Author’s response to reviews

Title: Trends in childhood and adolescent internalizing symptoms: results from Swedish population based twin cohorts

Authors:

Karolina Sorman (Karolina.Sorman@ki.se)

Natalie Durbeej (Natalie.Durbeej@ki.se)

Eva Norén-Selinus (Eva.Noren@ki.se)

Sebastian Lundström (sebastian.lundstrom@gnc.gu.se)

Paul Lichtenstein (Paul.Lichtenstein@ki.se)

Clara Hellner (Clara.Gumpert@ki.se)

Linda Halldner (Linda.Halldner@ki.se)

Version: 1 Date: 18 Apr 2019

Author’s response to reviews:

Author response to reviews of

Manuscript PSYO-D-19-00002R1

Trends in childhood and adolescent internalizing symptoms: results from Swedish population based twin cohorts

Karolina Sorman, Ph.D.; Natalie Durbeej; Eva Norén-Selinus; Sebastian Lundström; Paul Lichtenstein; Clara Hellner; Linda Halldner

re-submitted to BMC Psychology

April 18th, 2019

Dear Dr. Harris

Thank you very much for taking the time to review our manuscript for publication in BMC Psychology, and for the opportunity to re-submit the manuscript.
We have strived towards addressing all points raised by you and the Reviewers by making appropriate changes in the revised manuscript and by providing responses in this cover letter. In the following letter, we address all comments and concerns point-by-point and declare how we revised the manuscript accordingly.

Reviewer #1

Tracey Wade (Reviewer 1): The rationale for this manuscript is the mixed findings with respect to the increase over time across cohorts for internalizing symptoms, with a special focus on boys, given the literature is clearer with respect to girls. The design and results of the current manuscript do not really contribute to clarification of these mixed findings, and thus the manuscript makes only a small contribution to the literature in terms of adding more inconsistent results, albeit using a large sample with well-validated measures.

I have a few substantial queries for the authors to address:

1. Are there any predictors of non-involvement in the sample, that can explain why 40% did not respond?

Response: In the initial analyses of the CATSS study a response analysis showed: “Systematic analyses for differences between non-responders and responders (i.e., children for whom parents declined or consented to interviews) were performed on the basis of an anonymized merge between the CATSS database for the first 11,222 participants and official files such as the National Board of Health and Welfare database on socio-economic circumstances, in- and outpatient diagnostics, and pharmacological treatment. Non-responders to the CATSS 9/12 telephone interviews were more likely than responders to have: a parent treated in psychiatric settings (9.6% of the non-responders vs. 6.3% of the responders), a father convicted of a felony (11.2% vs. 7.2%), a mother convicted of a felony (1.6% vs. 0.7%), a divorced mother (16.4% vs. 12.5%), a divorced father (16.4% vs. 12.4%), or to belong to low socio-economic strata (26.6% vs. 21.9%). Non-responders to the telephone interviews also had 2.1% ADHD as compared to 1.6% among responders, 0.95% ASD versus 0.84%, 2.0% LDs versus 0.99%. Among non-responders, 1.8% had been prescribed psychopharmacological treatment for ADHD as compared to 1.4% of the responders.” (Anckarsäter et al 2011), as shortly referred to in the discussion (first paragraph p.12). There has been declining response rates in the CATSS study and its spin-off studies over time, as well as in all questionnaire studies in Sweden, and internationally (see e.g. Beullens, K., Loosveldt G., Vandenplas C. & Stoop I. (2018). Different reasons for this decline have been suggested (as the rise of online surveys, and information requests, greater awareness of privacy issues). Thus we can expect that the non-responders are to greater extent those with more societal challenges, and thus, as we point out in the discussion)
we are more likely to underestimate rather than overestimate, the increase of reported psychiatric symptoms in the present study.

2. Why rely only on the cut-off for SDQ, could not a total score be examined using linear regression? Dichotomization, whilst suitable for comparison, does decrease power.

Response: We have calculated linear regressions with cohort as the independent variable and individual total scale scores as dependent variables (see the statistical analyses section). These analyses were calculated in order to explore the relations between cohort and internalizing symptoms (as continuous variables). However, in order to explore whether proportions of participants with internalizing symptoms changed over time, we used cut-off scores of the scales to define the proportions of participants with elevated scores (as dichotomous variables), which were tested by Cochrane Armitage tests for trend. We are aware that dichotomization reduces power but given our research questions, we believe that we have analyzed our data in a correct way.

3. Can the authors provide other and more reliable, indicators of internal consistency for their measures, such as a H coefficient?

Response: We believe that we have provided the most widely used measure for internal consistency, namely Chronbach’s alpha (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4816140/). The main aim of the study was not to report on the reliability of the scales but we chose to report internal consistency for descriptive purposes. If the reviewer thinks that we should report the H coefficient instead of Chronbach’s alpha, could she inform us on how to calculate this and/or provide examples of further reading (articles, books etc), please? This would be highly appreciated.

4. Table 2: the R2 for sMFQ is not very informative? is there a more informative statistic?

Response: We agree that R2 is not informative. We have therefore removed all R2 values. Additionally, the aim of the study was not to explore the proportion of explained variance in the outcomes. Thus, we have not used another statistic instead of R2 (see Table 2).

5. Were any of the statistics corrected for correlated data (twin correlations)? it may be necessary to look at Twin 1 and 2 separately.
Response: Yes, the statistics were corrected for correlated data. A cluster robust sandwich estimator was applied to adjust the standard errors for the nested twin data when computing the regression models. We have made a clarification on this in the statistics section. Given that we have made this correction, we do not think it is necessary to conduct separate analyses with twin 1 and 2, respectively.

6. Throughout the manuscript there is too much reliance on p values, report ES and 95% CI instead.

Response: We have added some statistics to the results section (see Table 1), namely Cohen’s d for effect size to explore sex differences in the scale scores. For the linear regressions, we report both p-values and 95% CI (see Table 2). For the Cochrane Armitage Test for Trend, we only report p-values since it is not possible to obtain 95% CI from these analyses, as far as we are aware (see Table 3). We hope that these additions are satisfying according to the reviewer’s suggestion.

Minor issues:

1. Page 3, line 34: The word "international" is not required; more informative would be across how many countries have studied been conducted?

Response: We have clarified this by adding examples of countries included in the systematic review we discuss (Bor et al. 2014)

2. Pag4, lines 5-10: A single sentence does not make a paragraph - please join this to the previous paragraph.

Response: We agree, these sentences have been joined to the previous paragraph.

3. Page 7, line 5: a comma is missing

Response: We have added this comma.
Nora Trompeter (Reviewer 2): This manuscript addresses an important and interesting topic by investigating time-trends in internalising symptoms for boys and girls. The manuscript is quite strong, including a large, representative sample and a straightforward analytic plan. However, there are some minor issues warranting attention, as detailed below (by section).

Abstract: Clear description of the study, however, it would be helpful to state how many birth cohorts were examined and the time frame in which data were collected (i.e., which birth cohorts were included)

Response: This information has now been added to the abstract.

Introduction:

In the discussion regarding gender differences it would be helpful to include a discussion on why these discrepancies might be relevant and what this could mean in a clinical/policy setting. Furthermore, when discussing previous literature it would be appropriate to mention the international studies, which are referred to in the discussion, and integrating the discussion on findings from both Swedish research and international research in this section.

Response: We agree with this and have extended that section by adding (i) hypotheses about why these sex difference might occur and what the implications might be in a clinical/policy setting, (ii) previous findings from international literature regarding sex differences in internalizing symptoms, and (iii) a reference to one of the Swedish studies included, regarding potential explanations for why internalizing symptoms might be more common in girls specifically during adolescence.

Page 3, lines 17-22: Could you please clarify what is being compared. Are boys more likely to exhibit externalising compared to internalising or are they more likely to exhibit externalising compared to girls? Same applies to the following statement regarding girls.

Response: We have clarified the direction of the comparisons (i.e., that internalizing symptoms are more prevalent in girls than in boys particularly during puberty, and that externalizing symptoms are more prevalent in boys than in girls

Method: The authors make it clear in their description of the sample, that this sample has been reported on in previous publications. However, it would be nice to get a brief overview of the sample characteristics and/or be provided with a reference for another paper where the sample is
described in more detail. It would also be helpful to discuss what implications the use of a twin sample has on generalisability.

Response: The best description of the CATSS sample is in Anckarsäter et al, Twin Res Hum Genet 2011, which is referred to in the methods section. See also answer to Reviewer #1, Question 1.


Were there any participants who were captured at 9yo and 15yo? And if so, what are the ramifications for having captured participants twice?

Response: All twins starting with the cohorts 1994 (and their parents), were invited to take part in the questionnaire follow-up at age 15 (see Anckarsäter et al 2011). In a previous study on the CATSS sample (Törn P, Pettersson E, Lichtenstein P, Anckarsäter H, Lundström S, Hellner Gumpert C, Larsson H, Kollberg L, Långström N, Halldner L. Childhood neurodevelopmental problems and adolescent bully victimization: Population-based, prospective twin study in Sweden. Eur Child Adolesc Psychiatry. 2015; 24(9):1049-59) 60.5 % of the participants from the baseline interview had a follow-up at age 15. However, response rates at age 15 were lower for children screening positive for neurodevelopmental disorders in the CATSS parent interview at age 9 For children screening positive for ADHD at age 9, the response rate at age 15 was 46 %; for children screening positive for ASD at age 9, response rate at age 15 was 33 %; for children screening positive for Developmental Coordination Disorder [DCD] at age 9, response rate at age 15 was 53 %; for children screening positive for Tic Disorder [TD] at age 9, response rate at age 15 was 52 %; for children screening positive for Learning disorder [LD] at age 9, response rate at age 15 was 47 %, and also children who at age 9 were bullied according to the CATSS parent interview, the response rate at age 15 was 52.9 %. This may affect our results, but rather in the direction that we in the present study underestimate internalizing symptoms, as is the case
for a lower socioeconomic status for non-responders as already stated in the discussion in the present paper.

We are now planning a follow-up to the present study regarding the participants captured twice with the aim to identify any patterns in the development of internalizing symptoms, but we do not have those data yet.

A minor point: sMFQ has three Cronbach's alphas reported for boys and girls. Could you please clarify which alpha is the one for boys and which one is for girls?

The internal consistencies for the SDQ are quite low, how does this compare to other studies? Is this a common issue or specific to this sample?

Response: We have now clarified which alpha coefficients are for boys and girls, respectively (.80 for both boys and girls)

Statistical analyses: Statistical analyses appear well-suited and are described well.

However, there was no mention of assumption tests. Could you please elaborate on whether the assumptions for regressions were met?

Response: We have checked the distribution of the scales and they appear to be approximately normally distributed. Based on these results and the fact that linear regression models give the best power, we chose this kind of regression to explore the relation between birth cohort and internalizing symptoms.

Furthermore, if both sets of twins were used, what were the repercussions on the assumption of independence?

Response: Both sets of twins were used and the statistics were corrected for correlated data. More specifically, a cluster robust sandwich estimator was applied to adjust the standard errors for the nested twin data) when computing the regression models. We have made a clarification on this in the statistics section.

There appears to be high variability in scale completion, it would be good to comment on any differences between participants who completed vs not completed measures.
Response: We agree with the reviewer and have therefore investigated sex differences between participants who completed and who did not complete the scales (as we merely had data on sex as a comparison variable). These results have been added to the results section in the manuscript.

Results:

Clear and concisely written.

However, in your description of the results could you elaborate on regression analyses in terms of the direction of the relationship?

Response: Most beta coefficients of the regressions were positive (see Table 2). The exception was the regression of the SDQ self-report among boys for which the beta was negative (-.009). However, this regression was not significant. Overall, our results suggest that the relations between birth cohort and outcomes had positive directions. We have added this information to the result section of the manuscript.

Discussion

Well written.

It would be helpful to elaborate on difference between parent-reported and child-reported symptoms, as this appears to be an interesting finding in the current study. This might be especially interesting since results from the 9yo group were based solely on parent-report measures.

Response: We totally agree with this point and have made a comment on potential differences between parent-and self reports in two different sections in the discussion:

(1) under “levels of internalizing symptoms” we have added the sentence “It is worth highlighting that this data is based on parent-reports which could differ significantly from children’s own self-ratings (e.g., parents potentially over-or under estimating the distress in the child), when discussing levels of internalizing symptoms in the 9-year olds”.

(2) under “percentages of participants exceeding cut-off scores” we have added the sentence “Again, this finding should be seen in the light of parent-ratings potentially underestimating true levels of anxiety and depression symptoms in their child”

It would also be beneficial to link results back to studies discussed in the introduction, especially those conducted in Sweden.
Response: We agree with the reviewer and have added comparisons in the text to the Swedish studies brought up in the introduction, i.e., regarding the finding that girls had higher levels of internalizing symptoms than boys, at age 15.

Small point: please ensure to not use gender and sex interchangeably, as they are different constructs

Response: We have replaced gender with “sex” (or “boys and girls”) throughout the manuscript.

David Chinn (Reviewer 3): This is a generally well written and interesting paper with some important implications. Some editing, especially of language, is required. There are some limitations, expanded upon below, which need further description in the text prior to publication

Page 3 - the method could be clarified a little - this study involved gathering samples of both 9 and 15 year olds, each gathered from successive birth cohorts from 1998-2008 (for 9 year olds) and 1994-2001 (for 15 year olds).

Response: We agree with the reviewer and have clarified accordingly under “participants” and also in the abstract.

Page 4 - lines 2-3. Should read "some form of mental health problem."

Response: We have revised accordingly!

Page 4 - lines 10-12. Need to re-word. Rather than "internalising symptoms being associated with a range of symptoms," surely the negative effects cited are not "symptoms" but instead difficulties that may be related to the aforementioned internalising symptoms.

Response: We have re-worded this to “difficulties” (negatively affecting everyday life for youths).

Page 4 - lines 53-54. Clarify the text- this study presumably demonstrated an increase in girls but not boys demonstrating these psychosomatic health problems.
Response: We have clarified that the results demonstrated successive increases in psychosomatic health problems specifically in girls across the study period.

Page 5 - line 1. Reword - it would read better if it said "it is unclear if this reflects an increase in specific psychiatric symptoms or an increased propensity to seek mental health services.

Response: We have re-phrased this sentence to “It is unclear whether this reflects an increase in some form of mental health problem or an increased propensity to seek mental health services” to also adhere to your previous comment (see above).

Page 6 - line 14. The SCARED seemed a good instrument to use.

Response: In the description of the instrument, we point out that SCARED is one of the most commonly used scales to assess anxiety symptoms in children (14). SCARED is considered well-suited for the use in community samples (15), across a wide range of cultural context.

Pages 6-7. There appeared to be far greater depth to the investigation of anxious/depressive problems at age 9 (with two instruments being used that look into this specifically) than at age 15 - when the only measurement that was done was a 5 item sub scale into emotional symptoms. This there any reason for this significant discrepancy? This could be expanded upon in limitations.

Response: It would have been desirable to have separate scales for depression and anxiety measures also for the 15-year old participants but since the CATSS study’s arrangement has prioritized response rate, sacrifices have been done in regards of specified scales in the follow-up studies (i.e. for example at age 15). What is included in the follow-up at age 15 is reported in Table 2 in Anckarsäter et al, Twin Res Hum Genet 2011. Also it would of course have been preferable to have self-reported data from the 9-year-old participants too, but in an epidemiological study of this size this is rarely feasible.

We have now added information in the limitations section.

Page 8 - lines 43-44. Clarify text - this may be rewritten as "statistically significant for the total sample, and for both girls and boys respectively when their results were analysed separately." Or something similar.

Response: We have revised the text accordingly and agree that it made it clearer!
Page 9 - line 58-59 should be "percentages" given the reference to multiple groups. Similar on the next page in the same sentence this should read "cut-offs."

Response: We have revised the text accordingly and agree that it made it clearer!

When reporting these results on pages 9 and 10, it should be emphasised at this stage that the only data collected related to parent reports. Parent reports internalising difficulties differ significantly from children's own self-ratings (often under estimating the distress that children feel), which may have given very different results if there were collected.

Response: This is a valid and important point to raise. We have added a sentence stating that parent reports might be lower than self-reports, in the section discussing levels of internalizing symptoms at age 9.

The same point could be made on page 10, lines 21-22. Extending this, when discussing the different findings for 15 year olds between self and parent ratings, the inferior ability of parents to rate internalising symptoms compared to self ratings could again be noted. The point is well made on page 12 line 43-45, but could also be noted earlier in the paper too for ease of interpretation

Response: We agree that it is good to state this valid point at several places throughout the manuscript. We have added the sentence that the “finding should be seen in the light of parent-ratings potentially underestimating true levels of anxiety and depression symptoms in their child”, to this section.

Page 10 - line 51-52. Maybe don't use "robust finding" when referring to previous research that these findings appear to contradict in terms of the specificity of only girls experiencing increasing internalising symptoms at age 15.

Response: We have re-phrased this sentence to “These trends could challenge the replicated finding that the rise in symptoms of anxiety and depression during the past decades has been relatively specific to girls (Collishaw, 2015)”.

Page 11 - line 53-54 doesn't read well and could be reworded.
Response: This has been reworded to “the results are not easily generalized to other cultural contexts.

Page 12- line 2-3. This is a troubling finding in light of possible future psychiatric problems. These findings may increase the risk that a given youth will develop psychiatric problems, but it is by no means certain that they definitely will.

Response: We have adjusted this sentence accordingly.

Page 14 line 1. There may be clinical implications too. It is not realistic for clinicians to repeatedly assess boys and girls for internalising problems, but the results of this study should prompt clinicians to have a higher index of suspicion for the presence of internalising symptoms in children and young people, in both boys and girls.

Response: The last sentence in the conclusion was poorly phrased and has now been rephrased according to your much better suggestion.

Bryan Rodgers, Ph.D. (Reviewer 4): This manuscript reports analyses of data from successive birth cohorts of twins in Sweden, with the aim of identifying historical trends in internalizing symptoms for cohorts born 1994 to 2008. These data cover information on 9-year-old twins (cohorts 1998 to 2008) and 15-year-old twins (1994 to 2001) using parent reports for the younger age and both parent reports and self-reports at age 15. Analyses for 9-year-olds indicate an increasing trend in internalizing symptoms based on the SCARED measure but no or trivial trends based on the sMFQ measure. For 15-year-olds, increases were found for self-reports on the SDQ but there was little change using the parent-report SDQ. Where increasing trends were found, these appeared similar in boys and girls. The authors conclude that: "Trends of increasing internalizing symptoms are not exclusive to girls. A clinical implication of the results is that internalizing symptoms should be assessed across genders, from childhood and throughout adolescence."

The strengths of this study include the systematic collection of data from successive annual birth cohorts and the overall sample sizes (over 17,000 at age 9 and almost 11,000 at age 15). However, the manuscript has a number of weaknesses including: little consideration of why the study is important/significant, a very poor literature review, insufficient information on the
validity of the outcome measures, inadequate statistical tests to investigate possible gender differences, and insufficient discussion (or explanation) of apparently conflicting findings.

The final conclusions, quoted above, reflect this confusion. Is the primary implication of this research really clinical, and are the authors really suggesting some form of continuous assessment (research or general surveillance) of childhood and adolescent symptoms? How did the particular findings reported lead to such a conclusion?

At this stage, I think it would not be helpful to provide detailed feedback on all aspects of the manuscript when several fundamental features need to be addressed. The following are some key aspects for consideration in revising the report.

1) The literature in this field has suggested a number of reasons as to why it is important to investigate trends in the mental health of children and adolescents. Some acknowledgment of this literature would be helpful in pre-empting a "so what?" response. Even one or two sentences would set the scene for the manuscript.

Response: We have added several sentences to better “set the stage” for the manuscript, including: (i) understanding time trends in mental health symptoms among youth is one important aspect in the prevention of mental illness in this population, and that (ii) studying temporal trends adds important knowledge to the investigation of time related changed in diagnosis and treatment of youth psychiatric disorders.

2) There have been tricky conceptual issues with the measures used in this field in the past. The term "internalizing symptoms" is not well defined and the current paper is not clear in whether it is synonymous with "anxiety and depression" or covers other things. Historically, parent (and teacher) reports of internalizing behaviours have shown very low correlations with self-reports (as well as with each other) and self-reported introversion has been more consistently aligned with these measures than with measures of anxiety or depression. It would be good to know whether the measures reported on here have circumvented such difficulties but the information provided does not provide reassurance.

Response: The terms of externalizing and internalizing problems were introduced by Achenbach in 1966 (The classification of children’s psychiatric symptoms: a factor-analytic study. Psychol Monogr 1966;80:1-37). These are used for broad-band groupings of behavioral, emotional, and social problems, to be able to study psychopathology without limiting symptoms into narrow diagnostic categories. In the strict sense of the meaning of internalizing symptoms we have not used scales appropriate to measure this dimension fully, but the emotional domain of SDQ solely consists of items regarding internalizing problems (“I get a lot of headaches,
stomach-aches or sickness”, "I worry a lot", “I am often unhappy, depressed or tearful”, “I am nervous in new situations. I easily lose confidence”, and “I have many fears, I am easily scared”). Further, symptoms of depression and anxiety as measured by sMFQ and SCARED, respectively are reckoned internalizing behavior. It would have been desirable to have separate scales for depression and anxiety measures also for the 15-year old participants but since the CATSS study’s arrangement has prioritized response rate, sacrifices have been done in regards of specified scales in the follow-up studies (i.e. for example at age 15). Also it would of course have been preferable to have self-reported data from the 9-year-old participants too, but in an epidemiological study of this size this is not feasible. Thus, unfortunately, we have not been able to circumvent the problem of poor correlation between parent-, and self-reports for the measures use in the 9-years-old cohorts. Your point is of course very important, and also in line with our findings of different results in the adult-, and self-reports for the 15-years-old cohorts in our study. Our purpose of the study was however to study if internalizing symptoms assessed in the CATSS study, had been changing between cohorts, in the light of repeated media reports of poorer mental health in children and adolescents often referring to school questionnaires with few and non-validated items on “internalizing” symptoms, or to increasing number of health care visits.

The cited paper for the validity of the SCARED (ref 17) leaves doubts in terms of its generalizability to the present study and also because previous literature referred to in that publication is inconsistent with the findings reported by Hariz et al. The present study has the capacity to examine and report the correlations between self-reports and parent reports at age 15, and this would be reassuring if they indicate that the two sources are indeed assessing similar constructs.

Response: It would be preferable if there was a validation study on SCARED using a Swedish sample including 9-year-olds. Nevertheless, we believe that previous validation studies of the scale support that it is measuring anxiety symptoms rather than symptoms of depression.

The literature on correlation between self- and parent- reports is complicated and heterogeneous. For example, there was a marginal affect of age with higher agreement on the emotional symptoms in younger adolescents (Cleridou et al., 2017, Does parent-child agreement vary based on presenting problems? Results from a UK clinical sample. Child and Adolescent Psychiatry and Mental health 19:11:22). In another study on informant agreement in a clinical sample using SDQ, there was no age effect on parent-child agreement ratings (van der Meer et al., 2008, Parent and child agreement on reports of problem behavior obtained from a screening questionnaire, the SDQ. Eur Child Adolesc Psychiatry 17:491.497).
At age 15 in the present study, only the SDQ was used to measure (internalizing) symptoms of depression and anxiety. The aim of the study was to investigate any secular changes between cohorts, and therefore further validation of the scales used, were outside the scope of this paper.

3) Key references from the broader literature are not mentioned in the present report and the summary of previous research is weak. A good starting point is Collishaw's 2015 review and the relevant studies cited in this review (Annual Research Review: Secular trends in child and adolescent mental health, Journal of Child Psychology and Psychiatry 56:3 (2015), pp 370-393). Response: We have now included key points from this review to “set the stage” for the current manuscript, for example referring to (i) epidemiological studies indicating that mental health problems in youth are common and tend to persist into adulthood, (ii) studies on time trends in mental health symptoms among youth as one important aspect in the prevention of mental illness in this population, and (iii) cross-cohort comparisons are an important source to better understand patterns beyond diagnostic changes.

4) The statistical analyses used are not appropriate for investigating possible gender differences in trends (assuming these really are part of the study’s aims). Methods are required for testing possible interaction terms (cohort x sex) rather than simply splitting the sample and applying separate tests to boys and girls. Response: The aim of the study was not to explore sex differences and the separate statistical tests with boys and girls were computed as sensitivity analyses. This has been clarified in the manuscript, see the statistics section. Through the sensitivity analyses, it was possible to explore the trends in internalizing symptoms in the total sample and among boys and girls separately, which, adds value to the study. We are aware of the possibility to include an interaction term (cohort X sex) but given that it was not our main purpose to explore sex differences, we have chosen to keep our separate analyses.

5) There are a number of reported findings that are difficult to reconcile and can even seem contradictory. Why is the gender difference for the SCARED (Table 1) reversed for the sMFQ? Why is the trend for parent reports at age 9 not reflected in their reports for age 15? Might this be because of the different birth cohorts covered by the studies across the two age groups? Response: As shown by the effect sizes (Table 1), the difference between boys and girls were negligible with regard to the SCARED and sMFQ, i.e. at age 9. However, there were small and large differences between boys and girls with regard to the SDQ parent and self-report scales, i.e. at age 15. We have commented on this in the discussion section. Moreover, the trend for parent
reports at age 9 (SCARED) is actually reflected in their reports at age 15 (SDQ parent) among girls but not in the total sample and among boys. This suggests that the use of different cohorts across the age groups may not be a factor that explains our findings. We have not commented on this in the manuscript.

6) The manuscript does not give sufficient consideration as to whether the use of twin samples may not be generalizable to the total population. Given that the paper is about trends, the key issue is whether trends for twins can be generalized. The main potential threat to validity in this respect is the increased incidence in DZ twins resulting from fertility treatments. Can this distort trends over the particular cohorts used for this research? What do the authors know about changes in this population over time that are pertinent to the measured outcomes?

Response: Previous concerns about generalizability of research performed on twin samples has been addresses in various studies (see e.g. Evans, D.M. & Martin, N.G. (2000) The validity of twin studies. GeneScreen 1, 77–79. View ArticleGoogle Scholar; Barnes JC, Boutwell BB. A demonstration of the generalizability of twin-based research on antisocial behavior. Behav Genet 2013;43(2):120-131; Kendler KS, Martin NG, Heath AC, Eaves LJ. Self-report psychiatric symptoms in twins and their nontwin relatives: are twins different? Am J Med Genet 1995;60:588–91. View ArticlePubMedGoogle Scholar; Kendler KS, Pedersen NL, Farahmand BY, Persson PG. The treated incidence of psychotic and affective illness in twins compared with population expectations: a study in the Swedish twin and psychiatry registries. Psychol Med. 1996;26:1135–44.). In conclusion, little differences between twins and singletons were observed, as well as differences between DZ and MZ twins regarding psychiatric illnesses. (Kendler KS, Pedersen NL, Farahmand BY, Persson PG. The treated incidence of psychotic and affective illness in twins compared with population expectations: a study in the Swedish twin and psychiatry registries. Psychol Med. 1996;26:1135–44). Theoretically, if anything, an increasing proportion of DZ in relation to MZ would make the twin sample more similar to a singleton sample. However, in our analyses we have used a cluster robust sandwich estimator (to adjust the standard errors for the nested twin data)

It is true that the proportion of DZ twins increased in Sweden from 1980 and onwards probably mostly due to the introduction of IVF. Although this trend has reversed after successive introduction of single embryo IVF in later years (see Magnusson PK, Almqvist C, Rahman I, Ganna A, Viktorin A, Walum H, Halldner L, Lundström S, Ullén F, Långström N, Larsson H, Nyman A, Gumpert CH, Råstam M, Anckarsäter H, Cnattingius S, Johannesson M, Ingelsson E, Klareskog L, de Faire U, Pedersen NL, Lichtenstein P. The Swedish twin registry: establishment of a biobank and other recent developments. Twin Res Hum Genet. 2012 Feb;16(1):317-29.) The latter is somewhat reflected in our 9-year old sample:
Proportion DZ twins in 9-year old sample:
1998 73.5%
1999 74.1%
2000 71.3%
2001 70.5%
2002 73.6%
2003 72.1%
2004 67.8%
2005 65.5%
2006 65.7%
2007 59.8%
2008 65.6%

Proportion DZ twins in 15-year-old sample:
1994 64.5%
1995 66.3%
1996 67.1%
1997 68.8%
1998 68.6%
1999 73.1%
2000 68.9%
2001 63.0%
However, we do not think there is any reason to believe that changes in proportion of zygosity for twins from the cohorts studied would profoundly affect the measures of internalizing symptoms in our study; especially in the light of previous studies mentioned above.

In addition to these listed points, the manuscript can be made more accessible with some tidying up. The captions of the tables are sometimes incorrect (Table 2) or do not convey what the table is about (Table 3). I was perplexed as to why the reported Ns were so different for different measures (17,576 for the sMFQ and 14,979 for the SCARED) and what effect this might have on the findings reported. The paper reports response rates but not missing data rates or any analyses intended to shed light on possible biases arising from missing data.

Response: When item answers were missing (coded “does not know”, or “does not want to answer”) in the sMFQ or SCARED questionnaires from the parent interviews, cut-off and total score were not calculated. Due to the larger total number of items in the SCARED, missing answers were far more common in SCARED questionnaire than in the sMFQ. This is the explanation to the large discrepancy in Ns for the sMFQ and SCARED data.

Regarding missing data, analyses between responders and non-responders in the CATSS were made in Anckarsäter et al, Twin Res Hum Genet 2011 (which is referred to in the methods section): “Systematic analyses for differences between non-responders and responders (i.e., children for whom parents declined or consented to interviews) were performed on the basis of an anonymized merge between the CATSS database for the first 11,222 participants and official files such as the National Board of Health and Welfare database on socio-economic circumstances, in- and outpatient diagnostics, and pharmacological treatment. Non-responders to the CATSS 9/12 telephone interviews were more likely than responders to have: a parent treated in psychiatric settings (9.6% of the non-responders vs. 6.3% of the responders), a father convicted of a felony (11.2% vs. 7.2%), a mother convicted of a felony (1.6% vs. 0.7%), a divorced mother (16.4% vs. 12.5%), a divorced father (16.4% vs. 12.4%), or to belong to low socio-economic strata (26.6% vs. 21.9%). Non-responders to the telephone interviews also had 2.1% ADHD as compared to 1.6% among responders, 0.95% ASD versus 0.84%, 2.0% LDs versus 0.99%. Among non-responders, 1.8% had been prescribed psychopharmacological treatment for ADHD as compared to 1.4% of the responders.” Thus, we believe that the risk of underestimating internalizing symptoms in the cohorts studied is larger than the risk of overestimating internalizing symptoms. Further, this risk would also be valid for the time trend, in the light of decreasing rates of participation, which would if anything counteract any measures of increasing internalizing symptoms.

This is discussed in the Strengths and Limitations section.