-Kantor et al., 2014), gender and developmental invariance are not routinely evaluated (see Rosen et al., 2013; Shaw et al., 2013 for exceptions). Given the importance of identifying invariance to avoid potential bias, and the value of evaluating invariance for identifying gender and developmental differences in bullying manifestations, we recommend that that more bullying psychometric studies include these analyses in their validation protocols.

Finally, the scores showed good convergent and divergent validity with scores on proactive and reactive aggression, substance use, criminal violence, prosociality and internalising. Perpetration scores were more strongly correlated with the first five variables (negatively with prosociality; all in the small to moderate range) than was victimisation, . On the other hand, victimisation scores were more strongly correlated with internalising scores

(*r*=.32). These results are consistent with expectation. Past research has suggested that bullying perpetration is associated with lower prosociality and higher aggression and criminality (e.g., Gini et al. 2007; Ttof et al., 2011; Ttofi et al., 2016), while victimisation is particularly associated with internalising problems (e.g., Swearer, Song, Cary, Eagle, & Mickelson, 2001). The pattern of correlations observed in the current study are also similar to the convergent and divergent validity correlations reported for more comprehensive bullying measures such as the Forms of Bullying Scale (Shaw et al., 2013). This suggests that the brevity of the ZBBS does not undermine its convergent and divergent validity as compared to longer measures of bullying perpetration and victimisation.

Collectively, the present results provide some support the use of the scale in contexts where resource or time constraints preclude the use of comprehensive assessments. For example, the scales could be used as brief screens in school contexts, with full assessment indicated for adolescents scoring in the top percentiles. Similarly, when cases of bullying have been identified, they could be used to monitor progress on a more frequent basis than may be possible using more comprehensive assessments. In research contexts, the items would be useful in broadband studies of adolescent psychosocial functioning, where the assessment of bullying will have to compete with a wide range of constructs. Other situations where the brevity of the scales would be a major advantage include experience sampling studies of bullying, or in intervention studies in which bullying is a primary or secondary outcome of interest. Both involve repeated measurement of outcome variables and can thus involve significant burden to participants if the number of items is not kept to a minimum.

Given the gender and developmental differences in scale functioning identified and sometimes low reliabilities, in research contexts, scores on the instrument are best derived using a latent variable model that allows disattenuation for unreliability and non-invariant parameters to vary across gender and development.

Limitations

A limitation of the current study is that in using archival data we were unable to iterate the items within the sample to test progressive revisions of the scale. Thus, we were unable to test our hypotheses about how to improve the measure. Further, there were aspects of reliability and validity we were unable to test using archival data, such as test-retest reliability and correlations with a comprehensive gold standard measure. Our CFA also suggested potential multidimensionality in the instrument. Specifically, two items measuring social aggression were correlated with one another over and above their correlation with the bullying latent factor. Residual covariances were included between these items to avoid inflating the apparent reliability of the test scores. However, as these residual covariances were not specified a priori but determined based on modification indices, it would be useful to evaluate whether they replicate in future studies to ensure that their inclusion did not represent capitalisation on chance.

Finally, the lack of one to one translation for the word 'bullying' potentially complicates the cross-cultural use of the instrument and will require future research. In particular, it will be important to establish whether variations across different language versions of the scale in the seriousness of the connotations of the term for 'bullying' used are related to relative over- or under-reporting. This also applies to the behavioural descriptors in the items. For example, a recent study examining the cross-cultural invariance of the scale in Switzerland and Uruguay (where there is also no direct translation for the word 'bullying') suggested some violations of metric invariance across these societies (Kaiser et al., 2018). It is possible, given the difficulty of finding exact conceptual translations across languages, that linguistic differences played a role.

When translating scales intended to measure a concept such as bullying, it is necessary to consider that the meanings attributed to the construct may not fully overlap across cultures, and that the construct might be more differentiated in one culture than in another (Chen, 2008). Consequently, certain behavioural expressions of the construct described might not be appropriate for the cultural setting where the translated version of a questionnaire is to be applied (Behling & Law, 2000). Hence, future research intending to adapt the bullying questionnaire to new cultural settings should examine the extent to which the concept and its descriptors are construed in a similar manner and are equally relevant across these cultural contexts. Two practices from the field of cross-cultural adaptation of measurement instruments are well-suited for this purpose. First, researchers may draw from the expertise of academics familiar with the construct of bullying and the cultural setting where the questionnaire will be applied in order to explore the meaning, connotations and behavioural expressions of the construct of interest in the target language. Based on this assessment, researchers can determine which items are culturally relevant and should therefore be preserved and which ones ought to be adapted (Berry et al., 2002; Flaherty et al., 1988). Second, cognitive interviewing can be used to assess the appropriateness of a translated questionnaire and identify linguistic variations across original and translated versions (Beatty & Willis, 2007; Willis & Miller, 2011). By conducting cognitive interviews on the translated versions of questionnaires, researchers can gauge what respondents understand by bullying - or the term chosen for its translation - and thus explore the appropriateness of employing this concept and its measurement instrument in a new setting (Chan & Pan, 2011). Furthermore, they may assess whether respondents consider that the behaviours included in the questionnaire as indicators of bullying are actually indicative of this concept in their cultural context, and whether there are behavioural expressions of bullying which are relevant for their cultural setting which are not present in the

questionnaire. Aside from its use in z-proso, the ZBBS has been translated into and administered in large-scale surveys in other languages, including Spanish and Brazilian Portuguese. An English-language version of the ZBBS is provided in Supplementary Materials. Further validation of the ZBBS in these languages represent important future directions in the validation of the ZBBS as an international measure of bullying.

Conclusion

Overall, the 10-item ZBBS measure of bullying presented in the current study can be considered a good brief measure of general bullying for use in adolescent populations. It covers a range of bullying behaviours, including both cyber- and traditional and direct and indirect forms of bullying. It may, therefore, be useful in contexts in which there are significant constraints on the number of bullying items that can be administered and where more comprehensive measures are thus not feasible. However, improvements could be made by replacing or modifying the sexual bullying item, and by identifying and including another item that is better calibrated to detect bullying behaviour in females. Further, partial invariance models should be used in empirical analyses that utilise the items to account for gender and developmental differences in the functioning of the measure. For practical applications, age- and sex- adapted scoring algorithms could be used to the same end.

References

Ahmed, A., Moore, T. M., Lewis, J., Butler, L., Benton, T. D., & Boyd, R. C. (2018). Psychometric properties of bully, fighting, and victimization scales among clinically referred youth. *Journal of Aggression, Maltreatment & Trauma*, Early View.

Alsaker, F. D. (2012). *Mutig gegen Mobbing in Kindergarten und Schule*. Bern: Huber Verlag.

Archer, J. (2004). Sex differences in aggression in real-world settings: A meta-analytic review. *Review of General Psychology*, *8*(4), 291.

Arseneault, L. (2018). Annual research review: the persistent and pervasive impact of being bullied in childhood and adolescence: implications for policy and practice. *Journal of Child Psychology and Psychiatry*, *59*(4), 405-421.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71(2*), 287-311.

Behling, O., & Law, K. (2000). Translating questionnaires and other research instruments: Problems and solutions. *Sage University Papers Series on Quantitative Applications in the Social Sciences.* Thousand Oaks, California: Sage.

Berry, J.W., Poortinga, Y.H., Breugelmans, S.M., Chasiotis, A. & Sam, D.L. (2002). *Cross-cultural psychology: research and applications.* Cambridge: Cambridge University Press.

Bosworth, K., Espelage, D. L., & Simon, T. R. (1999). Factors associated with bullying behavior in middle school students. *The Journal of Early Adolescence*, *19*, 341-362.

Cantone, E., Piras, A. P., Vellante, M., Preti, A., Daníelsdóttir, S., D'Aloja, E., ... & Bhugra, D. (2015). Interventions on bullying and cyberbullying in schools: A systematic review. *Clinical Practice and Epidemiology in Mental Health: CP & EMH*, *11*(Suppl 1 M4), 58.

Cassidy, W., Faucher, C., & Jackson, M. (2013). Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice. *School Psychology International*, *34*(6), 575-612.

Chan, A. Y., & Pan, Y. (2011). The use of cognitive interviewing to explore the effectiveness of advance supplemental materials among five language groups. F*ield Methods, 23(4),* 342-361.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95(5),* 1005-1018.

Eisenberg, M. E., Neumark-Sztainer, D., & Perry, C. L. (2003). Peer harassment, school connectedness, and academic achievement. *Journal of School Health*, *73*(8), 311-316.

Eisner, M.; Manzoni, P., & Ribeaud, D. (2000). *Gewalterfahrungen von Jugendlichen: Opfererfahrungen und selbst berichtete Gewalt bei Schülerinnen und Schülern im Kanton Zürich.* Aarau: Sauerländer Verlag.

Eisner, N., Murray, A.L., Eisner, M., & Ribeaud, D. (2018). An analysis of non-response and attrition in the Zurich Project on Social Development from Childhood to Adulthood (z-proso). *International Journal of Behavioral Development.* In press.

Epskamp, S., Cramer, A.O.J., Waldorp, L.J., Schmittmann, V.D. & Borsboom, D. (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software, 48*(4), 1-18.

Flaherty, J. A., Gaviria, F. M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J. A., & Birz, S. (1988). Developing instruments for cross-cultural psychiatric research. *The Journal of Nervous and Mental Disease, 176(5),* 260-263.

Furlong, M. J., Sharkey, J. D., Felix, E. D., Tanigawa, D., & Green, J. G. (2010). Bullying assessment: A call for increased precision of self-reporting procedures. In S. R. Jimerson, S. Swearer, & D. L. Espelage (Eds.) *Handbook of bullying in schools: An International Perspective* (pp. 329–345). New York: Routledge.

Gable, R. K., Ludlow, L. H., McCoach, D. B., & Kite, S. L. (2011). Development and validation of the survey of knowledge of internet risk and internet behavior. *Educational and Psychological Measurement*, *71*, 217-230.

Gini, G., Albiero, P., Benelli, B., & Altoè, G. (2007). Does empathy predict adolescents' bullying and defending behavior? *Aggressive Behavior*, *33*(5), 467-476.

Gruber, J. E., & Fineran, S. (2008). Comparing the impact of bullying and sexual harassment victimization on the mental and physical health of adolescents. *Sex Roles*, *59*(1-2), 1.

Haynie, D. L., Nansel, T., Eitel, P., Crump, A. D., Saylor, K., Yu, K., & Simons-Morton, B. (2001). Bullies, victims, and bully/victims: Distinct groups of at-risk youth. *The Journal of Early Adolescence*, *21*, 29-49.

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

Huang, F. L., & Cornell, D. G. (2015). The impact of definition and question order on the prevalence of bullying victimization using student self-reports. *Psychological Assessment*, *27*, 1484.

Kaiser, D., Eisner, M., Ribeaud, D., Trajtenberg, N., Murray, A.,L., (2018). Moral neutralisation and bullying perpetration in adolescence: an evaluation of cross-cultural differences between Zurich and Montevideo. Manuscript submitted for publication.

Kert, A. S., Codding, R. S., Tryon, G. S., & Shiyko, M. (2010). Impact of the word "bully" on the reported rate of bullying behavior. *Psychology in the Schools, 47*, 193-204.

McMaster, L. E., Connolly, J., Pepler, D., & Craig, W. M. (2002). Peer to peer sexual harassment in early adolescence: A developmental perspective. *Development and Psychopathology*, *14*, 91-105.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11), S69-S77.

Miller, S., Williams, J., Cutbush, S., Gibbs, D., Clinton-Sherrod, M., & Jones, S. (2013). Dating violence, bullying, and sexual harassment: Longitudinal profiles and transitions over time. *Journal of Youth and Adolescence*, *42*(4), 607-618.

Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. New York: Routledge.

Mitsopoulou, E., & Giovazolias, T. (2015). Personality traits, empathy and bullying behavior: A meta-analytic approach. *Aggression and Violent Behavior, 21*, 61-72.

Modecki, K. L., Minchin, J., Harbaugh, A. G., Guerra, N. G., & Runions, K. C. (2014). Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying. *Journal of Adolescent Health*, *55*, 602-611.

Murray, A. L., Eisner, M., Obsuth, I., & Ribeaud, D. (2017a). Situating violent ideations within the landscape of mental health: Associations between violent ideations and dimensions of mental health. *Psychiatry Research, 249,* 70-77.

Murray, A. L., Eisner, M., Obsuth, I., & Ribeaud, D. (2017b). No evidence that substance use causes ADHD symptoms in adolescence. *Journal of Drug Issues, 47*, 405-410.

Murray, A. L., Eisner, M., & Ribeaud, D. (2017). Can the Social Behavior Questionnaire help meet the need for dimensional, transdiagnostic measures of childhood and adolescent psychopathology? *European Journal of Psychological Assessment*. Early view.

Murray, A. L., Obsuth, I., Eisner, M., & Ribeaud, D. (2017). Evaluating longitudinal invariance in dimensions of mental health across adolescence: An analysis of the Social Behavior Questionnaire. *Assessment*, Early view.,

Muthén, L. K., & Muthén, B. O. (2010). *Mplus: Statistical analysis with latent variables: User's guide* (pp. 1998-2007). Los Angeles: Muthén & Muthén.

Olweus, D. (1993). *Bullying at school: What we know and what we can do*. Oxford: Blackwell.

Olweus, D. (1996). The revised Olweus bully/victim questionnaire. University of Bergen, Research Center for Health Promotion.

Orpinas, P., & Frankowski, R. (2001). The Aggression Scale: A self-report measure of aggressive behavior for young adolescents. *The Journal of Early Adolescence*, *21*, 50-67.

Poteat, V. P. & Espelage, D. L. (2005). Exploring the relation between bullying and homophobic verbal content: The Homophobic Content Agent Target (HCAT) Scale. *Violence and Victims*, *20*, 513.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354-373.

Rosen, L. H., Beron, K. J., & Underwood, M. K. (2013). Assessing peer victimization across adolescence: Measurement invariance and developmental change. *Psychological Assessment*, *25*, 1-11

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8,* 23-74.

Shaw, T., Dooley, J. J., Cross, D., Zubrick, S. R., & Waters, S. (2013). The Forms of Bullying Scale (FBS): Validity and reliability estimates for a measure of bullying victimization and perpetration in adolescence. *Psychological Assessment*, *25*, 1045-1057.

Swearer, S. M., Song, S. Y., Cary, P. T., Eagle, J. W., & Mickelson, W. T. (2001). Psychosocial correlates in bullying and victimization: The relationship between depression, anxiety, and bully/victim status. *Journal of Emotional Abuse*, *2*, 95-121.

Thomas, H. J., Connor, J. P., & Scott, J. G. (2015). Integrating traditional bullying and cyberbullying: challenges of definition and measurement in adolescents–a review. *Educational Psychology Review, 27*, 135-152.

Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivee, S., & LeBlanc, M. (1991). Disruptive boys with stable and unstable high fighting behavior patterns during junior elementary school. *Journal of Abnormal Child Psychology*, *19*, 285-300.

Ttofi, M. M., Farrington, D. P., & Lösel, F. (2012). School bullying as a predictor of violence later in life: A systematic review and meta-analysis of prospective longitudinal studies. *Aggression and Violent Behavior, 17*, 405-418.

Ttofi, M. M., Farrington, D. P., Lösel, F., & Loeber, R. (2011a). Do the victims of school bullies tend to become depressed later in life? A systematic review and meta-analysis of longitudinal studies. *Journal of Aggression, Conflict and Peace Research, 3*,3-73.

Ttofi, M. M., Farrington, D. P., Lösel, F., & Loeber, R. (2011b). The predictive efficiency of school bullying versus later offending: A systematic/meta-analytic review of longitudinal studies. *Criminal Behaviour and Mental Health*, *21*(2), 80-89.

Ttofi, M. M., Farrington, D. P., Lösel, F., Crago, R. V., & Theodorakis, N. (2016). School bullying and drug use later in life: A meta-analytic investigation. *School Psychology Quarterly, 31*, 8-27.

Ttofi, M. M., & Farrington, D. P. (2011). Effectiveness of school-based programs to reduce bullying: A systematic and meta-analytic review. *Journal of Experimental Criminology*, *7*, 27-56.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486-492.

Vivolo-Kantor, A. M., Martell, B. N., Holland, K. M., & Westby, R. (2014). A systematic review and content analysis of bullying and cyber-bullying measurement strategies. *Aggression and Violent Behavior*, *19*(4), 423-434.

Wetzels, P., Enzmann, D., Mecklenburg, E. & Pfeiffer, C. (2001). *Jugend und Gewalt: Eine repräsentative Dunkelfeldanalyse in München und acht anderen deutschen Städten*. Baden-Baden: Nomos.

Willis, G. B., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods, 23(4),* 331-341.

Winter, C.R. & McKenzie, K. (2017). Teachers' Perceptions of Female Student Aggression at an All-Girls School. *Journal of Adolescent Research*, *32* (4), 509-525.

Yang, A., & Salmivalli, C. (2013). Different forms of bullying and victimization: Bully-victims versus bullies and victims. *European Journal of Developmental Psychology*, *10*(6), 723-738.

Ybarra, M. L., Boyd, D., Korchmaros, J. D., & Oppenheim, J. K. (2012). Defining and measuring cyberbullying within the larger context of bullying victimization. *Journal of Adolescent Health*, *51*(1), 53-58.

Zych, I., Ortega-Ruiz, R., & Del Rey, R. (2015). Systematic review of theoretical studies on bullying and cyberbullying: Facts, knowledge, prevention, and intervention. *Aggression and Violent Behavior*, *23*, 1-21.

Table 1

Parameters for multi-group longitudinal models of victimisation and perpetration

Wave	Item	Victimisation Model Female	Male	Perpetration Model Female	Male
		Loadings			
Age 11	1	0.659	0.659	0.544	0.544
	2	0.931	0.931	0.868	0.868
	3	0.382	0.763	0.576	0.770
	4	0.578	0.578	0.557	0.428
Age 13	1	0.659	0.659	0.544	0.544
	2	0.931	0.931	0.868	0.868
	3	0.382	0.600	0.576	0.576
	4	0.578	0.578	0.557	0.557
	5	0.128	0.128	0.136	0.136
Age 15	1	0.659	0.659	0.544	0.544
	2	0.931	0.931	0.868	0.868
	3	0.382	0.382	0.576	0.576
	4	0.578	0.578	0.557	0.557
	5	0.205	0.128	0.243	0.136
Age 17	1	0.659	0.659	0.544	0.544
	2	0.931	0.931	1.508	0.868
	3	0.159	0.358	0.254	0.576
	4	0.578	0.578	0.557	0.557
	5	0.128	0.128	0.090	0.136
		Intercepts			
Age 11	1	1.778	1.778	1.654	1.654
	2	2.112	2.112	1.734	1.977
	3	1.584	1.750	1.499	1.709
	4	1.640	1.640	1.343	1.386
Age 13	1	1.778	1.778	1.654	1.654
	2	2.112	2.112	1.734	1.734
	3	1.291	1.620	1.115	1.271
	4	1.640	1.640	1.151	1.151
	5	1.548	1.192	1.053	1.053
Age 15	1	1.958	1.686	1.654	1.654
	2	2.112	2.112	1.734	1.734
	3	1.304	1.304	1.115	1.115
	4	1.640	1.640	1.151	1.151
	5	1.670	1.164	1.005	1.019
Age 17	1	2.005	1.742	1.654	1.654
	2	2.112	2.112	1.693	1.856
	3	1.14	1.304	1.091	1.139
	4	1.478	1.543	1.155	1.155
	5	1.668	1.135	1.028	1.055

Figure 1

Network graph of nomological net
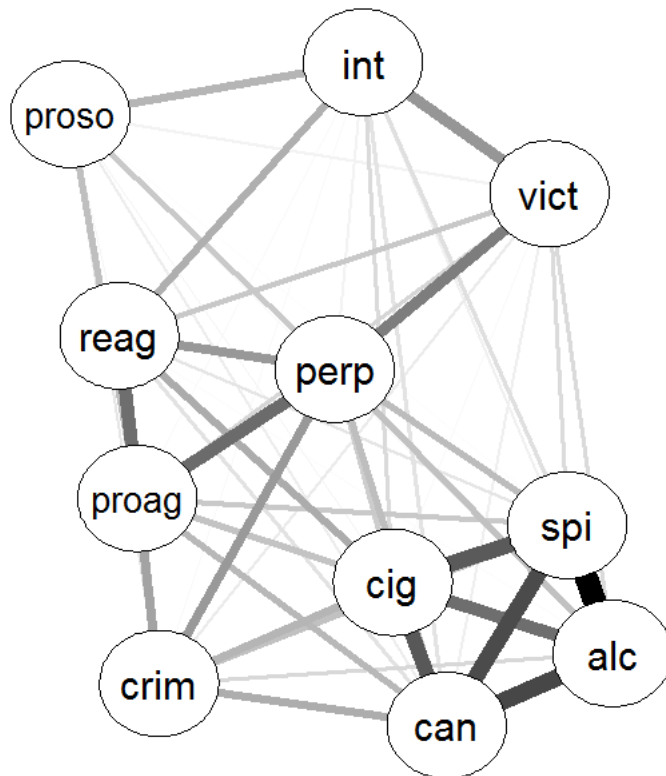
Note. proso= prosociality, int=internalising, vict= bullying victimisation, reag= reactive aggression, perp=bullying perpetration, proag= proactive aggression, cig=cigarette smoking, spi= alcohol (spirit-like), crim= criminal violence, can= cannabis, alc= alcohol (beer-like).

Supplementary Materials

Table S1

Descriptive statistics for bullying victimisation and perpetration

		Victimisation			Perpetration	
Item	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
			Age 11			
1	1145	1.82	1.11	1145	1.58	0.82
2	1145	2.06	1.23	1145	1.75	1.02
3	1145	1.66	1.02	1143	1.54	0.93
4	1144	1.58	0.88	1145	1.30	0.69
			Age 13			
1	1362	1.70	0.96	1364	1.87	1.01
2	1360	2.05	1.20	1360	2.11	1.20
3	1361	1.43	0.87	1365	1.48	0.90
4	1364	1.62	0.93	1364	1.43	0.79
5	1360	1.35	0.91	1364	1.15	0.64
			Age 15			
1	1444	1.74	1.01	1442	1.98	1.06
2	1438	1.98	1.14	1438	2.24	1.23
3	1441	1.29	0.70	1443	1.41	0.85
4	1444	1.56	0.88	1444	1.43	0.84
5	1445	1.38	0.90	1445	1.10	0.49
			Age 17			
1	1303	1.63	0.93	1304	1.71	0.90
2	1299	1.76	1.02	1302	1.82	1.08
3	1303	1.12	0.44	1302	1.18	0.53
4	1304	1.29	0.67	1304	1.21	0.55
5	1302	1.35	0.86	1304	1.06	0.38

Table S2

Model fits for the modified* single factor CFAs at each time point for males and females

	Females							Males						
	Model fit						Reliability	Model fit						Reliability
Wave	χ^2	p	CFI	TLI	RMSEA	SRMR	Omega	χ^2	p	CFI	TLI	RMSEA	SRMR	Omega
					Victimisation									
1	1.008	.315	.999	.995	.022	.008	.52	2.279	.131	.996	.973	.068	.011	.77
2	0.239	.993	1.000	1.013	<.001	0.002	.66	8.833	.065	.978	.946	.078	.024	.75
3	2.048	.727	1.000	1.005	<.001	.012	.55	4.604	.330	.990	.975	.045	.018	.64
4	3.684	.451	.999	.997	.014	.014	.55	.193	.526	1.000	.999	.009	.013	.65
					Perpetration									
1	2.091	.148	.994	.965	.059	.011	.57	1.492	.222	.996	.974	.068	.011	.76
2	5.249	.263	.997	.993	.026	.014	.61	9.144	.058	.989	.973	.058	.019	.73
3	4.564	.335	.992	.980	.047	.020	.61	1.024	.906	1.000	1.007	<.001	.009	.70
4	19.009	.001	.957	.892	.076	.036	.39	3.078	.545	1.000	1.000	.004	.011	.62

*Residual covariances between items 1 and 2 were added.

Table S3

Nomological net correlation matrix

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. **Bullying perpetration**	1	.40	.31	.45	.32	.04	-.17	.15	.19	.21	.20
2. **Bullying victimisation**	<.001	1	.06	.12	.17	.32	.04	.01	.11	.11	.05
3. **Criminal violence**	<.001	.03	1	.28	.27	-.02	-.08	.23	.12	.19	.25
4. **Proactive aggression**	<.001	<.001	<.001	1	.44	-.02	-.19	.18	.14	.17	.22
5. **Reactive aggression**	<.001	<.001	<.001	<.001	1	.25	-.03	.26	-.01	.08	.09
6. **Internalising**	.20	<.001	.55	.58	<.001	1	.23	.11	.05	.1	.09
7. **Prosociality**	<.001	.19	.01	<.001	.35	<.001	1	.03	-.03	.01	-.06
8. **Cigarette smoking**	<.001	.65	<.001	<.001	<.001	<.001	.30	1	.44	.51	.51
9. **Alcohol (beer-like)**	<.001	<.001	<.001	<.001	.80	.16	.43	<.001	1	.78	.56
10. **Alcohol (spirits)**	<.001	<.001	<.001	<.001	.01	<.001	.87	<.001	<.001	1	.55
11. **Cannabis use**	<.001	.10	<.001	<.001	<.001	<.001	.06	<.001	<.001	<.001	1

Note. Pearson correlations above the diagonal. P-values below the diagonal.

Appendix I: Bullying measure as administered to participants at age 15 (English translation)

This part is about **bullying**. Adolescents can be quite mean to each other sometimes. How about you?
In the last year, i.e., **since April 2012**, have you been bullied by other adolescents?
This could be, for example, at school, on the way to school, when being out in the evening, at home, or on the internet.

Since April 2012, how many times have other youths …	never	1 to 2-times	3 to 10-times	about once a month	about once a week	(almost) every day
… purposely ignored you or excluded you from something?	○	○	○	○	○	○
… laughed at you, mocked you, or insulted you?	○	○	○	○	○	○
… hit you, bitten you, kicked you, or pulled your hair?	○	○	○	○	○	○
… purposely stolen, broken, or hidden your things?	○	○	○	○	○	○
… sexually harassed you (e.g. hit on you, groped you).	○	○	○	○	○	○

And you? In the last year, i.e. **since April 2012**, have you bullied other adolescents?
This could be, for example, at school, on the way to school, when being out in the evening, at home, or on the internet.

Since April 2012, how many times have you…	never	1 to 2-times	3 to 10-times	about once a month	about once a week	(almost) every day
… purposely ignored or excluded another youth?	○	○	○	○	○	○
… laughed at, mocked, or insulted another youth?	○	○	○	○	○	○
… hit, bitten or kicked another youth, or pulled their hair.	○	○	○	○	○	○
… purposely stolen, broken or hidden another youth's things?	○	○	○	○	○	○
… sexually harassed (e.g. hit on, groped) another youth.	○	○	○	○	○	○

Appendix II: Bullying measure as administered to participants at age 15 (original German)

Jetzt geht es ums **Plagen**. Jugendliche können manchmal ziemlich gemein zueinander sein. Wie ist das bei dir?
Im letzten Jahr, das heisst **seit April 2012**, bist du da von anderen Jugendlichen geplagt worden?
Das kann zum Beispiel in der Schule, auf dem Schulweg, im Ausgang, zu Hause oder auch im Internet passiert sein.

Wievielmal haben andere Jugendliche seit April 2012 …	nie	1 bis 2-mal	3 bis 10-mal	etwa einmal pro Monat	etwa einmal pro Woche	(fast) jeden Tag
… dich absichtlich nicht beachtet oder ausgeschlossen?	❍	❍	❍	❍	❍	❍
… dich ausgelacht, beleidigt oder verspottet?	❍	❍	❍	❍	❍	❍
… dich geschlagen, gebissen, getreten oder an den Haaren gerissen?	❍	❍	❍	❍	❍	❍
… dir absichtlich Sachen weggenommen, kaputtgemacht oder versteckt?	❍	❍	❍	❍	❍	❍
… dich sexuell belästigt (z.B. angemacht, begrapscht).	❍	❍	❍	❍	❍	❍

Und du? Hast du im letzten Jahr, das heisst **seit April 2012,** andere Jugendliche geplagt?
Das kann zum Beispiel in der Schule, auf dem Schulweg, im Ausgang, zu Hause oder auch im Internet passiert sein.

Wievielmal hast du seit April 2012 einen/m anderen Jugendlichen …	nie	1 bis 2-mal	3 bis10-mal	etwa einmal pro Monat	etwa einmal pro Woche	(fast) jeden Tag
… absichtlich nicht beachtet oder ausgeschlossen?	❍	❍	❍	❍	❍	❍
… ausgelacht, beleidigt oder verspottet?	❍	❍	❍	❍	❍	❍
… geschlagen, gebissen, getreten oder an den Haaren gerissen?	❍	❍	❍	❍	❍	❍
… absichtlich Sachen weggenommen, kaputt-gemacht oder versteckt?	❍	❍	❍	❍	❍	❍
… sexuell belästigt (z.B. angemacht, begrapscht).	❍	❍	❍	❍	❍	❍