Gender differences in cross-informant discrepancies in aggressive and prosocial behaviour

A latent difference score analysis

Citation for published version:
Murray, AL, Nivette, A, Obsuth, I, Mirman, JH, Mirman, D, Ribeaud, D & Eisner, M 2022, 'Gender differences in cross-informant discrepancies in aggressive and prosocial behaviour: A latent difference score analysis', *Psychological assessment*. https://doi.org/10.1037/pas0001091

Digital Object Identifier (DOI):
10.1037/pas0001091

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Psychological assessment

Gender differences in cross-informant discrepancies in aggressive and prosocial behaviour: A latent difference score analysis

Aja Louise Murray[1*], Amy Nivette[2], Ingrid Obsuth[3], Jessica Hafetz Mirman[3], Daniel Mirman[1], Denis Ribeaud[4], Manuel Eisner[4,5]

[1]Department of Psychology, University of Edinburgh

[2] Department of Sociology, Utrecht University, The Netherlands

[3]School of Health in Social Science, University of Edinburgh

[4]Jacobs Center for Productive Youth Development, University of Zurich

[5]Violence Research Centre, University of Cambridge

*Corresponding author at: Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ; Email: aja.murray@ed.ac.uk

Funding: Funding for the z-proso study is gratefully acknowledged from the Jacobs Foundation and the Swiss National Science Foundation

Data and other study materials are available on reasonable request from the last author and scripts are available at: https://osf.io/3c2yq/.

Abstract

Cross-informant discrepancies (CIDs) in youth behaviour are common. Given that that these same behaviours often show or are perceived to show gender differences, it is important to understand how informant perceptions and their discrepancies are affected by gender. In n=1048 (51% male) grade 5 (age 11) Swiss youth, self- versus teacher- (*n*=261) CIDs were explored using latent difference score modelling. CIDs in prosociality ($\beta = -.15$) and aggression ($\beta=.14$) were predicted by child gender after adjusting for a range of covariates. Males reported more aggression than was attributed to them by teachers whereas females reported less aggression than was attributed to them. Both genders reported more prosociality than was attributed to them, with a larger discrepancy for males. Accounting for gender-related informant differences could help improve assessments used to ascertain whether clinically significant problems are present.

Keywords: cross-informant discrepancies; gender differences; aggression; prosociality

Public Significance: When assessing child social and behaviour problems, gathering information from multiple sources (e.g., both teachers and youth themselves) is recommended; however, these sources often disagree. We found that gender impacts the disagreements between youth and their teachers on youth aggressive and prosocial behaviour in a systematic way. This information helps improve the interpretation of scores from these two sources.

For the diagnosis of issues such as conduct disorder - characterised by features such as aggression, rule-breaking, and a lack of prosocial emotions (American Psychiatric Association, 2013) - it is considered critical to obtain a multi-informant perspective to ascertain the pervasiveness of problems and to take into account the context in which symptoms are occurring (Dirks et al., 2012). In research contexts, a multi-informant perspective is also considered the gold standard, where it helps to overcome potential errors and biases in the perspectives of individual informants, and to ensure that symptoms in different contexts and in interaction with different informants are captured (Achenbach, 2006; De Los Reyes, 2013).

The interpretation of multi-informant data is, however, made challenging by the fact that informants frequently disagree on whether and to what extent a young person is manifesting symptoms. Cross-informant discrepancies (CIDs) are common in studies of youth behaviour (Achenbach, 2006; De Los Reyes et al., 2015; Murray, Booth, et al., 2018). Identifying the features that influence CIDs is important for interpreting CIDs when they arise, providing insights into child-informant dynamics, and guiding strategies for dealing with them in research and clinical contexts (de Haan et al., 2018; Dirks et al., 2012). Previous research has found that the magnitude and direction of discrepancies relates to a range of features associated with the informant, the child, and their relationship (de Haan et al., 2018; Laird & LaFleur, 2016; Lohaus et al., 2020). However, the majority of previous studies have relied on suboptimal difference score or moderated multiple regression-based approaches (de Haan et al., 2018). Further, the impact of youth gender on CIDs has been under-studied given that gender is among the most important predictors of differences in dimensions of many social behaviours (e.g., aggression).

Gender differences in youth behaviour are an important object of study, evident in a broad range of normal and abnormal behaviour and potentially providing insights into etiological mechanisms (Rutter et al., 2003). Robust gender differences have been demonstrated in a range of traits with important long-term implications for the success and well-being (Booth & Murray, 2018); however, particularly large gender differences have been observed in conduct problems, with an estimated male:female gender ratio of 2.4-3.1:1 during middle childhood (Erskine et al., 2013). This gender difference is also reflected in population-level variation in constituent traits such as prosocial and aggressive behaviour (Archer, 2004; Card et al., 2008; Crick, 1996; Fabes & Eisenberg, 1998; Nivette et al., 2014). For example, based on a meta-analysis by Archer (2004), primary school-aged boys are consistently more aggressive than their female peers with a gender difference equivalent to an effect size of d=0.53. Meta-analytic investigations of childhood prosociality suggest a smaller but still substantive gender differences favouring girls in prosocial behaviours such as helping, being kind, and sharing (Fabes & Eisenberg, 1998).

The well-established nature of these gender differences may, however, make these behaviours vulnerable to gender-stereotyped ratings i.e., a greater tendency towards endorsement of male-typical behaviours in males and female-typical behaviours in females irrespective of child behaviour (Harvey et al., 2013; Pellegrini, 2011). It has been suggested, for example, that over-reporting of behaviour consistent with respective gender norms may lead some informants to over-report aggressive behaviour to a greater extent for males than for females, and to over-report prosocial behaviour for females to a greater extent than for males (Fabes & Eisenberg, 1998). This may include reports by youth themselves, as evidence suggests that gender stereotypical beliefs are detectable already by preschool (Baker et al., 2016).

Youth, teacher, parents, and untrained observer reports have been shown to be vulnerable to biases consistent with gender stereotyping (Pellegrini, 2011). It is not clear, however, whether the ratings from some informants tend to be affected by gender stereotypes more than others. Considering teachers and youth self-reports, for example, there are arguments to suggest that either informant may be more affected by gender stereotyping in reports. Teachers, for example, observe children only in certain (highly structured) and supervised contexts, they may not observe some behaviour which children themselves are aware of (Pellegrini, 2011). Teachers may, therefore, rely to a greater extent on implicit theories about which behaviours occur in girls versus boys to impute the missing information. This problem may become exacerbated when youth enter secondary school and spend more limited time with each teacher, especially for behaviours such as aggression which children may make efforts to conceal from adults (Pellegrini & Bartini, 2000). Youth however, may be biased in favour of reporting behaviours that are consistent with their gender identity.

To the extent that different informants (e.g., self- versus teacher-reports) are differentially affected by stereotyped rating, these biases may contribute to cross-informant discrepancies (Harvey et al., 2013). Such differences would also be relevant for assessments that seek to determine whether a young person shows clinically significant problems. It may be beneficial to make differential adjustments to self- versus teacher/parent reports of behavioural problems for males versus females in order to facilitate more accurate diagnosis if systematic ratings differences are observed.

While previous research has examined gender-stereotyping in perceptions of aggression and to a lesser extent, prosociality (Condry & Ross, 1985; Coyne et al., 2008; Pellegrini et al., 2011) there is very little research examining the potential role of gender in cross-informant discrepancies in these perceptions. In this study we, therefore, assessed the effects of youth gender on CIDs in both general aggressive and prosocial behaviour, adjusting for other potential sources of CID, namely, teacher connection with the youth, teacher gender, youth socially desirable responding and youth internalising problems.

Methods

Participants

Participants were drawn from the Zurich project on social development from childhood to adulthood (z-proso). Z-proso is a longitudinal cohort study mainly focused on the development of aggression, delinquency and other problem behaviour as well as of prosocial behaviour. The study began in 2004 when the participants were aged 7. Stratified random sampling at the school level, taking into account school size and location as stratification variables, was used to define a target sample of pupils entering 56 schools in Zurich, Switzerland. Location, specifically, school district, was used as a stratification variable to ensure the representation of pupils with varying socioeconomic background. Fourteen groups of schools were created defined by these stratification variables and 4 were randomly drawn from each. All pupils attending one of these selected schools were invited to participate. The target sample size was 1675 and 1572 participants provided some information for at least one of the measurement waves. Previous analyses suggested that those who participated at baseline were largely representative of those who declined to participate. The most important difference was that children whose parents did not speak German as a first language were under-represented in the study sample. Thus, the children at baseline could be considered broadly representative of the underlying population of same-aged children in (N. L. Eisner et al., 2018) and via the z-proso website: https://www.jacobscenter.uzh.ch/en/research/zproso.html and in the cohort profile (Ribeaud et al., 2021).

The analytic dataset for the current study was created by selecting all cases from the 4th measurement wave when the youth were of a median age of 11 years old. We focussed on this wave because it was the first point at which youth completed paper and pencil measures that were comparable to those administered to teachers (at previous waves, children completed computerised assessments). Children and their parents took part in interventions prior to the measurement wave utilised in the current study; however, because there was no evidence that the interventions had any short- or long-term effects we judged it justifiable to treat the data from this measurement wave as observational (Averdijk et al., 2016; Malti et al., 2011). At this stage, participation was based on informed consent from the youth's parents.

Thus, there were 1048 youth (51% male) who provided self-reported information about their prosociality and aggressive behaviours, and 261 teachers (35% male) who provided parallel ratings of these children's behaviours. At this stage the youth were in school grade 5 when most of the children were aged 11. The majority of participants (>80%) had been in the same classroom and with teacher who provided the ratings for just short of two years.

Measures

Prosociality and Aggression

Prosociality was measured using five items common to the child self-report and teacher versions of the Social Behavior Questionnaire (SBQ; Tremblay et al., 1991). These items referred to: volunteering to help (item 41), trying to stop disputes (item 42), trying to help someone who is hurt (item 43), comforting upset peer (item 46) and sharing things (item 49). Both versions were administered in German (the official language in the study location) in paper and pencil format and responses were given on a 5-point Likert scale from *never* to *very often.*

General aggression was measured using nine items common to the teacher and child versions of the SBQ measuring reactive aggression, specifically: aggressive if teased (item 53), aggressive if something is taken from them (item 54) and aggressive if contradicted (item 55); proactive aggression, specifically: making threats (item 37), trying to dominate (item 51) and scaring other children (item 52); and physical aggression, specifically: fighting (item 33), attacking (item 34) and kicking/biting/hitting (item 35). We did not include any items measuring forms of aggression such as relational or 'indirect' aggression that are not considered gender-stereotypic or if they are, are more likely to be (stereotypically if not empirically) associated with females (e.g., Pellegrini, 2011).

The translation process for the SBQ is described in previous publications (M. Eisner & Ribeaud, 2007). In brief, translation from English into German was based on an 'expert team approach' that involved formally qualified translators in an initial step. These translators received written guidance, including: general criteria for item translation, the need to document potentially ambiguous wordings, and brief descriptions of the purpose of each scale. An initial translation completed by these translators was then given to an independent translator who, where possible, had a background in social science and some experience of translation in the context of psychometric instruments. They were asked to review the initial translation and suggest any revisions where necessary. Finally, the initial translator, reviewer, and translation co-ordinator met to resolve any areas of disagreement/ambiguity that had arisen in the process. Study fieldworkers were also involved at this stage to provide an additional perspective to further ensure the clarity and conceptual equivalence of the translations.

The psychometric properties of the version of the SBQ used in the current sample have been investigated and supported in several previous studies in the current sample (Murray, Eisner, Obsuth, et al., 2017; Murray, Eisner, & Ribeaud, 2017; Murray, Obsuth, et al., 2017) as well as previous research in independent samples (Tremblay et al., 1991, 1992). For example, one study established that the SBQ as administered in the current sample has a wide ‘reliable range of measurement’ meaning that it yields reliable scores for a wide range of trait levels. (Murray, Eisner, & Ribeaud, 2017), while another suggested that it yielded measurement invariant scores up to the metric level longitudinally over adolescence (Murray, Obsuth, et al., 2017). The factorial validity of the SBQ has also been investigated and supported (Murray, Eisner, Obsuth, et al., 2017), and its criterion validity can be inferred from evidence across numerous studies that show that its aggression items correlate well with other relevant measures such as aggressive behaviour measured using an ecological momentary assessment study design, self-control, moral neutralisation, and violent ideations (Murray, Eisner, et al., 2018; Murray et al., 2019).

Covariates

Self-reported internalising problems were also included to adjust for negative response biases on the part of the youth. Internalising problems are more common in females than males and may be related to symptom over-reporting (Fergusson et al., 1993). Internalising problems were measured using 8 items from the self-report version of the SBQ tapping anxiety and depression. These items referred to crying a lot (item 1), being nervous/tense (item 2), being fearful/anxious (item 3), being worried (item 4), being unhappy (item 5), being miserable/distressed/unhappy (item 8), having trouble sleeping (item 62), feeling bored (item 63) and feeling alone (item 65).

The teacher-reported connection between the youth and their teacher was also adjusted for as past research has shown that CIDs are associated with the quality of relationship with informants (De Los Reyes & Kazdin, 2006). This was measured using a single item assessing whether the teacher felt they had a good connection with the young person. Responses were on a 5-point Likert scale from *strongly disagree* to *strongly agree*.

Teacher gender was adjusted for because some past research has suggested that male raters tend to attribute more aggression to boys relative to female raters (Pellegrini et al., 2011). In addition, female teachers may rate students generally more favourably than male teachers do, especially when rating boys (Quaglia et al., 2013).

Socially desirable responding by the youth was adjusted for because over- or under-reporting behaviours may affect gender differences when the desirability of a behaviour is gender-linked. Previous research has shown that adolescents' socially desirable responding scores are correlated with their aggression scores (Barry et al., 2017); however, the same study found no association between socially desirable responding and prosociality. Socially desirable responding was measured using three items referring to always being nice to other people, always telling the truth to everyone, and always sharing their favourite things with everyone. These behaviours were measured on a 4-point Likert scale from *not at all true* to *completely true* and coded such that *completely true* answers obtain a score of 1 and all others obtain a score of 0. These measures were developed on the same principles as other major social desirability scales such as the Marlow-Crowne Social Desirability Scale (Crowne & Marlowe, 1960). Their use is predicated on the probability of a child being able to genuinely answer *completely true* is so low that responding in this way can be assumed to indicate socially desirable responding rather than reflecting true behaviour. A composite was formed from these three items by summation of individual item scores.

Statistical Procedure

As a first step, measurement invariance across informants was assessed using a series of multi-group confirmatory factor analyses (de Haan et al., 2018). Residual covariances between the same item across informants were freely estimated across all models. First, in a configural model, the patterns of loadings were fixed equal across informants, the mean and variance of the teacher factor fixed to 0 and 1 respectively, and the loading and intercept of one item fixed equal across informants. Next, metric invariance constraints (equal loadings across informants) were added. Metric invariance was judged supported if there was <0.010 decrease in CFI, <0.015 increase in RMSEA and <0.030 increase in SRMR with the addition of metric invariance constraints to a configural model (Chen, 2007). Finally, provided that metric invariance held, scalar invariance constraints (equal intercepts across informants) were added. Scalar invariance was judged supported if there was <0.010 decrease in CFI, <0.015 increase in RMSEA and <0.010 increase in SRMR with the addition of scalar invariance constraints (Chen, 2007). For aggression, we used a general factor model using residual covariances to account for the expected excess covariance between items measuring reactive, proactive and physical aggression respectively.

We used the measurement invariance criteria proposed by (Chen, 2007) based on the goal of striking a balance between identifying non-invariance that could have a biasing effect but without rejecting models with only trivial (non-invariance) mis-specifications. As such, we did not use a chi-square difference testing approach as this is prone to reject models with only trivial mis-specifications in large sample sizes (Yuan & Chan, 2016). It is, however, important to acknowledge that other criteria have been suggested (both stricter and less strict) and the choice of which criteria to adopt is subjective and depends on the relative concerns about the consequences of over- versus under-identification of measurement non-invariance in a given context (see e.g., Svetina et al., 2020).

Latent difference score modelling

Using the model with the highest level of invariance attainable in the above-described cross-informant measurement invariance analyses, a latent difference score (LDS) model (de Haan et al., 2018) was fit. A simplified version of the LDS model is shown in Figure 1 to illustrate its key features, using the example of a three-item measure of prosociality. In brief, in the LDS, teacher-reported and self-reported prosociality/aggression can each be defined as separate latent factors, ideally with cross-informant scalar invariance. Then, using a specific set of parameter constraints in a regression of the self-reported factor on the teacher-reported factor a latent informant discrepancy factor can be defined that represents the directional discrepancy between the informants. Positive scores on this factor represent higher self-reported than teacher-reported ratings while negative scores represent higher teacher-reported than self-reported scores. The informant discrepancy factor can then serve as an outcome in a regression in order to illuminate the effect of factors such as gender.

The LDS model has several advantages over traditional methods of predicting cross-informant discrepancies, discussed in detail elsewhere (de Haan et al., 2018; Meisel et al., 2019). In brief, unlike using observed difference scores and regression residuals, LDS models provide more reliable measures of difference due to using a latent measurement model. Further, unlike observed difference scores, they do not place untested constraints on the relations between the difference score components (here teacher versus self-reports) nor confound the effects of the difference score components with the difference itself. Compared with polynomial regression (Laird & De Los Reyes, 2013), they leverage the numerous advantages of the structural equation model framework and are thus considerably more flexible.

In the current study, an LDS model was fit in order to allow us to explore whether gender is related to self-teacher discrepancies in prosociality and aggression. Separate models were fit to explore whether gender predicts discrepancies in prosociality and aggression and for each construct two models were fit. First, an unadjusted model estimated the raw relation between gender and informant discrepancy with no covariates. In these models the relevant informant discrepancy factor derived as described above was regressed on gender alone (coded female=0, male=1). Second, covariate-adjusted models were fit to explore the impact of additional potential explanatory factors: child internalising problems, teacher-child relationship, and socially desirable responding on the part of the child. Internalising was specified as single factor latent variable, scaled and identified by fixing its variance to 1 but all other constructs were treated as observed. These covariates were adjusted for by including them as addition predictors of the latent discrepancy factor. Analyses were conducted in *Mplus 8.4* (Muthén & Muthén, 2014).

To account for clustering within classrooms (and teachers) robust maximum likelihood estimation (MLR) was used. There are two broad methods that can be used to account for this kind of clustering: model-based techniques (multi-level modelling) and design-adjustments (e.g., Murray et al., 2021; Wu & Kwok, 2012). MLR falls into the latter category and uses a sandwich estimator to provide a correction to the standard errors of the parameter estimates and model fit statistics that counteracts their inflation in clustered data due to the violation of the assumption of independence (Stapleton, 2006; Wu & Kwok, 2012) . It alone does not make any adjustment to the parameter estimates themselves. Design-based adjustments such as MLR are useful when only the level-1 model is of interest as in such cases they are both easier to fit and provide a more straightforward interpretation than model-based techniques. In the case of the current data this method was also selected over the model-based option of multi-level structural equation modelling (ML-SEM) for the practical reason that previous research in the sample has suggested that there is very little between-classroom variance in aggression and prosociality self-reports (with most items showing ICCs <.01), making it difficult to estimate a multi-level model (which relies on sufficient variance at both the within- and between- level) (Murray, Obsuth, et al., 2019). *Mplus* scripts for these analyses are available at: https://osf.io/3c2yq/

Ethics

The z-proso study underwent ethical review and approval was received from the University of Zurich Faculty of Arts and Humanities ethics committee. The research was conducted in accordance with the ethical standards laid down by the 1964 Declaration of Helskini and its later amendments. Data and other study materials are available on reasonable request from the last author and scripts are available at: https://osf.io/3c2yq/

Results

Descriptive Statistics

Descriptive statistics are provided in Table 1. In general, responses to the aggression items tended to suggest low frequencies of aggression (especially physical aggression) and moderate frequencies of prosociality.

Measurement models

Prosociality

Neither metric nor scalar invariance across teacher and self-reports of prosociality were fully supported based on the model comparison criteria of Chen (2007). Using modification indices and parameter estimates from the scalar rand metric models to guide the release of constraints, we found a partial invariance (Pokropek et al., 2019) could be achieved by allowing the intercept and loading of the item measuring trying to comfort others when upset and the intercept of the item measuring trying to settle disputes to vary across informants. Indeed, recent simulation studies have suggested that unbiased inferences on structural parameters (latent means, variances) are possible even when up to 80% of the items are non-invariant (Pokropek et al., 2019). This model showed good fit by conventional standards (RMSEA=.06, CFI=.97, TLI=.95, SRMR=.04) and all loadings were salient (>|.3|). This model suggested that children reported more prosocial behaviour than their teachers attributed to them (M = 0.64, p<.001).

Aggression

Following a similar procedure as for the multi-informant prosociality model, we found a partially invariant model for aggression in which the loadings and intercepts for the item measuring threatening others to get something and the item measuring scaring others to force them to do something and the intercepts for the item measuring hitting/biting/kicking , the item measuring responding aggressively when something is taken from them and the item measuring responding aggressively when not getting their way were freely estimated across informants. This model fit well by conventional standards (RMSEA=.04, CFI=.96, TLI=.94, SRMR=.06) and suggested that youth on average reported no less aggression than teachers attributed to them (M = -.06, p=.406).

Internalising problems

We also found that our proposed unidimensional measurement model for child-reported child internalising problems fit well (RMSEA=.04, CFI=.98, TLI=.97, SRMR=.03) and that all items showed salient loadings and statistically significant loadings.

Latent difference score models

Prosociality

Raw gender and informant differences in latent prosociality based on the above-described multi-informant prosociality measurement model are shown in Figure 2. Teacher-reported prosociality for females was used as the reference, with all differences relative to that group. Using the same multi-informant measurement model in the context of a latent different score model, we found a significant effect of child gender on latent discrepancies ($\beta = -.20$, $p<.001$). The effect suggested that males tend to have more negative (teacher-self) discrepancies. Specifically, males and females both report more prosocial behaviour than their teachers attribute to them; however, the discrepancy is larger for males. When adjusting for teacher gender ($\beta= -.04$, $p=.11$), child internalising problems ($\beta=.04$, $p=.28$), social desirability ($\beta=.27$, $p<.001$) and teacher-child connection ($-.01$, $p=.76$), the effect of child gender was attenuated but still significant ($\beta = -.15$, $p<.001$).

Aggression

Raw gender and informant differences in latent aggression based the above-described multi-informant aggression measurement model are shown in Figure 2. Teacher-reported aggression for females was used as the reference, with all differences relative to that group. Using the same above-described multi-informant aggression measurement model in a latent difference score model we found that males had more positive discrepancies than females ($\beta=.19$, $p<.001$). Figure 2 shows that this is a result of the fact that male (self-teacher) discrepancies are positive, i.e., they report more aggression than their teachers attribute to them, while female (self-teacher) discrepancies are negative, i.e., they report less aggression than teachers attribute to them. When adjusting for teacher gender ($\beta=.00$, $p=.91$), child internalising problems ($\beta=-.13$, $p=.017$), social desirability ($\beta=-.03$, $p=.53$) and teacher-child connection ($\beta=-.29$, $p<.001$), the effect of child gender was attenuated but remained significant ($\beta=.14$, $p<.001$).

Discussion

In this study, we found evidence that the direction or magnitude of cross-informant discrepancies between self- and teacher-reports in adolescence are dependent on gender. Specifically, males reported more aggression than was attributed to them by their teachers but females reported less. Both males and females reported more prosociality than was attributed to them by their teachers but the discrepancy was greater for males. Both of these findings held after adjusting for other potential influences on inter-rater discrepancies, namely: teacher gender, child internalising problems, child socially desirable response tendencies, and teacher-child connection.

Our results are generally consistent with previous evidence which suggests that many symptoms and behaviours are reported by adolescents to a greater extent than they are attributed to them by informants and that this varies by gender. Penney & Skilling (2012), for example, found that females tended to report more somatic complaints compared to their caregivers, whereas males did not. Similarly, Sourander et al., (1999) found a greater tendency to report more internalising problems relative to caregiver reports in females compared to males. However, few previous studies have focused on aggression and prosociality and none have yet examined gender effects on their CIDs using a latent difference score model to overcome the limitations of traditional modelling approaches (de Haan et al., 2018). Our study thus arguably provides some of the most direct and compelling evidence to date of the impact of gender impacts on aggression and prosociality CIDs.

Prosociality and aggression reporting (except indirect aggression) have been previously proposed to be affected by gender because it is argued that males and females are socialised to express these respective behaviours (Card et al., 2008; Fabes & Eisenberg, 1998). However, the discrepancies in the present study suggest that to the extent that there is a gender stereotyping effect, it is not straightforward: males self-reported relatively more aggression than teachers attributed to them; however, they also reported relatively more prosociality than their teachers attributed to them. This could suggest that males are more liable to show larger informant discrepancies in general, or it could reflect the effect of other differences between the traits such as observability by informants.

It was particularly noteworthy that females reported less aggression than was attributed to them by their teachers despite males showing the opposite reporting tendency. Further illumination on the source of these discrepancies in future research would be valuable and have potentially important implications for the detection of conduct problems in females. In particular, any tendency for female early adolescents to under-report aggressive behaviour would contribute to challenges in identifying and characterising their difficulties. It is already known that females who experience gender atypical problems such as autism spectrum disorder or attention-deficit/hyperactivity disorder symptoms tend to have to display more severe problems or additional co-morbidities in order to receive a diagnosis, and they are more likely to experience delayed diagnosis (Russell et al., 2011; Rutherford et al., 2016; Williamson & Johnston, 2015). Explanations commonly discussed for this gender gap in diagnosis include potential referrers (e.g., frontline healthcare staff, teachers) being less attuned to gender atypical issues and young people camouflaging of behaviours that they perceive to be gender-inconsistent (Hull et al., 2020). However, our findings point to the need to further investigate the role of under-reporting of gender-atypical behaviours at the point of assessment in the under-identification of females with ‘male-typical’ problems.

It is also important to consider the potential contribution of teacher report biases to the CIDs observed in the current study. For example, while adolescents observe their behaviour across all settings, teachers observe only a limited classroom-based sample of adolescent behaviours, therefore, it is conceivable that aggressive and prosocial behaviours occur that are not noted by a teacher. Much aggression may, for example, occur covertly as adolescents seek to avoid sanctions (Furlong et al., 2010). This could potentially help explain the attenuated gender difference in aggression that teachers reported in the present study relative to self-reports. Indeed, our findings suggests that self-reports will tend to yield larger gender differences in aggression than teacher reports. Thus, the nature of the informant will be important to take into account when interpreting gender differences identified in research contexts.

Given our finding of systematic informant differences in adolescent behaviour, it may ultimately be beneficial for the direction and magnitude of gender-related CIDs to be taken into account when assessing adolescents for social and behavioural problems. A multi-informant perspective is considered beneficial (and in some cases mandatory) for the diagnosis of emotional, social, and behavioural problems in children and adolescents (Achenbach, 2006); however, when informants disagree on the severity of problems, it can be difficult to determine whether a young person is experiencing clinically significant issues. There remains much confusion and debate regarding how the perspectives of different informants should be combined when assessing young people as part of the diagnostic process (e.g., Yeguez & Sibley, 2016). However, knowledge of the systematic differences that are likely to exist between informants and how they are related to gender can help clinicians to appropriately weight or compare different informant perspectives. Our results suggest that for male self-reports of aggression and prosociality to be comparable to teacher reports the former would require adjustment downwards (or the latter upwards). For females, self-reports of aggression would require adjustment downwards and prosociality upwards (or teacher reports upwards and downwards respectively) to provide comparability across informants. This would be particularly relevant in cases where reports are available from only the youth or the teacher. For instance, this could affect youth who are home-schooled, have experienced a permanent school exclusion, or for whom for other reasons there are no teachers who have had extensive enough contact to provide a reliable rating of their behaviour. Similarly, in a research context there may be missing data from one informant, for example, when a study is unable to receive a response from teachers or where resources allow the collection of data from only a subset of teachers (e.g., Mostafa, 2013). Here, treating a youth self-report report as comparable to a teacher report could lead to an over- or under-estimation of problem severity for a youth with comparable severity but scores from the other or both informants. Our findings suggest that this would further differ by gender, potentially biasing or obscuring gender effects depending on patterns of missingness. Including gender as an auxiliary variable as part of the missingness treatment may, therefore, be important for mitigating bias when some participants are missing data on one informant. However, further research will be required to confirm the replicability of effect and establish more precise estimates of its magnitude to inform corrections at the level of the individual youth in clinical and educational contexts.

Finally, it is informative to interpret the magnitude of the effect of gender on CIDs ($\beta = -.15$ for prosociality and $\beta=.14$ for aggression) in the context of the effects of other previously established predictors of CIDs when mutually adjusting for all predictors. For example, internalising problems have previously been noted as an important influence on CIDs due to their association with a more pessimistic response style (Lohaus et al., 2020). For aggression, the effect of gender was comparable to that of youth internalising problems ($\beta=-.13$). For prosociality, however, the effects of internalising problems were minimal ($\beta=.04$) and gender had the larger standardised effect. In contrast, socially desirable responding had a larger standardised effect on prosociality CIDs ($\beta=.27$) than gender but only a minimal effect on aggression CIDs $\beta=-.03$. Finally, teacher-reported teacher-child connection appeared to be an additional important predictor of aggression CIDs ($\beta=-.29$), consistent with previous research that suggests that the quality/closeness of relationship between child and another informant is related to their level of agreement (Lohaus et al., 2020).

Limitations

The primary limitation of the current study concerns the reliance on only two informants. Peer, parent, and trained observed reports could help adjudicate between these informants, as well as provide a broader perspective on early adolescent prosociality and aggression (Clemans et al., 2014). A second limitation concerns the brevity of the available prosociality and aggression measures, which did not include sufficient items of prosociality and aggression subdimensions to allow us to reliably examine CIDs in specific prosocial (e.g., helping versus empathy) and aggressive (e.g., reactive versus proactive) subtypes. Finally, we had only limited data on the teacher informants, therefore, we were unable to adjust for teacher depression, burnout, attitudes, and other factors that may impact their ratings of child behaviour.

Conclusion

Early adolescent males report more aggression than is attributed to them by their teachers while females report less aggression than is attributed to them by their teachers. Both males and females report more prosociality than is attributed to them by their teachers but the discrepancy is greater for males. These findings highlight the importance of considering the gender of a young person when interpreting reports from different informants.

References

Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*(2), 94–98.

Archer, J. (2004). Sex differences in aggression in real-world settings: A meta-analytic review. *Review of General Psychology*, *8*(4), 291–322.

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Averdijk, M., Zirk-Sadowski, J., Ribeaud, D., & Eisner, M. (2016). Long-term effects of two childhood psychosocial interventions on adolescent delinquency, substance use, and antisocial behavior: A cluster randomized controlled trial. *Journal of Experimental Criminology*, *12*(1), 21–47.

Baker, E. R., Tisak, M. S., & Tisak, J. (2016). What can boys and girls do? Preschoolers' perspectives regarding gender roles across domains of behavior. *Social Psychology of Education*, *19*(1), 23–39.

Barry, C. T., Lui, J. H., & Anderson, A. C. (2017). Adolescent narcissism, aggression, and prosocial behavior: The relevance of socially desirable responding. *Journal of Personality Assessment*, *99*(1), 46–55.

Booth, T., & Murray, A. L. (2018). Sex Differences in Personality Traits. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1265-1

Card, N. A., Stucky, B. D., Sawalani, G. M., & Little, T. D. (2008). Direct and indirect aggression during childhood and adolescence: A meta-analytic review of gender differences, intercorrelations, and relations to maladjustment. *Child Development*, *79*(5), 1185–1229.

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

Clemans, K. H., Musci, R. J., Leoutsakos, J.-M. S., & Ialongo, N. S. (2014). Teacher, parent, and peer reports of early aggression as screening measures for long-term maladaptive outcomes: Who provides the most useful information? *Journal of Consulting and Clinical Psychology*, *82*(2), 236.

Condry, J. C., & Ross, D. F. (1985). Sex and aggression: The influence of gender label on the perception of aggression in children. *Child Development*, 225–233.

Coyne, S. M., Archer, J., Eslea, M., & Liechty, T. (2008). Adolescent perceptions of indirect forms of relational aggression: Sex of perpetrator effects. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *34*(6), 577–583.

Crick, N. R. (1996). The role of overt aggression, relational aggression, and prosocial behavior in the prediction of children's future social adjustment. *Child Development*, *67*(5), 2317–2327.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349.

de Haan, A., Prinzie, P., Sentse, M., & Jongerling, J. (2018). Latent difference score modeling: A flexible approach for studying informant discrepancies. *Psychological Assessment*, *30*(3), 358.

De Los Reyes, A. (2013). Strategic objectives for improving understanding of informant discrepancies in developmental psychopathology research. *Development and Psychopathology*, *25*(3), 669–682.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*(4), 858.

De Los Reyes, A., & Kazdin, A. E. (2006). Informant discrepancies in assessing child dysfunction relate to dysfunction within mother-child interactions. *Journal of Child and Family Studies*, *15*(5), 643–661.

Dirks, M. A., De Los Reyes, A., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Embracing not erasing contextual variability in children's behavior: Theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *53*(5), 558.

Eisner, M., & Ribeaud, D. (2007). Conducting a criminological survey in a culturally diverse context: Lessons from the Zurich Project on the Social Development of Children. *European Journal of Criminology*, *4*(3), 271–298.

Eisner, N. L., Murray, A. L., Eisner, M., & Ribeaud, D. (2018). A practical guide to the analysis of non-response and attrition in longitudinal research using a real data example. *International Journal of Behavioral Development*, 0165025418797004.

Erskine, H. E., Ferrari, A. J., Nelson, P., Polanczyk, G. V., Flaxman, A. D., Vos, T., Whiteford, H. A., & Scott, J. G. (2013). Research Review: Epidemiological modelling of attention-deficit/hyperactivity disorder and conduct disorder for the Global Burden of Disease Study 2010. *Journal of Child Psychology and Psychiatry*, *54*(12), 1263–1274.

Fabes, R. A., & Eisenberg, N. (1998). Meta-analyses of age and sex differences in children's and adolescents' prosocial behavior. *Handbook of Child Psychology*, *3*, 1–29.

Fergusson, D. M., Lynskey, M. T., & Horwood, L. J. (1993). The effect of maternal depression on maternal ratings of child behavior. *Journal of Abnormal Child Psychology*, *21*(3), 245–269.

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.*

Harvey, E. A., Fischer, C., Weieneth, J. L., Hurwitz, S. D., & Sayer, A. G. (2013). Predictors of discrepancies between informants' ratings of preschool-aged children's behavior: An examination of ethnicity, child characteristics, and family functioning. *Early Childhood Research Quarterly*, *28*(4), 668–682.

Hull, L., Lai, M.-C., Baron-Cohen, S., Allison, C., Smith, P., Petrides, K. V., & Mandy, W. (2020). Gender differences in self-reported camouflaging in autistic and non-autistic adults. *Autism*, *24*(2), 352–363.

Laird, R. D., & De Los Reyes, A. (2013). Testing informant discrepancies as predictors of early adolescent psychopathology: Why difference scores cannot tell you what you want to know and how polynomial regression may. *Journal of Abnormal Child Psychology*, *41*(1), 1–14.

Laird, R. D., & LaFleur, L. K. (2016). Disclosure and monitoring as predictors of mother–adolescent agreement in reports of early adolescent rule-breaking behavior. *Journal of Clinical Child & Adolescent Psychology*, *45*(2), 188–200.

Lohaus, A., Rueth, J.-E., & Vierhaus, M. (2020). Cross-informant discrepancies and their association with maternal depression, maternal parenting stress, and mother-child relationship. *Journal of Child and Family Studies*, *29*(3), 867–879.

Malti, T., Ribeaud, D., & Eisner, M. P. (2011). The effectiveness of two universal preventive interventions in reducing children's externalizing behavior: A cluster randomized controlled trial. *Journal of Clinical Child & Adolescent Psychology*, *40*(5), 677–692.

Meisel, S. N., Fosco, W. D., Hawk, L. W., & Colder, C. R. (2019). Mind the gap: A review and recommendations for statistically evaluating Dual Systems models of adolescent risk behavior. *Developmental Cognitive Neuroscience*, 100681.

Mostafa, T. (2013). *Technical Report on Response in the Teacher Survey in MCS 4 (age 7)*.

Murray, A. L., Booth, T., Ribeaud, D., & Eisner, M. (2018). Disagreeing about development: An analysis of parent-teacher agreement in ADHD symptom trajectories across the elementary school years. *International Journal of Methods in Psychiatric Research*, e1723.

Murray, A. L., Eisner, M., Obsuth, I., & Ribeaud, D. (2017). 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., & 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*.

Murray, A. L., Eisner, M., & Ribeaud, D. (2018). Development and validation of a brief measure of violent thoughts: The Violent Ideations Scale (VIS). *Assessment*, *25*(7), 942–955.

Murray, A. L., Eisner, M., Ribeaud, D., & Booth, T. (2019). *Validation of a brief measure of aggression for experience sampling research: The Aggression-ES-A*.

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*, 1073191117721741.

Murray, A. L., Obsuth, I., Eisner, M., & Ribeaud, D. (2019). Disaggregating between-and within-classroom variation in student behavior: A multilevel factor analysis of teacher ratings of student prosociality and aggression. *The Journal of Early Adolescence*, *39*(7), 993–1019.

Murray, A. L., Ushakova, A., Wright, H., Booth, T., & Lynn, P. (2021). Structural equation modelling with complex sampling designs and non-random attrition: A tutorial using lavaan and Mplus. *PsyArXiv. February*, *16*.

Nivette, A. E., Eisner, M., Malti, T., & Ribeaud, D. (2014). Sex differences in aggression among children of low and high gender inequality backgrounds: A comparison of gender role and sexual selection theories. *Aggressive Behavior*, *40*(5), 451–464.

Pellegrini, A. D. (2011). “In the Eye of the Beholder” Sex Bias in Observations and Ratings of Children’s Aggression. *Educational Researcher*, *40*(6), 281–286.

Pellegrini, A. D., Bohn-Gettler, C. M., Dupuis, D., Hickey, M., Roseth, C., & Solberg, D. (2011). An empirical examination of sex differences in scoring preschool children's aggression. *Journal of Experimental Child Psychology*, *109*(2), 232–238.

Pellegrini, A. u, & Bartini, M. (2000). An empirical comparison of methods of sampling aggression and victimization in school settings. *Journal of Educational Psychology*, *92*(2), 360.

Penney, S. R., & Skilling, T. A. (2012). Moderators of informant agreement in the assessment of adolescent psychopathology: Extension to a forensic sample. *Psychological Assessment*, *24*(2), 386.

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A monte carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 724–744.

Quaglia, R., Gastaldi, F. G., Prino, L. E., Pasta, T., & Longobardi, C. (2013). *The pupil-teacher relationship and gender differences in primary school*.

Ribeaud, D., Murray, A. L., Shanahan, L., Shanahan, M., & Eisner, M. (2021). COHORT PROFILE: THE ZURICH PROJECT ON THE SOCIAL DEVELOPMENT FROM CHILDHOOD TO ADULTHOOD (Z-PROSO). *PsyArXiv. August*, *17*.

Russell, G., Steer, C., & Golding, J. (2011). Social and demographic factors that influence the diagnosis of autistic spectrum disorders. *Social Psychiatry and Psychiatric Epidemiology*, *46*(12), 1283–1293.

Rutherford, M., McKenzie, K., Johnson, T., Catchpole, C., O'Hare, A., McClure, I., Forsyth, K., McCartney, D., & Murray, A. (2016). Gender ratio in a clinical population sample, age of diagnosis and duration of assessment in children and adults with autism spectrum disorder. *Autism*, *20*(5), 628–634.

Rutter, M., Caspi, A., & Moffitt, T. E. (2003). Using sex differences in psychopathology to study causal mechanisms: Unifying issues and research strategies. *Journal of Child Psychology and Psychiatry*, *44*(8), 1092–1115.

Sourander, A., Helstelä, L., & Helenius, H. (1999). Parent-adolescent agreement on emotional and behavioral problems. *Social Psychiatry and Psychiatric Epidemiology*, *34*(12), 657–663.

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling*, *13*(1), 28–58.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using M plus and the lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111–130.

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*(3), 285–300.

Tremblay, R. E., Vitaro, F., Gagnon, C., Piché, C., & Royer, N. (1992). A prosocial scale for the Preschool Behaviour Questionnaire: Concurrent and predictive correlates. *International Journal of Behavioral Development*, *15*(2), 227–245.

Williamson, D., & Johnston, C. (2015). Gender differences in adults with attention-deficit/hyperactivity disorder: A narrative review. *Clinical Psychology Review*, *40*, 15–27.

Wu, J.-Y., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 16–35.

Yeguez, C. E., & Sibley, M. H. (2016). Predictors of informant discrepancies between mother and middle school teacher ADHD ratings. *School Mental Health*, *8*(4), 452–460.

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, *21*(3), 405.

Table 1: Descriptive Statistics

Variable	N	Mean	SD	Min	Max
Child-reported prosociality - helps clear up	1010	3.308	1.096	1	5
Child-reported prosociality - tries to settle a dispute	1003	3.473	1.168	1	5
Child-reported prosociality - tries to help when someone hurt	1008	4.130	0.953	1	5
Child-reported prosociality - tries to comfort upset peer	1009	3.979	1.010	1	5
Child-reported prosociality - shares readily with others	1009	3.892	0.969	1	5
Teacher-reported prosociality - helps clear up	1001	3.021	1.166	1	5
Teacher-reported prosociality - tries to settle a dispute	1004	3.088	1.092	1	5
Teacher-reported prosociality - tries to help when someone hurt	993	3.481	1.052	1	5
Teacher-reported prosociality - tries to comfort upset peer	1000	3.383	0.996	1	5
Teacher-reported prosociality - shares readily with others	1006	3.433	0.896	1	5
Child-reported aggression - brawl	1022	1.364	0.738	1	5
Child-reported aggression - violent attack	1023	1.333	0.702	1	5
Child-reported aggression - hit/bite/kick	1022	1.249	0.622	1	5
Child-reported aggression - threaten others to get something	1021	1.255	0.629	1	5
Child-reported aggression - boss others around	1021	1.537	0.874	1	5
Child-reported aggression - scare others to force them	1021	1.303	0.710	1	5
Child- reported aggression - aggressive when teased	1023	1.873	1.060	1	5
Child-reported aggression - aggressive when something taken	1019	1.851	1.009	1	5
Child-reported aggression - aggressive when not getting way	1022	1.631	0.899	1	5
Teacher-reported aggression - brawl	1018	1.417	0.785	1	5
Teacher-reported aggression - violent attack	1013	1.427	0.758	1	5

Teacher-reported aggression - hit/bite/kick	1019	1.525	0.799	1	5
Teacher-reported aggression - threaten others to get something	1014	1.123	0.451	1	5
Teacher-reported aggression - boss others around	1016	1.357	0.686	1	5
Teacher-reported aggression - scare others to force them	1018	1.203	0.563	1	5
Teacher- reported aggression - aggressive when teased	1014	2.728	1.085	1	5
Teacher-reported aggression - aggressive when something taken	1011	1.378	0.675	1	5
Teacher-reported aggression - aggressive when not getting way	1020	1.724	0.842	1	5
Internalising problems- crying	1011	2.089	0.955	1	5
Internalsing problems - fear	1013	1.616	0.915	1	5
Internalising problems - worry	1015	2.091	1.104	1	5
Internalising problems - depressed	1013	1.717	1.073	1	5
Internalising problems - miserable	1014	2.080	1.068	1	5
Internalising problems - sleep difficulty	1012	2.431	1.295	1	5
Internalising problems - boredom	1014	2.588	0.950	1	5
Internalising problems - loneliness	1014	1.735	0.999	1	5
Teacher-child bond	1024	4.034	0.876	1	5
Social desirability score	1024	2.920	0.575	1	4

Figures

Figure 1: A basic latent difference score model to model the discrepancy between teacher and self-reported prosociality

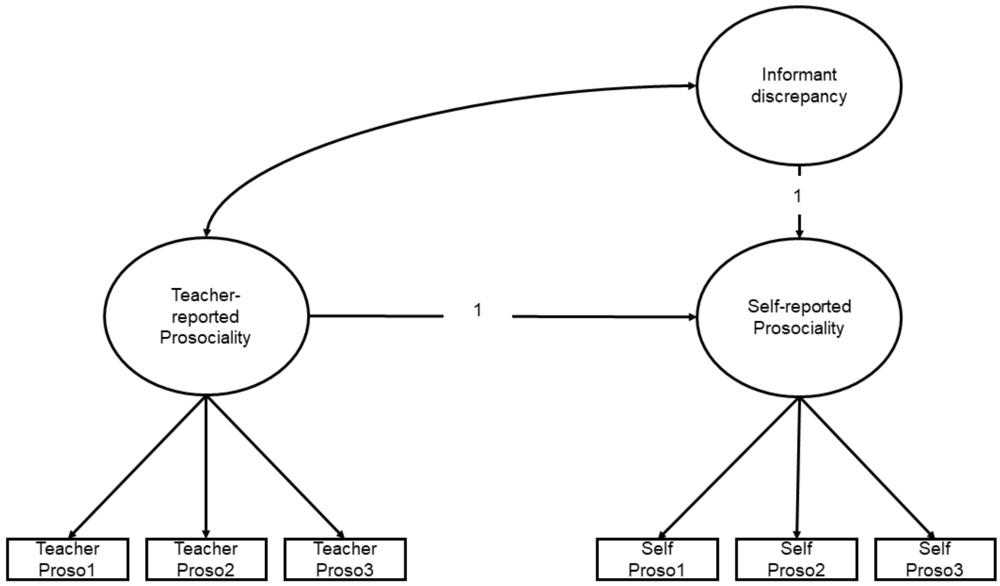
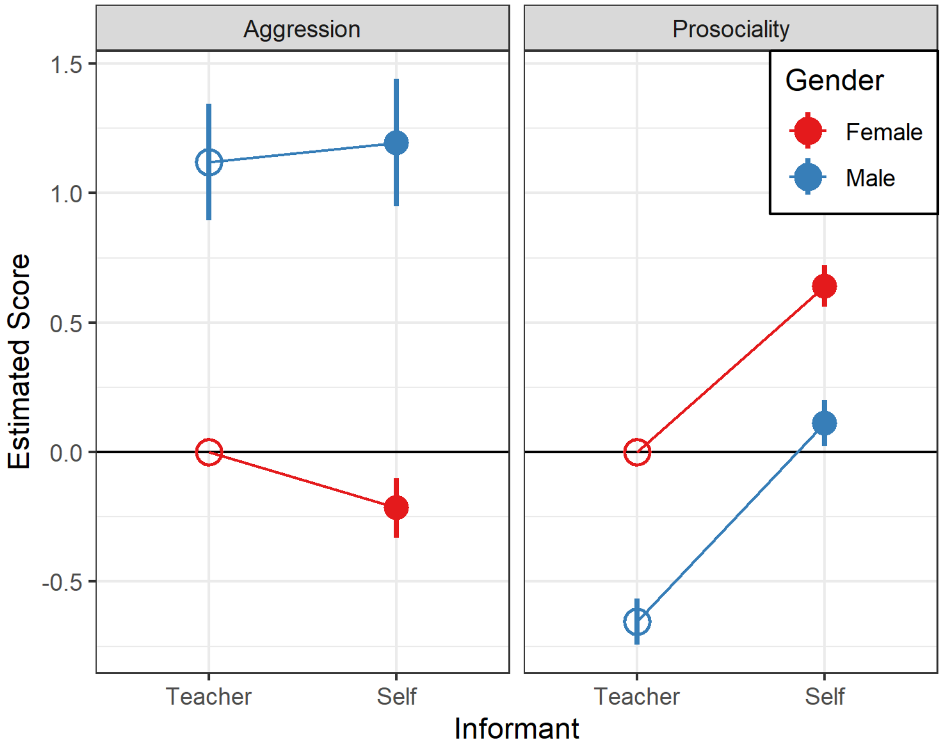


Figure 2: Latent means for teacher- and self-reported prosociality and aggression by gender.



Note. Teacher-reported scores for females was the reference level, therefore, it has an estimated latent mean of 0.