Investigating sex bias in the AQ-10

Short report: Investigating sex bias in the AQ-10: A replication study

Aja Louise Murray

University of Cambridge, UK

Tom Booth

University of Edinburgh, UK

Bonnie Auyeung

University of Edinburgh, UK

Karen McKenzie

Northumbria University, UK

Renate Kuenssberg

NHS Fife, UK

Investigating sex bias in the AQ-10: A replication study

Abstract

There are concerns that females with autism spectrum disorders (ASD) may be under-identified because of factors such as better camouflaging and poorer recognition of the signs of ASD in females. One stage at which females may be under-identified is during screening. In this study we, therefore, evaluated whether the AQ-10, a brief recommended screening instrument for ASD in adults suspected of having ASD, showed any evidence of under-estimating symptoms in females. Our results broadly replicate those of an earlier study in finding no strong evidence that the AQ-10 is biased against females. However, to achieve better performance in females we suggest that one item be replaced with an item measuring more 'female' manifestations of ASD.

Keywords: Autism screening; AQ-10; sex bias; female autism

Autism spectrum disorder (ASD) is defined by clinically significant impairments in social communication and interaction and restrictive repetitive activities; however, the presentation of symptoms varies considerably from person to person (American Psychiatric Association, 2013). Timely recognition of impairment is important for obtaining a diagnosis and access to relevant support and services as early as possible. It is, however, thought that there are a large number of individuals who would qualify for a clinical diagnosis but who have not presented at clinical services (e.g. Baron-Cohen et al., 2009). Screening for ASD can help those who may not otherwise be identified to come to the attention of relevant clinical services. At the same time, it can help avoid referral of individuals for full diagnostic assessment who are unlikely to ultimately receive a diagnosis.

It is, however, important to ensure that screening practices do not disadvantage females with ASD. In several areas, concerns have been voiced about the under-identification of females with ASD. Females may, for example, show better camouflaging of symptoms, be more susceptible to diagnostic overshadowing, or may be harder to identify simply because ASD in females is less well understood and more stereotypically associated with males (e.g. Krieser & White, 2014; Lai, Lombardo, Auyeung, Chakrabarti & Baron-Cohen, 2015; Lai et al., 2016; Russell, Steer, & Golding, 2011). As a result, females seem to need to show more severe problems in obtaining a diagnosis and are generally diagnosed at an older age than males (e.g. Beeger et al., 2013; Russell et al., 2011; Rutherford et al., 2016).

In relation to screening instruments, there is a concern that they may be less able to detect ASD in females. The concern derives from the fact that because assessments for ASD have historically been based on a 'male-typical' view of ASD and have generally been validated in predominantly male samples, they may not be well calibrated to detect ASD in females. Only a small number of studies have attempted to address these concerns. Kopp and Gillberg (2011) presented the Autism Spectrum Screening Questionnaire-Revised Extended

Version which was designed to include more items that were more sensitive to ASD in girls. They found several items that were more likely to be endorsed in the case of girls, including: interacting mostly with younger children, avoiding demands, having a different voice/speech, and having difficulty completing simple daily activities. Boys, however, were more likely to lack best friends. Although not representing a direct test for bias, these results imply that males and females may have different patterns of responses to ASD assessments. Given the focus on male-typical manifestations of ASD in most assessment this suggests a potential systematic bias against females in terms of the detection of ASD-related impairment.

Murray et al. (2017) conducted a direct test of possible bias against females in the AQ-10. The AQ-10 is an abbreviated version of the autism spectrum quotient (AQ; Baron-Cohen et al., 2001) and is recommended by the National Institute for Health and Care Excellence (NICE) as a screen for ASD for use by frontline professionals in cases of suspected ASD (NICE, 2014). They tested differential item functioning (DIF) and differential test functioning (DTF) by sex. DIF by sex is when the expected score on an item differs for males and females with the same underlying level of the trait being measured. DTF is when test scores differ for males and females of the same trait level. DIF suggests that particular items are biased; however, it is possible for DIF to be present without an overall bias in the test (DTF) if the DIF goes in both directions and cancels out at the level of the overall test.

In a combined sample of individuals with and without a clinical diagnosis of ASD (N=1237 with a clinical diagnosis, N=7356 controls) the study found that individual items of the AQ-10 showed DIF. Some were biased in favour of males and some were found to be biased in favour of females. These biases in individual items cancelled out at the level of total test scores, meaning that in spite of DIF, no DTF was in evidence. The lack of DTF applied not only at the cut-off point used to determine whether an individual should be referred for full diagnostic assessment, but across the entire range of test scores. This

suggested that individual items could not be relied upon to give comparable estimates of symptom levels across males and females. When summed, however, the overall test scores were not biased against females. As such, the study supported the NICE (2014) recommendation to use the AQ-10 as a brief screen for ASD.

Given the widespread impact of the recommendation for use of the AQ-10 by frontline health professionals, it is vitally important to ensure the generalisability of the results of this earlier study. Assessing whether the same items consistently show bias across different samples also helps to identify items that are candidates for revision. Moreover, these items may reveal differences in how ASD manifests differently across males and females. This kind of knowledge can contribute to future test design as well as contribute to a better understanding of male versus female ASD phenotypes. Here we replicate this previous study in an independent sample. to further assess whether there may be sex bias in this measure.

Method

Sample

Participants were a combined sample of individuals with a clinical diagnosis of ASD and individuals from the community with no clinical diagnosis of ASD. We took this approach to ensure a broad range of ASD trait levels. We analysed both samples together because in practice the AQ-10 is administered in contexts where diagnostic status is not yet known.

The clinically ascertained sub-sample included 107 males and 41 females with a mean age of 33.34 (SD = 10.70). Participants were recruited from a specialist regional ASD consultancy (n=140) service and clinical psychology services (n=13) in Scotland. Services identified clients who had received a clinical diagnosis of ASD and data was retrieved for these individuals. All ASD diagnoses were based on DSM-IV-TR and made by experienced

clinicians. Diagnoses were based on clinical interviews, informant interviews (where available) and individual assessments such as neuropsychological testing, where such assessments were indicated. Each case was discussed at a multidisciplinary clinic before a final diagnosis was made. Those included in the current study received a diagnosis of high functioning autism (HFA) or Asperger's Syndrome (AS). AS was defined as meeting the criteria for HFA but with no history of language delay. HFA was defined as meeting the criteria for autism with normal intellectual functioning. Those included in the current study were those with at least some data on the AQ available.

In all non-clinical subsamples, participants were recruited from the University community, and via social media and voluntary research participation websites. The first non-clinical subsample (n = 165, 21% male) was recruited for a study of sex differences in autistic-like traits (unpublished). This sub-sample had a mean age of 28.07 (SD = 12.18). Of the 164 who supplied occupational status data, 44 reported their occupational status as employed or self-employed, 17 as retired or not in work, 93 as student and 10 as 'other'. As the data collection did not rule out completion of the questionnaire by those with ASD, ten of this subsample self-reported having a clinical diagnosis of ASD. The second non-clinical subsample (n = 164; 24% male) was recruited for a psychometric study of the AQ (Murray, Booth, McKenzie, Kuenssberg & O'Donnell, 2014). This subsample has a mean age of 29.37 (SD = 10.96). Fifty-nine of this sub-sample reported their occupational status as employed, 14 as unemployed, 84 as student, and 4 as school pupil. The third non-clinical sub-sample (n = 238, 28% male) was recruited for a study on emotion recognition (McKenzie et al., 2018). This subsample has a mean age of 29.8 (SD = 13.17). Among those who supplied occupational status data, 87 reported their occupational status as employed, 33 as unemployed, 54 as student, and 13 as retired. Combinations of the above-described datasets

have been also been used in several previous publications e.g. Murray, McKenzie, Kuenssberg and Booth (2015); Booth et al., (2013).

Measures

AQ-10

The AQ-10 is a brief 10-item measure recommended as a screen for ASD in adults where ASD is suspected (Allison, Auyeung & Baron-Cohen, 2012). It is dichotomously scored, with total scores above 6 indicating referral for full diagnostic assessment. The AQ-10 is derived from the 50-item Autism Spectrum- Quotient (AQ; Baron-Cohen et al., 2001). Items were selected in order to ensure that all five domains of the AQ were represented: Attention to Detail, Attention Switching, Communication, Imagination, and Social. The items that showed the best discrimination between cases and controls within each of these domains were selected to form the AQ-10.

For a screening instrument, the most important property is whether it can correctly classify individuals as meeting diagnostic criteria or not. Only a few previous studies have examined the sensitivity and specificity of the AQ-10 at its cut-point compared against the gold standard of independent clinical diagnosis. In the original study by Allison et al. (2012), sensitivity and specificity were .88 and.91 respectively and the positive predictive value was .85. Booth et al. (2013) broadly replicated these results, finding a sensitivity and specificity of .88 and .87. Ashwood et al. (2016), however, reported a sensitivity of .77 but a specificity of only .29. These divergent results likely reflect the fact that the former used a sample of clinically diagnosed and community-sampled individuals while the latter used a sample of individuals referred for assessment for ASD. No study has yet evaluated the sensitivity and specificity of the AQ-10 as it is recommended for use in practice, namely as a screen in cases where ASD is suspected prior to referral.

Statistical Procedure

Preliminary analyses.

We began by assessing the assumption of unidimensionality using parallel analysis with principal components analysis (PA-PCA), the minimum average partial (MAP) test and visual inspection of a scree plot. We also assessed the fit of a single factor CFA model estimated using weighted least squares means and variances (WLSMV) estimation in *Mplus 7.13* (Muthén & Muthén, 2014).

DIF and DTF.

We assessed DIF and DTF using a multi-group 2 parameter logistic model. To provide a common scale and facilitate the comparison of parameters across groups, the slope and difficulty parameters of two anchor items were constrained to equality across the groups. Murray et al. (2017) used items 5 and 20 as anchors as these were identified as non-DIF based on a process of iteratively removing DIF items and retesting DIF until suitable anchors could be found. For comparability with this study, we used these same items as anchors.

Using this multi-group model, we tested for DIF in the remaining items by comparison of the fit of nested models with and without the slope and difficulty parameters of items fixed equal across groups. A significant chi-square difference test was used to determine statistically significant DIF and a BIC difference >|10| was used to determine practically significant DIF (Raftery, 1995).

Using this same model, we tested DTF, estimated by summing the item response functions across all the items for each group to obtain a test response functions. Signed DTF (sDTF) was computed using the method described in Chalmers, Counsell and Flora (2006). This is a measure of the directional bias of a test score and can range from -10 (completely biased in favour of females) to 10 (completely biased in favour of males). We also assessed

sDTF specifically at the female latent trait value corresponding to the score of 6 to assess whether the test was biased at its cut-point. Statistical significance of sDTF overall and at its cut-point of 6 was assessed using the imputation-based method described in Chalmers, et al. (2016).

Results

Descriptive statistics

Item endorsement proportions for males and females are provided in Table 1. Males endorsed all items at higher rates than females. Differences were statistically significant in all cases except item 20, which refers to understanding fictional character intentions.

Preliminary tests

The first:second eigenvalue ratio was 4.42. Parallel analysis suggested two factors to retain, MAP suggested one and visual inspection of a scree plot suggested one. A 1-factor CFA model fit well (RMSEA=.06; TLI=.93; CFI=.94; WRMR=1.83). On the basis of these results, we judged it reasonable to assume unidimensionality.

DIF and DTF analysis

Model parameters for the male and female groups from fitting a multi-group 2PL are provided in Table 2. Three items: AQ32, AQ41 and AQ45 showed statistically significant DIF (p<.05) but the DIF for only one item (AQ41) remained significant after Bonferonni correction for multiple comparisons and no items showed DIF according to the ΔBIC criterion. Test response functions are plotted for males and females in Figure 1. These indicate some evidence for a bias favouring males i.e. around the middle range of latent trait levels, males would be expected to score slightly higher than females of the same trait level. However, sDTF was not statistically significant (sDTF = 0.26, p =.064). sDTF at the male

latent trait value corresponding to the cut-point of 6 (= 0.02), was 1.06 but not statistically significant (p =.052). This suggests that at the trait values at which males are obtaining scores that would indicate referral for full diagnostic assessments, males are tending to endorse on average one additional item. We checked whether a cut-off of 5 for females better corresponded to latent trait scores associated with a score of 6 for males. At a cut-off of 6, female latent trait values were 0.31 (versus 0.02 for males), while at a cut-off of 5, female latent trait levels were -0.306. Thus, using a cut-off of 5 for females and 6 for males would tend to bias the test in favour of females by about as much as using a cut-off of 6 for both sexes would bias it in favour of males.

Discussion

In response to concerns that females with ASD are under-identified, we evaluated whether the AQ-10 screen for ASD is biased against females. Our study is a replication of a recent study by Murray et al. (2017). They found no substantial bias in test scores overall and at the recommended cut-point of 6. In this study, our results broadly replicated this result: we found no evidence for statistically significant bias at the test score level overall, or at the trait level corresponding to a test score of 6 in females. Taken together with the results of Murray et al. (2017) the balance of evidence currently suggests that the AQ-10 is not biased against females with respect to screening for ASD.

Given the importance of ensuring fairness; however, it is worth noting that, though not statistically significant, the magnitude of bias at the cut-point was approximately equivalent to males of this trait level endorsing on average an additional item compared with females. Given that the total test score ranges only from 0 to 10, this could have an important effect on under-referral in practice. One option would be to lower the cut-off point for females to 5. Our analyses indicated that this would result in a relative over-referral of

females compared with males and would, of course, result in more referrals overall. A better option may be to adapt the AQ-10 to include more female-relevant items, as Kopp and Gillberg (2011) did for the Autism Spectrum Screening Questionnaire-Revised Extended Version. The goal would be to develop a female AQ-10 with good sensitivity and specificity, as well as an equivalence of latent trait scores with a male AQ-10 version around at the optimal cut-point. Here, item 41, which showed a lack of sex invariance and a poor discrimination parameter in females would represent a good candidate for substitution with an item reflecting more 'female' manifestations of ASD. Item 41 (collecting information in categories of things) also showed significant DIF in the analysis by Murray et al. (2017). As such, future research should aim to develop and validate a replacement for item 41. In addition, though apparently invariant across males and females, item 5 (noticing small sounds) showed poor discrimination in both this study and the study by Murray et al. (2017). Replacing this item may improve the performance of the AQ-10 overall.

Of course, screening is only one area in which bias against females may occur; bias could also occur if teachers and parents are less attuned to female symptoms, if diagnostic instruments are less sensitive to female symptoms, if females are better at concealing their symptoms, or if females are more likely to be misdiagnosed with related issues (e.g. Krieser & White, 2014; Lai et al., 2016). Better understanding and awareness of the signs of ASD in females can help to reduce bias at all stages along the pathway to diagnosis.

Finally, the current study utilised DSM-IV based diagnoses and it is not yet clear how the AQ-10 performs when measured against the criterion of DSM 5 diagnosis. DSM 5 criteria differ from DSM-IV in important ways. It, for example, no longer includes Asperger Syndrome as a diagnostic category and combines what was previously a triad of impairments (social interaction, communication and restricted behavioural repertoire; APA, 1994) into a dyad (social communication and restricted repetitive activities). More importantly, and as

mentioned above, it is not yet clear whether diagnostic criteria themselves may disadvantage females with ASD.

Conclusion

In this study, our results broadly replicated that of a previous study in suggesting there is not significant male bias in the AQ-10 screen for ASD. Nonetheless, our results hinted at ways in which the AQ-10 could be made more suitable for females. In particular, we would recommend that future research explores replacing item 41 with an item that captures more 'female' manifestations of ASD.

Compliance with Ethical Standards

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*, 202-212.

APA 1994 American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (DSM-IV). *American Psychiatric Association*.

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

Ashwood, K. L., Gillan, N., Horder, J., Hayward, H., Woodhouse, E., McEwen, F. S., ... & Cadman, T. (2016). Predicting the diagnosis of autism in adults using the Autism-Spectrum Quotient (AQ) questionnaire. *Psychological Medicine*, *46*, 2595-2604.

Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., Matthews, F. E., & Brayne, C. (2009). Prevalence of autism-spectrum conditions: UK school-based population study. *The British Journal of Psychiatry*, *194*, 500-509.

Begeer, S., Mandell, D., Wijnker-Holmes, B., Venderbosch, S., Rem, D., Stekelenburg, F., & Koot, H. M. (2013). Sex differences in the timing of identification among children and adults with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *43*, 1151-1156.

Booth, T., Murray, A. L., McKenzie, K., Kuenssberg, R., O'Donnell, M., & Burnett, H. (2013). Brief report: An evaluation of the AQ-10 as a brief screening instrument for ASD in adults. *Journal of Autism and Developmental Disorders*, *43*, 2997-3000.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It Might Not Make a Big DIF: Improved Differential Test Functioning Statistics That Account for Sampling Variability. *Educational and Psychological Measurement*, *76*, 114-140.

Kopp, S., & Gillberg, C. (2011). The Autism Spectrum Screening Questionnaire (ASSQ)-Revised Extended Version (ASSQ-REV): an instrument for better capturing the autism phenotype in girls? A preliminary study involving 191 clinical cases and community controls. *Research in Developmental Disabilities*, *32*, 2875-2888.

Kreiser, N. L., & White, S. W. (2014). ASD in females: are we overstating the gender difference in diagnosis?. *Clinical Child and Family Psychology Review*, *17*, 67-84.

Lai, M. C., Lombardo, M. V., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/gender differences and autism: setting the scene for future research. *Journal of the American Academy of Child & Adolescent Psychiatry*, *54*, 11-24.

Lai, M. C., Lombardo, M. V., Ruigrok, A. N., Chakrabarti, B., Auyeung, B., Szatmari, P., ... & MRC AIMS Consortium. (2016). Quantifying and exploring camouflaging in men and women with autism. *Autism*, 1362361316671012.

McKenzie, K., Murray, A. L., Wilkinson, A., Murray, G. C., Metcalfe, D., O'Donnell, M., & McCarty, K. (2018). The relations between processing style, autistic-like traits, and emotion recognition in individuals with and without Autism Spectrum Disorder. *Personality and Individual Differences*, *120*, 1-6.

Murray, A. L., Booth, T., McKenzie, K., Kuenssberg, R., & O'Donnell, M. (2014). Are autistic traits measured equivalently in individuals with and without an autism spectrum disorder? An invariance analysis of the Autism Spectrum Quotient Short Form. *Journal of autism and developmental disorders*, *44*, 55-64.

Murray, A. L., McKenzie, K., Kuenssberg, R., & Booth, T. (2015). Do the Autism Spectrum Quotient (AQ) and Autism Spectrum Quotient Short Form (AQ-S) primarily reflect general ASD traits or specific ASD traits? A bi-factor analysis. *Assessment*. Epub ahead of print.

Murray, A. L., Allison, C., Smith, P. L., Baron-Cohen, S., Booth, T., Auyeung, B. (2017). Investigating diagnostic bias in autism spectrum conditions: An item response theory analysis of sex bias in the AQ-10. *Autism Research*, *10,* 790-800.

NICE (2014). Autism: recognition, referral, diagnosis and management of adults on the autism spectrum. (Clinical guideline 142.) 2014. https://www.nice.org.uk/guidance/qs51

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111-163.

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*, 1283-1293.

Rutherford, M., McKenzie, K., Johnson, T., Catchpole, C., O'Hare, A., McClure, I., …& Murray, A.L. (2016). Gender ratio in a clinical population sample, age of diagnosis and duration of assessment in children and adults with Autism Spectrum Disorder. *Autism*. Early view.

Tables

Table 1: AQ-10 item proportion endorsement across males and females

Item no.	Content	P Males	P Females	χ^2	p
28	I usually concentrate more on the whole picture, rather than the small details. R	.57	.47	7.23	.007
5	I often notice small sounds when others do not.	.70	.56	11.46	<.001
32	I find it easy to do more than one thing at once. R	.59	.40	23.23	<.001
37	If there is an interruption, I can switch back to what I was doing very quickly. R	.54	.44	5.62	.017
27	I find it easy to 'read between the lines' when someone is talking to me. R	.56	.42	13.10	<.001
31	I know how to tell if someone listening to me is getting bored. R	.48	.39	4.83	.028
20	When I'm reading a story I find it difficult to work out the characters' intentions.	.54	.48	1.79	.182
41	I like to collect information about categories of things (e.g., types of car, types of bird, types of train, types of plant, etc.).	.60	.38	31.77	<.001
36	I find it easy to work out what someone is thinking or feeling just by looking at their face. R	.59	.40	22.05	<.001
45	I find it difficult to work out people's intentions.	.63	.44	24.28	<.001

Note. Item numbers refer the position of the item numbers from the full 50 item AQ. P= proportion endorsement where R indicates that an item has been reverse coded so that for all items endorsement means having higher levels of ASD traits.

Table 2:

Male versus female 2PL parameters and DIF analyses

Item	Male a	Male b	Female a	Female b	χ^2	*p*	ΔBIC
5	0.041	0.454	0.041	0.454	-	-	-
28	0.717	0.038	0.865	0.348	2.364	.31	-10.792
32	1.432	-0.215	1.027	0.458	7.435	.02*	-5.7212
37	1.037	-0.012	1.279	0.191	1.037	.60	-12.119
27	2.532	-0.045	2.24	0.39	1.331	.51	-11.825
31	2.147	-0.264	1.203	-0.1	3.557	.17	-9.599
20	1.146	0.21	1.146	0.21	-	-	-
41	1.031	-0.357	0.909	0.496	12.648	.002*	-0.507
36	2.902	-0.191	2.029	0.567	4.484	.11	-8.672
45	1.422	0.034	1.959	0.866	6.239	.044*	-6.916

Note. a=discrimination parameter; b= difficulty parameter. Items 5 and 20 were used as anchors and parameters fixed equal across groups.

*statistically significant at $p<.05$.

Figure 1:

Test characteristic curves for males (=0) and females (=1)