Do we understand our daughters' depression and anxiety? A bifactor modeling approach

Abstract

Background: Depression and anxiety are common in adolescence, but adolescents' reports of their symptoms are often at odds with their parents' report. The extent to which reporting difference is a function of gender and context (i.e., clinical or community) has yet to be established. Since discrepancies have been found to be predictive of poor long-term health outcomes, it is critical to determine the extent of the problem in a community sample, and that disagreement itself be measured with the upmost precision. **Methods**: Lifetime symptoms of depression, generalized anxiety and social anxiety were modeled with a bifactor structure, and various quality assessments were performed. Adolescent and caregiver reports of the adolescent's symptoms came from the GOASSESS on the Philadelphia Neurodevelopmental Cohort (4,812 adolescent-caregiver pairs; adolescents aged 11-17 years; 52.2% female; 57.1% White, 31.7% Black). **Results**: Controlling for lifetime internalizing severity, females still endorse crying more frequently than males, which indicates item bias (χ^2=134.13, p_{Bon}<.05). The caregivers' reports for the adolescents also evidence this pattern, but to a lesser extent (χ^2=29.49, p_{Bon}<.05). Caregivers tended to underestimate the severity of adolescents' internalizing symptoms, and when they disagreed with adolescents on a specific symptom, it was often twice as likely that the adolescent endorsed the symptom and the caregiver denied it than the reverse. This effect was markedly more pronounced for female than male adolescents. **Conclusions**: Researchers should build on this work by exploring potential sources of the gender differences in disagreement, and by using measures that assess recent symptomatology using Likert scales.

Keywords: informant, bifactor model, depression, anxiety, adolescence

Introduction

The three most common classes of mental disorders among adolescents are mood, anxiety, and behavioral disorders (Merikangas et al., 2010). In the United States in 2015, 11% of adolescents between the ages of 13 and 18 had major depressive disorder in their lifetime (MDD), and 7.5% had MDD in the past year (Avenevoli, Swendsen, He, Burstein, & Merikangas, 2015). By young adulthood, about 1 in 5 people have had an anxiety disorder, while annual prevalence estimates for adolescents range from 5.7% to 18.6% (Copeland, Angold, Shanahan, & Costello, 2014; Essau & Conradt, 2000; World Health Organization, 2004). Youth with a depressive disorder often also have an anxiety disorder, with estimates around 30%, while around 20% of youths with anxiety disorders have comorbid depressive disorders (Angold, Costello, & Erkanli, 1999; Axelson & Birmaher, 2001; Essau & Conradt, 2000). Usually, anxiety precedes depression during development, with generalized and social anxiety disorders being two common antecedents (Avenevoli, Stolar, Li, Dierker, & Merikangas, 2001; Fichter, Quadflieg, Fischer, & Kohlboeck, 2010).

Males and females are differentially affected by anxiety and depression. Differences in rates of depression across the sexes are evident in early adolescence, during which two to three times as many females suffer from depression as males (Hankin, 2009; Nolen-Hoeksema & Jackson, 2001). Similar but attenuated patterns have been observed with anxiety (Aune & Stiles, 2009; Axelson & Birmaher, 2001; Essau & Conradt, 2000; Leikanger, Ingul, & Larsson, 2012).

Depression and anxiety symptoms are particularly common among adolescents who perceive having little social support available to them from their parents (Rueger, Malecki, Pyun, Aycock, & Coyle, 2016). They may think that this support is not available because they do not

disclose their emotions, which could lead parents to underestimate the presence of their child's symptoms and subsequently fail to express willingness to provide social support. Alternatively, adolescents may not disclose their symptoms because they anticipate that it would not result in social support from their parents. Therefore, discrepancies between parent and adolescent reports of the adolescent's symptoms could reflect multiple types of inadequate family functioning (De Los Reyes & Kazdin, 2006).

Not surprisingly, these discrepancies have been shown to be predictive of poor outcomes. For instance, Ferdinand, Van Der Ende, & Verhulst (2004) found that when adolescents reported having worse anxiety and depression than their parents did for them, they were more likely to report having behavioral or emotional problems four years later. The authors proposed that this result may reflect lack of parental interest in or inability to recognize issues with their children. Disturbingly, informant discrepancies are the rule, not the exception. A meta-analysis by De Los Reyes et al. (2015) found that the correlation between parent and child reports of the child's internalizing symptoms is .29.

De Los Reyes & Kazdin (2005) proposed that the marked discrepancy may reflect contextual biases on the part of adolescents and their parents. Studies of informant discrepancies have largely been conducted in treatment-seeking families. In the case of adolescents, the parent usually makes the decision about whether the adolescent should see a mental health professional (Kazdin, 1989). If the adolescent agrees, both parties should be motivated to report symptoms such that the clinician will decide that they warrant treatment. If the adolescent disagrees, they may intentionally under-report symptoms to avoid treatment. Therefore, studies of treatment-seeking families may be particularly prone to finding low informant correspondence. To determine the magnitude of informant discrepancies, it is necessary to perform analyses on a

community sample with no pretext of treatment. Any remaining discrepancies will better represent the extent of the problem in the general population, and are more likely to reflect true differences in view between parents and their children.

In addition to contextual factors, the previously observed low correlation between adolescent and caregiver assessments may be due to inaccurate measurement of psychopathology severity using Classical Test Theory (CTT). CTT assumes exchangeability of item responses given the underlying trait level (Bechger, Maris, Verstralen, & Beguin, 2003). Assuming exchangeability of item responses means treating every item like it reflects the same level of a construct. Consider a depression measure that asks the following questions: "Has there ever been a time when you cried a lot, or felt like crying?" and "Have you ever thought about killing yourself?" Two people complete this questionnaire, responding "yes" to all the same items, except person #1 says yes to only the first item above, and person #2 says yes to only the second. Since suicidality is associated with more serious depression than having cried a lot, one would want to assign person #2 a greater depression severity score than person #1 (Hetrick, Parker, Robinson, Hall, & Vance 2012). Under CTT, though, they would get the exact same score.

Exchangeability is rarely, if ever, true in psychopathology measurement, and assuming exchangeability when it does not exist inevitably leads to worse estimates of psychopathology trait scores. Item response theory (IRT) presents a solution to the problem, because IRT-based approaches for estimating trait levels do not make the exchangeability assumption; instead, each item has its own set of parameter estimates, ensuring optimal weighting in score calculation. The exact parameters to be estimated depends on the IRT model, but in a unidimensional two-parameter logistic model, parameters reflecting the severity of the item and the extent to which the item is related to the trait being measured are estimated (S. P. Reise, 2014).

Several studies have used IRT to develop measures of adolescents' internalizing symptoms (Bevans, Diamond, & Levy, 2012; Irwin et al., 2010; Simms, Grös, Watson, & O'Hara, 2008), but few have done so with a caregiver's assessment of the subject's symptoms (Ebesutani et al., 2012). This is problematic because items might function differently in caregiver reports. For instance, female adolescents, especially when they are close with their family, are more likely to disclose their emotions to relatives than are male adolescents (Papini, Farmer, Clark, Micka, & Barnett, 1990). Therefore, caregivers of male adolescents may be prone to under-reporting symptoms that are hard to detect relative to caregivers of female adolescents, which would result in differential item functioning (i.e., item bias). This means that, *controlling for internalizing severity*, caregivers of male adolescents may be less likely to endorse an item that is less observable than caregivers of female adolescents. While many internalizing symptoms are unobservable, symptoms that may be less likely to change behaviors, such as worrying about world events far removed from day-to-day life, may be particularly hard for caregivers to detect. The fact that adolescents do not always disclose their symptoms could also lead their caregivers to guess about their symptoms, increasing error variance and decreasing the magnitude of any observed item bias effects.

Importantly, to meaningfully compare caregivers' assessments of the adolescents' symptoms, it is important to establish which, if any, of the items in a measure exhibit bias as a function of sex. It is also essential to establish if any pattern of item bias observed in the caregivers' reports are also apparent in the adolescents'. If this is the case, item bias in the caregivers' reports may be due simply to sex differences in the true emotional experience of the adolescents, as opposed to a lack of expressed emotion. For instance, one common item bias finding is that, after controlling for internalizing severity, females are more likely to endorse

questions that ask about crying than males (Gelin & Zumbo, 2003; Van Beek, Hessen, Hutteman, Verhulp, & Van Leuven, 2012). This implies that females may experience crying-related problems more readily than males at lower levels of internalizing severity. If the same pattern is found in caregivers' reports, this effect is probably due to the actual emotional experiences of the adolescents, opposed to observability issues or increased error variance. The issue of the source of the item bias can be probed further by explicitly testing if, given that an adolescent-caregiver pair disagrees on the presence of a symptom, caregivers are less likely to endorse said symptom. This would help establish if a discrepancy in the magnitude of an item bias effect is a function of caregivers being unaware of the symptom, which may be due to a lack of expressed emotion on the part of the adolescent.

In the present study, we hypothesize that, 1) there is one primary dimension that explains a large portion of the variance in depressive and anxious symptoms in both the caregivers' and the adolescents' reports of the adolescents' symptoms, but there are also noticeable group factors in a bifactor model; 2), *controlling for symptom severity dimensions*, female adolescents and their caregivers endorse crying more frequently than male adolescents and their caregivers; 3) item bias effects are larger in adolescents' reports than caregivers'; 4) caregivers' assessments of adolescents' symptoms is moderately correlated with adolescents' assessments of their own symptoms, with caregivers tending to underestimate severity; and 5) given that caregivers and adolescents are discrepant on a given symptom, it is more likely that the caregiver does not report its presence than the adolescent. As exploratory analyses, we will test whether the effects hypothesized in points #4 and #5 differ by gender.

Methods

Participants: Details of the recruitment protocol have been reported elsewhere (Calkins et al., 2015). Briefly, prospective participants (N=9,498) in the Philadelphia Neurodevelopmental Cohort (PNC) were recruited through the Children's Hospital of Philadelphia (CHOP) pediatric (non-psychiatric) health care network. Potential participants from this pool were excluded if they were not proficient in English, had significant developmental delays or other conditions that would interfere with their ability to complete study procedures, or could not be contacted. Subjects were further excluded for this study if they did not report on their own symptoms, they did not have a caregiver report on their symptoms, if skip logic was violated in the utilized questions from the depression, generalized anxiety or social anxiety sections of their assessments, or if participants did not complete the full version of the GOASSESS. The total sample for the current analyses included youths aged 11-17 and their caregivers (N=4,812 pairs of participants; Adolescents: mean age=14.52 (SD=1.97), 52.2% female, 57.1% White, 31.7% Black; Caregiver Relations: 4181 mothers, 490 fathers, 50 maternal grandmothers, 31 legal guardians, 19 paternal grandmothers, 8 maternal aunts, 3 maternal grandfathers, 2 paternal aunts, 16 other not biologically related, 9 other biologically related, and 3 unknown) assessed between March 2010 and August 2013. The University of Pennsylvania and CHOP Institutional Review Boards approved all procedures.

Measures: Adolescents (age 11-17) and caregivers were independently administered a computerized structured interview (GOASSESS). Psychopathology screen was conducted through an abbreviated computerized version of the NIMH Genetic Epidemiology Research Branch Kiddie-Schedule for Affective Disorders and Schizophrenia (K-SADS) (Kaufman et al., 1997), that was modified to collect information on symptoms, duration, distress and impairment

for lifetime mood, anxiety, behavioral, psychosis spectrum and eating disorders, suicidal thinking and behavior, as well as treatment history.

Scale Development: The DEP, GAD and SOC sections of the GOASSESS were analyzed together in a bifactor model in R as implemented by 'mirt' (Chalmers, 2012; R Development Core Team, 2014; S. P. Reise, Moore, & Haviland, 2010). A confirmatory factor structure was specified such that all items were allowed to load on the general factor, there were group factors for each specific domain of psychopathology, and all dimensions were orthogonal to one another. The model was fit using an EM algorithm (Gibbons & Hedeker, 1992).

Prior to any other analyses, monotonicity and dimensionality were evaluated. Using two-parameter logistic models estimated using the 'psych' package in R (Revelle, 2017), plots of the proportion of participants with a given trait estimate who responded "yes" to each item (empirical response functions) were examined visually. Dimensionality was assessed by creating a scree plot. The extent to which scores reflect a single variable (i.e., lifetime internalizing severity) was assessed by comparing the loadings on the factors of the confirmatory bifactor models (each item loading onto its respective section of the GOASSESS) to the loadings on unidimensional IRT models (S. Reise, Moore, & Maydeu-Olivares, 2011; S. P. Reise, Cook, & Moore, 2014). Item bias was assessed for each item using the following procedure: one item was removed, and then a bifactor model was fit to the remaining items. Then, a chi-squared statistic was calculated between the nested logistic models (INT=general factor, DEP=depression group factor; SOC=social anxiety group factor; GAD=generalized anxiety group factor):

Model 1: Item ~ INT + DEP + SOC + GAD

Model 2: Item ~ INT + DEP + SOC + GAD + gender + INT*gender

Chi-square statistics were then inspected. If a given statistic was an outlier, that item was then split by gender, allowing for separate parameters to be estimated for that item for each gender, and the procedure was repeated. If an item was not dichotomous, it was dichotomized by coding all non-zero values as one. This procedure was done on the adolescents and caregivers. Once an acceptable model was reached on the adolescents, the adolescents and caregivers were scored according to it using a quasi-Monte Carlo method (Jank, 2005). This method was selected because it performs well on factor models with greater than three dimensions (Chalmers, 2012). The adolescent model was chosen over the caregiver model because it has been shown that adolescents' reports of their own symptoms are more associated with diagnoses based on a structured interview than parents' reports of the adolescents' symptoms (Hope et al., 1999). Finally, convergent validity was examined by testing if females, as reported by adolescents and caregivers, had worse lifetime internalizing severity than males. Given the zero-heavy distribution of data, permutation tests were used (Wheeler, Torchiano, & R Development Core Team, 2016).

Informant Discrepancies: A correlation was calculated between the adolescents' and the caregivers' estimates of the adolescents' lifetime internalizing severity. As exploratory analyses, correlations were calculated separately for female and male adolescents, and the direction of differences in informant estimates were evaluated by gender. Then, proportion tests were conducted for every item using a Bonferroni correction comparing the two different types of disagreement: adolescent "yes" and caregiver "no", and adolescent "no" and caregiver "yes". As an exploratory analysis, these proportion tests were repeated splitting by gender.

Results

Scale Development: Item content can be found in Online Resource Table S1. Probability of responding "yes" to every item monotonically increased with trait internalizing severity, using a two-parameter logistic model for the initial trait estimates, according to both the adolescent and caregiver reports. Scree plots indicated that in both the adolescent and caregiver reports three dimensions explain most of the systematic variance in item responses (See Online Resource Figure S1). Further, comparisons of item loadings on the general dimension of the bifactor models with loadings on the one-dimensional models and the loadings in a one-factor solution indicated that, while there is a general construct (i.e., internalizing severity) underlying all item responses, there are potentially meaningful group factors as well (Table 1).

Among adolescents, item 2 (crying) displayed the largest item bias effect such that females were more likely to respond "yes" than males, given trait estimates from the bifactor model (χ^2=134.13, p_{Bon}<.05) (See Table 2). Once item 2 was split by gender, item 4 (nothing fun) still displayed noticeable item bias, but the magnitude was deemed too small to be problematic for the purposes of creating factor scores (χ^2=25.63, p_{Bon}<.05). Among caregivers, there were no large item bias effects, though six items did show some bias (2, 8, 15, 21, 25 and 32) (ITEM002: χ^2=29.49, p_{Bon}<.05). Controlling for psychopathology dimensions, caregivers of female adolescents were more likely than males to endorse items 2 (crying) and 32 (performance anxiety), while caregivers of male adolescents were more likely to endorse items 8 (tired while experiencing depressive symptoms), 15 (worry a lot for age), 21 (worry about world events) and 25 (concentration problems while worrying). After splitting item 2 by gender in the caregiver report, all item biases remained significant.

Informant Discrepancies: The correlation between adolescent and caregiver reports was .399 (95% CI: .376-.423) (Note: All correlations are Pearson correlations and all CIs are computed using Fisher's z-transformation). Within the female sample the correlation was .427 (95% CI: .394-.458), and within the male sample the correlation was .366 (95% CI: .330-.401) (See Online Resource Figure S2). In general, caregivers tended to underestimate adolescent symptoms, with an average difference of .238 (95% CI: .212-.264). This effect was larger for females than males (See Figure 2). Proportion tests for individual items revealed that, given that an adolescent and caregiver pair disagreed on an item, it was much more likely that the adolescent endorsed the symptom and the caregiver did not. This effect was much larger for females than males (Table 3). For instance, 21.5% of female endorsed item 3 (irritability) while their caregiver did not, while only 10.4% of females denied item 3 while their caregiver endorsed it. A similar but less dramatic effect is seen in the males, with 17.5% and 11.7%, respectively.

Discussion

Our results reproduce many common findings: depression, generalized anxiety and social anxiety reflect a general internalizing dimension, as well as specific dimensions for each domain (Hypothesis #1); items that ask about crying display bias such that females endorse it more frequently than males, controlling for symptom severity dimensions (Hypothesis #2); females report worse internalizing symptoms than males; and adolescents and their caregivers are largely discrepant in their reports of adolescents' internalizing symptoms (Hypothesis #4).

The current study adds to the literature by modeling lifetime internalizing severity with a bifactor model, which is better suited for the structure of psychopathology than classical test theory models (Thomas, 2011); demonstrating that item bias for "crying" is larger when using

adolescents' reports of their own symptoms than caregivers' reports of their symptoms (Hypothesis #3); establishing that the direction of disagreement between adolescents and their caregivers in a community sample is opposite that of pairs in a clinical sample, such that caregivers tend to underestimate internalizing symptoms relative to adolescents' reports (Hypothesis #4); demonstrating that caregivers of females underestimate the adolescents' internalizing symptoms more so than caregivers of males (Exploratory #4); and showing that it is much more common for the caregiver to not endorse the presence of a symptom when the adolescent endorses it than the reverse (Hypothesis #5), with this effect being much more pronounced among females than males (Exploratory #5).

While adolescent females did endorse the crying item more often than males, after controlling for symptom dimensions, there was no obvious pattern of observability in the caregivers' item bias effects. Additionally, while none of the item bias effects were as large in the caregivers' reports as the effect for the crying item was in the adolescent report, there were more biased items in the caregiver report, indicating that item bias effects may not have been diluted by random error in the caregivers' reports. Why the item bias for the crying item was so much larger in the adolescents' reports than the caregivers' reports should be investigated further. Since the item asks whether there had ever been a time where the adolescent cried a lot or felt like crying, it is possible that caregivers generally underestimated the prevalence in part because of the "felt like crying" portion of the question. It is also possible that gender norms surrounding what entails "a lot" of crying influenced caregivers' reports. Notably, the disagreement portions were very different: 26.4% of female adolescents endorsed the crying item while the caregivers did not, while only 7.9% of adolescents did not endorse while the caregivers

did. Knowing that caregivers of female adolescents under-report crying at such a rate indicates that caregivers may not be aware of the problem.

The fact that caregivers under-report female adolescents' internalizing symptoms relative to caregivers of male adolescents is striking. Based on the treatment-seeking literature, we would expect that there would be no sex differences in informant discrepancies (De Los Reyes & Kazdin, 2005), and based on evidence that females are more likely to disclose their emotions than males (Papini et al., 1990), we would expect that if there were any sex differences, they would have been in the opposite direction (i.e., caregivers of male adolescents underestimate their symptoms to a greater degree than caregivers of female adolescents). This result therefore raises several questions. Are female adolescents indeed more likely to disclose their emotions to caregivers than males? If so, and assuming this is the case in our sample, why did the caregivers of the female adolescents under-report their symptoms? Did they forget emotions that had been disclosed to them? Did they underestimate the severity of the symptoms such that they believed the symptoms did not warrant reporting? Or did females legitimately over-report the severity of their symptoms? Future studies should probe this issue further.

Results from this study cannot be expected to generalize to surveys that ask about recent symptomatology. Psychopathology fluctuates over time, and people often forget symptoms they have experienced in the past, especially if the symptoms were mild (Wells, 2004). It is also possible that caregivers tend to forget symptoms that their adolescent experienced at a higher rate than adolescents forget their own past symptoms. Future research should study this explicitly, and evaluate potential mediators such as family conflict and disclosure of symptoms.

This study is limited by the feature that most items were dichotomous, and multiple constructs were assessed with some items. For instance, item 5 assesses multiple different

problems with sleep, including difficulty falling asleep, waking up too early, and sleeping too much. It is possible that each of these sleep problems reflect internalizing severity to different degrees, and some may be associated with worse levels of internalizing severity than others. Even if the items are ultimately combined to represent a single construct, assessing them separately would allow for these questions to be investigated. Similarly, using items with Likert-scales and fitting a graded-response model (Samejima, 1969) would increase confidence in measures of lifetime internalizing severity.

Future studies should examine how adolescent-informant discrepancies on internalizing symptoms predict psychopathology severity. De Los Reyes et al. (2015) has proposed that informant discrepancies reflect meaningful social and contextual variables. Discrepancies can be due to the adolescent withholding information about their symptoms, which could reflect distrust, which in turn could be caused by caregivers, a) failing to indicate willingness to provide emotional support, b) displaying stigmatizing attitudes towards internalizing psychopathology, or even c) having committed physical abuse against the adolescent. Since informant discrepancies have the potential to reflect so many kinds of family dysfunction, they may serve as potent predictors of a variety of poor outcomes. Therefore, it is important that all research groups interested in adolescent health—mental health in particular—evaluate and exploit informant discrepancies in their quest to create a happier, healthier world.

References

Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *40*(1), 57–87.

Aune, T., & Stiles, T. C. (2009). The effects of depression and stressful life events on the development and maintenance of syndromal social anxiety: Sex and age differences. *Journal of Clinical Child and Adolescent Psychology*, *38*(4), 501–512. https://doi.org/10.1080/15374410902976304

Avenevoli, S., Stolar, M., Li, J., Dierker, L., & Ries Merikangas, K. (2001). Comorbidity of depression in children and adolescents: Models and evidence from a prospective high-risk family study. *Biological Psychiatry*, *49*(12), 1071–1081. https://doi.org/10.1016/S0006-3223(01)01142-8

Avenevoli, S., Swendsen, J., He, J. P., Burstein, M., & Merikangas, K. R. (2015). Major Depression in the National Comorbidity Survey–Adolescent Supplement: Prevalence, Correlates, and Treatment. *Journal of the American Academy of Child and Adolescent Psychiatry*, *54*(1), 37-44.e2. https://doi.org/10.1016/j.jaac.2014.10.010

Axelson, D. A., & Birmaher, B. (2001). Relation between anxiety and depressive disorders in childhood and adolescence. *Depression and Anxiety*, *14*(2), 67–78. https://doi.org/10.1002/da.1048

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Beguin, A. A. (2003). Using Classical Test Theory in Combination With Item Response Theory. *Applied Psychological Measurement*, *27*(5), 319–334. https://doi.org/10.1177/0146621603257518

Bevans, K. B., Diamond, G., & Levy, S. (2012). Screening for adolescents' internalizing symptoms in primary care: Item response theory analysis of the behavior health screen depression, anxiety, and suicidal risk scales. *Journal of Developmental and Behavioral Pediatrics*, *33*(4), 283–290. https://doi.org/10.1097/DBP.0b013e31824eaa9a

Calkins, M. E., Merikangas, K. R., Moore, T. M., Burstein, M., Behr, M. A., Satterthwaite, T. D., … Gur, R. E. (2015). The Philadelphia Neurodevelopmental Cohort: Constructing a deep phenotyping collaborative. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *56*(12), 1356–1369. https://doi.org/10.1111/jcpp.12416

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(April 2012). https://doi.org/10.18637/jss.v048.i06

Copeland, W. E., Angold, A., Shanahan, L., & Costello, E. J. (2014). Longitudinal patterns of anxiety from childhood to adulthood: The great smoky mountains study. *Journal of the American Academy of Child and Adolescent Psychiatry*, *53*(1), 21–33. https://doi.org/10.1016/j.jaac.2013.09.017

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*(4), 858–900. https://doi.org/10.1037/a0038498

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: a critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*(4), 483–509.

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. https://doi.org/10.1007/s10826-006-9031-3

Ebesutani, C., Regan, J., Smith, A., Reise, S., Higa-McMillan, C., & Chorpita, B. F. (2012). The 10-item Positive and Negative Affect Schedule for Children, Child and parent shortened versions: Application of item response theory for more efficient assessment. *Journal of Psychopathology and Behavioral Assessment*, *34*(2), 191–203. https://doi.org/10.1007/s10862-011-9273-2

Essau, C. A., & Conradt, J. (2000). Psychosocial Impairment of Anxiety Disorders in German Adolescents. *Journal of Anxiety Disorders*, *14*(3), 263–279.

Ferdinand, R. F., Van Der Ende, J., & Verhulst, F. C. (2004). Parent-Adolescent Disagreement Regarding Psychopathology in Adolescents from the General Population as a Risk Factor for Adverse Outcome. *Journal of Abnormal Psychology*, *113*(2), 198–206. https://doi.org/10.1037/0021-843X.113.2.198

Fichter, M. M., Quadflieg, N., Fischer, U. C., & Kohlboeck, G. (2010). Twenty-five-year course and outcome in anxiety and depression in the upper bavarian longitudinal community study. *Acta Psychiatrica Scandinavica*, *122*(1), 75–85. https://doi.org/10.1111/j.1600-0447.2009.01512.x

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement*, *63*(1), 65–74. https://doi.org/10.1177/0013164402239317

Gibbons, R. D., & Hedeker, D. R. (1992). *Full-information item bi-factor analysis*. *57*(3), 423–436.

Hankin, B. L. (2009). Development of sex differences in depressive and co-occurring anxious symptoms during adolescence: Descriptive trajectories and potential explanations in a multiwave prospective study. *Journal of Clinical Child and Adolescent Psychology*, *38*(4), 460–472. https://doi.org/10.1080/15374410902976288

Hetrick, S. E., Parker, A. G., Robinson, J., Hall, N., & Vance, A. (2012). Predicting suicidal risk in a cohort of depressed children and adolescents. *Crisis*, *33*(1), 13–20. https://doi.org/10.1027/0227-5910/a000095

Hope, T. L., Adams, C., Reynolds, L., Powers, D., Perez, R. A., & Kelley, M. Lou. (1999). Parent vs. self-report: Contributions toward diagnosis of adolescent psychopathology. *Journal of Psychopathology and Behavioral Assessment*, *21*(4), 349–363. https://doi.org/10.1023/A:1022124900328

Irwin, D. E., Stucky, B., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J. S., … DeWalt, D. A. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research*, *19*(4), 595–607. https://doi.org/10.1007/s11136-010-9619-3

Jank, W. (2005). Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM. *Computational Statistics and Data Analysis*, *48*(4), 685–701. https://doi.org/10.1016/j.csda.2004.03.019

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., … Ryan, N. (1997). Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*(7), 980–988. https://doi.org/10.1097/00004583-199707000-00021

Kazdin, A. E. (1989). Identifying depression in children: A comparison of alternative selection criteria. *Journal of Abnormal Child Psychology*, *17*(4), 437–454.

Leikanger, E., Ingul, J. M., & Larsson, B. (2012). Sex and age-related anxiety in a community sample of Norwegian adolescents. *Scandinavian Journal of Psychology*, *53*(2), 150–157. https://doi.org/10.1111/j.1467-9450.2011.00915.x

Merikangas, K. R., He, J. P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., … Swendsen, J. (2010). Lifetime prevalence of mental disorders in U.S. adolescents: Results from the national comorbidity survey replication-adolescent supplement (NCS-A). *Journal of the American Academy of Child and Adolescent Psychiatry*, *49*(10), 980–989.

https://doi.org/10.1016/j.jaac.2010.05.017
Nolen-Hoeksema, S., & Jackson, B. (2001). Mediators of the gender difference in rumination. *Psychology of Women Quarterly*, *25*(1), 37–47. https://doi.org/10.1111/1471-6402.00005
Organization, W. H. (2004). *Prevention of Mental Disorders: Effective Interventions and Policy Options*. Geneva.
Papini, D. R., Farmer, F. F., Clark, S. M., Micka, J. C., & Barnett, J. K. (1990). Early adolescent age and gender differences in patterns of emotional self-disclosure to parents and friends. *Adolescence*, *25*(100), 959–976.
R Development Core Team. (2014). *A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing*.
Reise, S., Moore, T., & Maydeu-Olivares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement*, *71*(4), 684–711. https://doi.org/10.1177/0013164410378690
Reise, S. P. (2014). Item Response Theory. In *The Encyclopedia of Clinical Psychology* (pp. 1–10). https://doi.org/10.1002/9781118625392.wbecp357
Reise, S. P., Cook, K. F., & Moore, T. M. (2014). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In *Handbook of item response theory modeling* (pp. 31–58). Routledge.
Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor Models and Rotations. *Journal of Personality Assessment*, *92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477.Bifactor
Revelle, W. R. (Photographer). (2017). *psych: Procedures for Personality and Psychological Research*.
Rueger, S. Y., Malecki, C. K., Pyun, Y., Aycock, C., & Coyle, S. (2016). A meta-analytic review of the association between perceived social support and depression in childhood and adolescence. *Psychological Bulletin*, *142*(10), 1017–1067. https://doi.org/10.1037/bul0000058
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100.
Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety*, *25*(7), 34–46. https://doi.org/10.1002/da.20432
Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*(3), 291–307. https://doi.org/10.1177/1073191110374797
Van Beek, Y., Hessen, D. J., Hutteman, R., Verhulp, E. E., & Van Leuven, M. (2012). Age and gender differences in depression across adolescence: Real or "bias"? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *53*(9), 973–985. https://doi.org/10.1111/j.1469-7610.2012.02553.x
Wheeler, R. E., Torchiano, M., & R Development Core Team. (2016). lmPerm: Permutation tests for linear models. *Http://CRAN.R-Project.Org/Package=lmPerm.*, 2–24.

Table 1. Item loadings for the unidimensional and bifactor models

	Adolescents					Caregivers				
Item	One-Factor	INT-Bifactor	DEP-Bifactor	GAD-Bifactor	SOC-Bifactor	One-Factor	INT-Bifactor	DEP-Bifactor	GAD-Bifactor	SOC-Bifactor
ITEM001	0.75	0.889	0.092	0	0	0.764	0.647	0.616	0	0
ITEM002 (F)	0.63	0.847	0	0	0	0.699	0.634	0.533	0	0
ITEM002 (M)	0.545	0.76	-0.075	0	0	0.695	0.676	0.421	0	0
ITEM003	0.593	0.835	-0.009	0	0	0.64	0.597	0.556	0	0
ITEM004	0.601	0.778	0.023	0	0	0.733	0.625	0.562	0	0
ITEM005	0.981	0.857	0.461	0	0	0.964	0.565	0.806	0	0
ITEM006	0.974	0.83	0.493	0	0	0.939	0.526	0.811	0	0
ITEM007	0.971	0.836	0.472	0	0	0.95	0.571	0.777	0	0
ITEM008	0.984	0.847	0.49	0	0	0.96	0.567	0.801	0	0
ITEM009	0.985	0.853	0.482	0	0	0.967	0.592	0.786	0	0
ITEM010	0.977	0.866	0.43	0	0	0.949	0.601	0.759	0	0
ITEM011	0.979	0.823	0.546	0	0	0.949	0.555	0.821	0	0
ITEM012	0.978	0.821	0.549	0	0	0.95	0.552	0.823	0	0
ITEM013	0.947	0.785	0.519	0	0	0.942	0.571	0.774	0	0
ITEM014	0.803	0.431	0	0.829	0	0.84	0.661	0	0.711	0
ITEM015	0.765	0.547	0	0.636	0	0.825	0.72	0	0.523	0
ITEM016	0.766	0.403	0	0.799	0	0.809	0.529	0	0.775	0
ITEM017	0.701	0.391	0	0.732	0	0.728	0.491	0	0.722	0
ITEM018	0.736	0.458	0	0.715	0	0.804	0.58	0	0.702	0
ITEM019	0.812	0.464	0	0.788	0	0.822	0.725	0	0.532	0
ITEM020	0.802	0.55	0	0.695	0	0.812	0.75	0	0.467	0
ITEM021	0.707	0.418	0	0.707	0	0.735	0.633	0	0.539	0
ITEM022	0.875	0.535	0	0.82	0	0.904	0.714	0	0.68	0
ITEM023	0.846	0.57	0	0.758	0	0.896	0.74	0	0.638	0
ITEM024	0.839	0.588	0	0.697	0	0.881	0.745	0	0.6	0
ITEM025	0.853	0.566	0	0.761	0	0.89	0.768	0	0.581	0
ITEM026	0.849	0.573	0	0.749	0	0.9	0.733	0	0.642	0
ITEM027	0.793	0.534	0	0.699	0	0.837	0.687	0	0.63	0
ITEM028	0.83	0.561	0	0.75	0	0.883	0.733	0	0.629	0
ITEM029	0.426	0.424	0	0	0.662	0.501	0.487	0	0	0.711
ITEM030	0.407	0.38	0	0	0.611	0.517	0.504	0	0	0.657
ITEM031	0.348	0.284	0	0	0.778	0.459	0.377	0	0	0.812
ITEM032	0.409	0.366	0	0	0.757	0.53	0.457	0	0	0.752
ITEM033	0.418	0.397	0	0	0.74	0.494	0.4	0	0	0.798
ITEM034	0.446	0.426	0	0	0.727	0.565	0.547	0	0	0.729
ITEM035	0.64	0.521	0	0	0.658	0.65	0.609	0	0	0.643

Note: INT=general factor, DEP=depression group factor; SOC=social anxiety group factor; GAD=generalized anxiety group factor

Table 2. Chi-squared statistics for item bias effects

	Adolescents		Caregivers	
Item	χ^2 #1	χ^2 #2	χ^2 #1	χ^2 #2
ITEM001	5.47	7.17	1.67	1.06
ITEM002	134.13* (F)		29.49* (F)	
ITEM003	0.16	1.36	5.5	6.63
ITEM004	36.56* (M)	25.63* (M)	3.14	3.08
ITEM005	1.1	2.15	0.7	0.71
ITEM006	11.05	11.84	1.18	1.26
ITEM007	4.9	3.68	1.73	1.83
ITEM008	2.84	3.72	15* (M)	15.9* (M)
ITEM009	17.02* (M)	15.98* (M)	12.61	13.07* (M)
ITEM010	1.03	0.43	2.53	2.2
ITEM011	4.84	3.63	1.4	1.27
ITEM012	1.68	2.5	0.82	0.93
ITEM013	1.06	1.09	1.32	1.51
ITEM014	5.46	4.94	0.47	0.49
ITEM015	3.24	2.35	33.94* (M)	32.79* (M)
ITEM016	4.08	4.02	4.48	4.23
ITEM017	0.7	0.47	4.87	3.97
ITEM018	0.16	0.19	0.31	0.5
ITEM019	0.97	1.12	5.73	6.45* (F)
ITEM020	2.44	2.17	3.46	4.28
ITEM021	5.5	5.23	14.14* (M)	14.2* (M)
ITEM022	2.3	2.24	4.59	4.51
ITEM023	2.26	2.69	2.1	2.2
ITEM024	5.01	5.72	6.78	6.98* (M)
ITEM025	4.49	4.89	27.28* (M)	25.52* (M)
ITEM026	2.07	2.64	12.98	13.14* (F)
ITEM027	13.25* (M)	12.93	3.3	3.34
ITEM028	7.36	8.65	1.47	1.63
ITEM029	0.73	1.18	7.99	7.75* (F)
ITEM030	4.97	4.51	0.07	0.05
ITEM031	4.85	4.68	1.39	1.38
ITEM032	9.38	9.75	15.08* (F)	14.06* (F)
ITEM033	2.27	2.11	3.05	2.60
ITEM034	6.31	5.94	10.66	9.53* (M)
ITEM035	0.05	0.05	4.8	5.04

*F=Main effect for gender is positive for females, M=Main effect for gender is positive for males; The nested models are significantly different (Bonferroni-adjusted p < .05)

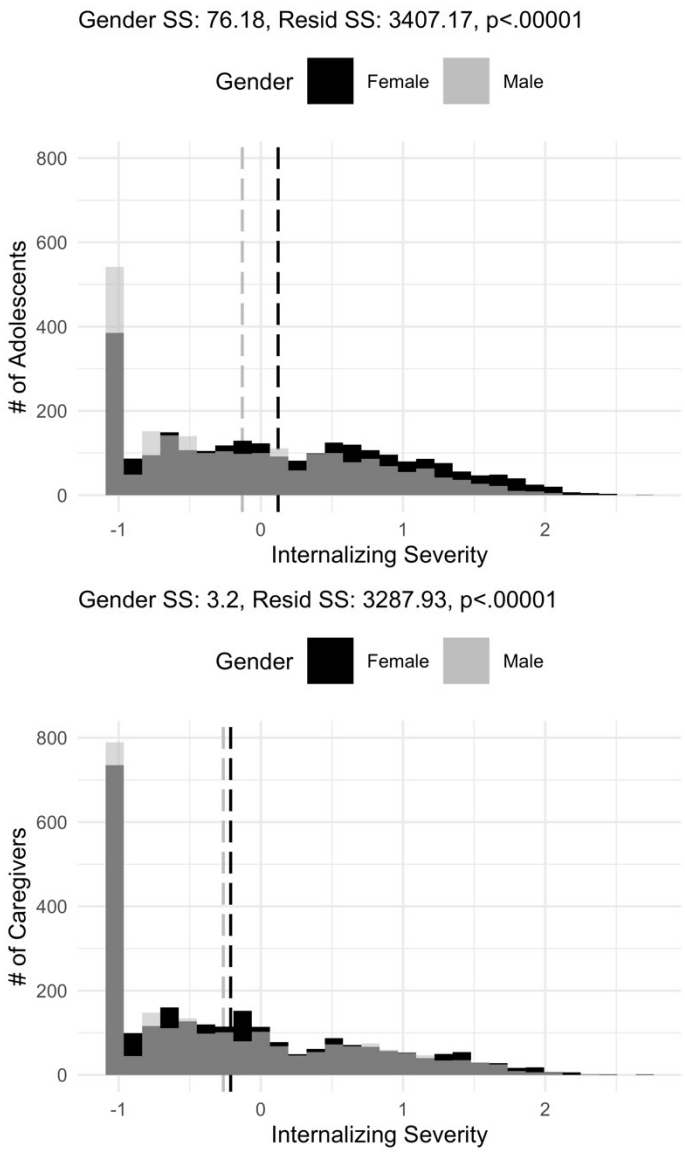
Table 3. Agreement and disagreement proportions for each item by adolescents' gender

ale Adolescents

Agree		Disagree	
N C N	A Y C Y	A Y C N	A N C Y
5.90%	10.50%	15.50%	8.1%*
53%	12.70%	26.40%	7.9%*
3.30%	14.80%	21.50%	10.4%*
66%	7.60%	18.10%	8.3%*
7.90%	5.20%	10.50%	6.4%*
0.10%	3.90%	10.60%	5.4%*
3.30%	2.50%	9.10%	5.1%*
6.70%	5.40%	11.80%	6.1%*
8.60%	4.10%	11.50%	5.9%*
2.70%	3.60%	9.70%	4%*
1.20%	6.80%	14.40%	7.6%*
1.80%	6.70%	13.90%	7.5%*
8.20%	2.40%	5.50%	3.8%*
0.70%	33.30%	23.50%	12.5%*
5.30%	15.70%	17.50%	11.5%*
5.40%	18.70%	23.30%	12.6%*
7.10%	6.50%	18.50%	7.9%*
2.10%	8.40%	17.30%	12.2%*
6.40%	17.60%	25.10%	10.9%*
8.60%	12.40%	18.50%	10.5%*
9.50%	4.60%	20%	5.8%*
0.50%	21.10%	27.40%	10.9%*
7.20%	8.10%	15.70%	9%*
72%	5.70%	13.30%	9.1%*
9.80%	9.80%	22.40%	8%*
6.90%	11.60%	17.90%	13.6%*
2.70%	2.70%	7.70%	7%
67%	8.40%	15.50%	9.1%*
6.20%	11.80%	19.30%	12.7%*
3.60%	4.80%	13.80%	7.8%*
2.50%	14.80%	23.40%	9.2%*
5.30%	17.60%	26.50%	10.5%*
5.40%	10.70%	23.90%	10%*
0.60%	14.60%	23.70%	11.1%*
0.80%	1.10%	4.20%	3.90%

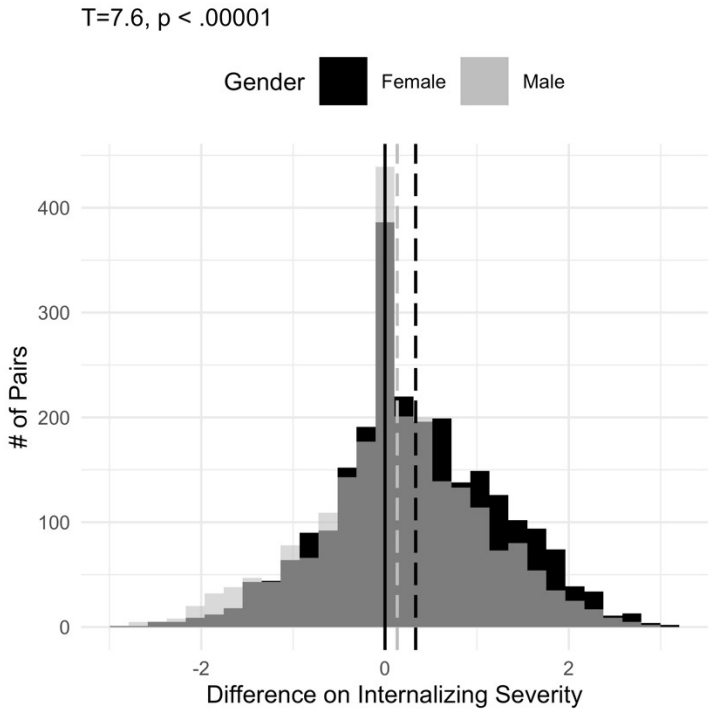
proportions within gender are significantly different

a. Adolescents: Internalizing Severity, b. Caregivers: Internalizing Severity



Note: Vertical dashed lines indicate means for female and male adolescents and caregivers, respectively. Sums of squares (SS) for permutation tests using 5000 iterations to test for mean differences between genders are given.

Difference Between Informants



Note: Vertical dashed lines indicate means for differences on the lifetime internalizing severity dimension between females and their caregivers and males and their caregivers. The two-sided t-test statistic is given.