Original Article

autism

Autism
2019, Vol. 23(7) 1752–1764

DOI: 10.1177/1362361319828361
journals.sagepub.com/home/aut
SAGE

Negatively phrased items of the Autism Spectrum Quotient function differently for groups with and without autism

Joost A Agelink van Rentergem[1,2], Anne Geeke Lever[1,2] and Hilde M Geurts[1,2,3]

Abstract

The Autism Spectrum Quotient is a widely used instrument for the detection of autistic traits. However, the validity of comparisons of Autism Spectrum Quotient scores between groups may be threatened by differential item functioning. Differential item functioning entails a bias in items, where participants with equal values of the latent trait give different answers because of their group membership. In this article, items of the Autism Spectrum Quotient were studied for differential item functioning between different groups within a single sample ($N = 408$). Three analyses were conducted. First, using a Rasch mixture model, two latent groups were detected that show differential item functioning. Second, using a Rasch regression tree model, four groups were found that show differential item functioning: men without autism, women without autism, people 50 years and younger with autism, and people older than 50 years with autism. Third, using traditional methods, differential item functioning was detected between groups with and without autism. Therefore, group comparisons with the Autism Spectrum Quotient are at risk of being affected by bias. Eight items emerged that consistently show differences in response tendencies between groups across analyses, and these items were generally negatively phrased. Two often-used short forms of the Autism Spectrum Quotient, the AQ-28 and AQ-10, may be more suitable for group comparisons.

Keywords

adults, age differences, autism spectrum disorders, Autism Spectrum Quotient, differential item functioning, measurement, measurement invariance, sex differences

The Autism Spectrum Quotient (AQ; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) is a widely used instrument for the detection of autistic traits. In a systematic review and meta-analysis, Ruzich et al. (2015) gathered evidence from 73 studies that used the AQ that tested participants with and without autism.[1] They concluded that people with autism score higher on autistic traits than people without autism. Also, men without autism score higher on autistic traits than women without autism, while no sex difference was found for people with autism. However, the validity of comparisons of AQ scores between groups may be threatened by differential item functioning (DIF).

DIF denotes that there are group differences in responses to a particular item. DIF analyses typically assume that there is a single psychological trait that underlies differences between people (Mazor, Hambleton, & Clauser, 1998). If two people are identical on this psychological trait and they answer items on a questionnaire measuring that trait, they should provide the same answers. However, if there is a particular item for which the person who happens to be male has a higher chance to respond "agree" than the person who happens to be female, there is something wrong. Then, group membership also determines the response, not just the latent trait. Formally, DIF means that if we consider

[1]University of Amsterdam, The Netherlands
[2]Dutch Autism & ADHD Research Center, The Netherlands
[3]Dr. Leo Kannerhuis, The Netherlands

Corresponding author:
Joost A Agelink van Rentergem, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129 B, 1018 WS Amsterdam, The Netherlands.
Email: j.a.agelink@uva.nl

participants with equal values on the latent trait, group membership still affects the response to a particular item (Mellenbergh, 1989).

As an example, consider item 10 of the AQ: "In a social group, I can easily keep track of several different people's conversations." This item measures autistic traits and those with higher autistic traits will generally answer "disagree," while those with low autistic traits will generally answer "agree." Two people with identical low values on autistic traits should tend to both answer "agree." However, if one of the two is from an elderly group, he may instead tend to answer "disagree," because of diminished hearing ability. Responses would then be dependent not just on autistic traits, but also on age group membership. If we would use the AQ to compare elderly and young groups, we could due to this item erroneously conclude that the elderly exhibit more autistic traits. Similarly, consider item 13: "I would rather go to a library than a party." Those with identical low autistic traits should generally respond the same: "disagree." However, those with low autistic traits and high education may actually answer "agree," because not just autistic traits but also education could determine whether a person likes libraries. This would bias the results in the direction of concluding that those with high education have more autistic traits. Therefore, when group membership affects the responses directly, that is, DIF, this may distort results of group comparisons.

Note that DIF does not refer to group differences in responses: There will be differences between groups due to differences in how groups score on the latent trait. Autistic and non-autistic people are likely to answer items 10 and 13 differently, due to differences in autistic traits between these groups. Such group differences in autistic traits are not what we mean by DIF, but to be able to study group differences in autistic traits, we need to study DIF first.

So far, five[2] studies on DIF in the AQ have been conducted, but these studies were limited to autism diagnosis and sex, and are difficult to compare because of differences in the materials and methods. Two studies examined DIF between autism and non-autism groups (Lundqvist & Lindner, 2017; Murray, Booth, McKenzie, Kuenssberg, & O'Donnell, 2014). Three studies examined DIF between men and women (Grove, Hoekstra, Wierda, & Begeer, 2017; Murray, Allison, et al., 2017; Murray, Booth, Auyeung, McKenzie, & Kuenssberg, 2017). In some of these studies, the groups are confounded with age and sex, with one group being older than the other, or containing more men. To our knowledge, no studies investigated the influence of both sex and autism on DIF in a single sample, or at DIF between age groups or levels of education. Therefore, we do not know whether there are threats to the validity of comparisons of young and old age groups, or comparisons of different levels of education. Also, the validity of comparisons between men and women without autism is understudied. In this article, DIF is examined for autism and non-autism groups, men and women, young and old age groups, and different levels of education, within a single sample. This allows us to relate the different findings, and look at interactions between these factors.

Second, it is difficult to compare studies as some studies consider the full AQ which has 50 items (used by Lundqvist & Lindner, 2017), while others consider shortened versions, like the AQ-28 (Hoekstra et al., 2011; used by Grove et al., 2017; Murray et al., 2014), or the AQ-10 (Allison, Auyeung, & Baron-Cohen, 2012; used by Murray, Allison, et al., 2017; Murray, Booth, et al., 2017). Because they use shortened versions, these studies do not examine DIF given equal levels of autistic traits as measured by the full AQ, and we do not have information on DIF between men and women for all 50 items of the AQ. In this article, we examine DIF between men and women on the full AQ, in addition to other comparisons between groups.

Third, studies are difficult to compare because they used different scoring and analysis methods. The AQ is administered with four response options: "definitely disagree," "slightly disagree," "slightly agree," and "definitely agree." The AQ was designed to be scored dichotomously, with "definitely disagree" and "slightly disagree" taken together, and "slightly agree" and "definitely agree" taken together. Three DIF studies use this coding scheme (Lundqvist & Lindner, 2017; Murray, Allison, et al., 2017; Murray, Booth, et al., 2017). Two studies consider four possible scores for every item (Grove et al., 2017; Murray et al., 2014), because there is more information in a four-point scale than in a dichotomous scale (Austin, 2005; Hoekstra, Bartels, Cath, & Boomsma, 2008; Murray, Booth, McKenzie, & Kuenssberg, 2016; Stevenson & Hart, 2017). These two studies used different analysis methods because of the difference in measurement scale. Therefore, it is difficult to lay the results of the studies side-by-side, and we do not know the impact of the results of the four-point scale studies for cases where the dichotomous scale is used. Because the dichotomous scale is how the instrument is used in clinical practice, we examine DIF with this scoring rule.

DIF analyses typically assume unidimensionality, that is, there is a single construct "autistic traits" that is measured by the AQ. This assumption is implicitly made in all articles that use the sum score on the AQ in their analyses. Multidimensionality could lead to inaccurate findings of DIF in an analysis that assumes unidimensionality, because legitimate group differences on an unmeasured extra dimension can cause differences in responses between groups (Mazor et al., 1998). However, research has shown that there may be more than one dimension underlying the AQ scores. In fact, the original AQ paper stated that the questionnaire was designed to consist of five domains (Baron-Cohen et al., 2001). For the full AQ,

a confirmatory factor analysis approach showed that a hierarchical structure fits best, with two main dimensions, social interaction, and attention to detail, of which the former is subdivided into four further dimensions (Hoekstra et al., 2008). Exploratory factor analyses showed that three to five dimensions may be necessary to accurately describe the structure of the AQ (Austin, 2005; Freeth, Sheppard, Ramachandran, & Milne, 2013; Kloosterman, Keefer, Kelley, Summerfeldt, & Parker, 2011; Lau et al., 2013, Lau, Kelly, & Peterson, 2013; Russell-Smith, Maybery, & Bayliss, 2011; Stewart & Austin, 2009). Because of their exploratory nature, these models may be overfitting the data, and not all items are included in the factor solutions. However, because these methods are data-driven, they are less affected by prior assumptions. Therefore, both confirmatory and exploratory results are useful.

In this article, we study the differences in item functioning between different groups, with the full version of the AQ and dichotomous scoring. Two of the factors that we include in our analyses have not been studied yet: age and level of education. Also, to make comparisons between groups, we use recently developed statistical techniques for detecting DIF that are less restricted by the way groups are defined than previously used methods. Finally, we study whether the assumption of unidimensionality is tenable. This allows us to answer the question of whether group comparisons on the full AQ are valid, or may be biased by DIF.

Methods

Materials

The Dutch translation of the AQ was used for this study (Hoekstra et al., 2008). All 50 items were used in the analysis, because we investigated DIF in all 50 items, given autistic traits as measured by the entire AQ. We used dichotomous scoring, meaning that one can obtain a score of either 0 or 1 on each item, because this is the original and most common way of scoring (Baron-Cohen et al., 2001).

Participants

We made use of data from 435 participants that were collected in an earlier conducted project (Lever & Geurts, 2018). Autistic participants were recruited through mental health institutions and advertisements on client organization websites. They were required to have a clinical diagnosis on the spectrum according to the Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM-IV) criteria, which was generally established by a multidisciplinary team including a psychiatrist and/or psychologist. Non-autistic participants were recruited through advertisements on the university website, social media, and within the social environment of the original authors. They were required to not have an autism, attention-deficit/hyperactivity disorder or schizophrenia diagnosis, or close relatives with an autism or schizophrenia diagnosis.

Table 1. Table of cell counts, for autism diagnosis and sex.

	Men	Women	Total
Autism	144	70	214
Non-autism	108	86	194
Total	252	156	408

Four participants were missing a value on the education variable (three autism and one non-autism). For 23 participants, one or more responses on the AQ were missing (20 autism, 3 non-autism; maximum number missing per item was three). One of the three analyses required complete data on demographic variables, and all analyses required complete response data. Data from 27 cases with missing data were removed for all analyses, and data from 408 participants were analyzed. The distribution of men and women was different between autism and non-autism groups, $\chi^2(1) = 5.3$, $p = 0.021$, with more men in the autism group. Cell counts are given in Table 1.

The mean age was 45.3 years, with a standard deviation of 15.2, range 19–79. Age was not different between men and women, $t(406) = 1.53$, $p = 0.127$, $d = -0.16$, nor between autism and non-autism groups, $t(406) = 0.13$, $p = 0.894$, $d = -0.01$.

Education was scored on an ordinal rating scale (Verhage, 1964), with 1 denoting not having completed primary school, and 7 denoting a university education. The sample was highly educated, with 102 participants scoring 7, 188 scoring 6, 100 scoring 5, and 18 participants scoring between 2 and 4. The non-autism group was more highly educated than the autism group, $\chi^2(5) = 16.2$, $p = 0.006$, with the majority of the non-autism group scoring 6 or 7, and the majority of the autism group scoring 5 or 6. Education level did not differ between the sexes, $\chi^2(5) = 5.1$, $p = 0.400$.

Analyses

Three analyses were conducted to determine whether items of the AQ show DIF, using Rasch mixture models, Rasch regression trees, and traditional methods for detecting DIF. We used the Rasch model in all three analyses, because this is the psychometric model that best fits the idea of using the sum score as a measure of autistic traits, which is the way the AQ is used in practice.

In the first analysis, Rasch mixture model were fitted using the *psychomix* R-package (Frick, Strobl, Leisch, & Zeileis, 2012). This method splits the total sample into a number of latent groups that show DIF. This method has the advantage, compared with traditional methods, that it can detect whether there are groups that show DIF, even if

these groups have not been measured or defined (Frick et al., 2012). To return to our example of item 10, suppose participants with normal and low hearing ability respond differently to the question of whether they can keep track of conversations, even given equal values on autistic traits. This would constitute DIF, as answers depend on group membership, but this will not be detected by a traditional DIF analysis comparing men and women. With the Rasch mixture model, the two hearing ability groups that do show DIF can be identified even when hearing ability is not measured and is not included as a factor in the analyses. Therefore, mixture models which do not use observed, but latent, groups can provide a more accurate test of whether a questionnaire displays DIF, by searching for the optimal way to split participants into groups.

For the second analysis, we fitted Rasch regression trees, which split the sample into groups using demographic variables. For this analysis, we used the *psychotree* R-package (Strobl, Kopf, & Zeileis, 2015). The disadvantage of Rasch regression trees in comparison to Rasch mixture models is that grouping variables have to be predefined, for example, by entering age and sex. This means that variables that are important to DIF may not be entered and DIF may go undetected. However, if the entered grouping variables do result in DIF, results are more interpretable for the groups in the Rasch regression tree compared with the latent groups in the Rasch mixture model.

There are a number of advantages to using trees in comparison to traditional methods of detecting DIF. First, for continuous variables like age, no predefined cutoff point to split groups needs to be chosen. Without regression trees, one often-chosen strategy is to use the median as a cutoff point, defining everyone below the median age as young, and everyone above the median age as old (Strobl et al., 2015). However, the median is theoretically uninformative and sample-dependent. It is better to use regression trees to find cutoff points that maximize DIF, subjecting items to the most critical test. Second, regression trees allow for interactions: If there is DIF between young men and young women, but not between old men and old women, this would be difficult to detect with traditional DIF analyses, as those require that two groups are chosen a priori. Within regression trees, such interactions are naturally accommodated.

In the third analysis, we examined DIF using methods that compare parameters of the Rasch model between two predefined groups. Many different methods of comparing two groups are available, which have been bundled in the R-package *difR* (Magis, Béland, Tuerlinckx, & De Boeck, 2010). This package was also used by Murray, Allison, et al. (2017) in their analysis of DIF between men and women. We will use these methods to compare groups with and without autism. As mentioned above, there are a number of disadvantages to the traditional methods of detecting DIF, compared with Rasch mixture and Rasch regression tree methods. However, there are advantages as well.

The first advantage is that using traditional methods makes it easier to compare our results with those from the literature. The second advantage is that if there are two predefined groups that are of theoretical interest, like groups with and without autism, traditional methods provide a more powerful test of differences between these groups. The third advantage is that we can use an algorithm for item purification that has been developed for these methods, which is not yet available for the newer methods (Magis et al., 2010). Because we were testing the significance of DIF for 50 items separately, a Bonferroni correction was performed to correct for multiple testing.

After the three analyses, we considered the content of the items, and considered whether items were included in published short forms of the AQ. This allows us to consider causes of possible DIF, and the impact of possible DIF on short forms. We also compared our item-level results with earlier DIF results. Finally, we focused on unidimensionality. As mentioned in the Introduction to this article, DIF analyses typically assume that there is a single latent trait underlying scores. We tested whether the assumption of unidimensionality is tenable, and examined DIF for different subsets of items from multidimensional models.

Results

Rasch mixture results

First, Rasch mixture models were fitted for various numbers of latent groups. The models with one, two, three, four, five, six, and eight latent groups converged, while the models with seven, nine, and ten latent groups had not converged after 5000 iterations. Of the seven models that converged, the model with two groups fits best according to the Bayesian Information Criterion (Schwarz, 1978). These two groups generally correspond to autism and non-autism groups, $\chi^2(1) = 250.6$, $p < 0.001$, but also somewhat correspond to men and women, $\chi^2(1) = 12.7$, $p < 0.001$. Cell counts are given in Table 2.

The difference between item parameters estimated in the two latent groups is displayed in Figure 1(a), designated by the triangles. The items on the *x*-axis are ordered by the amount of DIF they showed in this analysis. The 10 items that showed the biggest difference in item parameters between latent groups were 30, 29, 11, 49, 26, 22, 24, 21, 45, and 14. The content of the items is given in Table 3, also for the items that display DIF in the following analyses.

Stability. To check whether the item ordering was stable, we reran the model 100 times with different starting values. The item ordering did change somewhat between these 100 reruns of the model, but only for the 32 items

Table 2. Table of cell counts, for latent group assignment and autism diagnosis, and latent group assignment and sex.

	Autism	Non-autism	Total		Men	Women	Total
Latent group 1	194	23	217	Latent group 1	152	65	217
Latent group 2	20	171	191	Latent group 2	100	91	191
Total	214	194	408	Total	252	156	408

that showed the least DIF. The identification of the items that showed DIF was stable.

Rasch regression tree results

Second, a Rasch regression tree model was fitted, with four variables as possible predictors: autism diagnosis, age, sex, and level of education. A regression tree model was obtained that divided the sample into four groups, by autism diagnosis, age, and sex. The first split was by autism diagnosis. The autism group was then split into two groups, one 50 years and younger, and the other older than 50 years. Note that the age of 50 was not a preset value in our analysis, even though it is named as a divider between the young and old participants in the autism literature (e.g. Totsika, Felce, Kerr, & Hastings, 2010). The non-autism group was split by sex. Therefore, four groups were detected: autism 50 years and younger, autism older than 50 years, women without autism, and men without autism. The regression tree is depicted in Figure 2.

As a measure of severity of DIF for every item, we computed the largest difference between item parameters for the four groups. This difference in item parameters is displayed in Figure 1(a), designated by the circles.

The 10 items that showed the largest DIF effects between two of the four groups were items 30, 29, 49, 11, 26, 22, 35, 21, 14, and 46 (see Table 3). Item 30 is especially noticeable, as particularly women without autism have a stronger tendency to provide the response that is typically associated with autistic traits, while the groups with autism have a stronger tendency to provide the response not associated with autistic traits.

Stability. Regression tree-based methods are known to be unstable (Strobl, Malley, & Tutz, 2009), which means that small differences in the sample can lead to large differences in the conclusions that are drawn. To estimate the stability of the present results, the analysis was repeated for 1000 samples of 90% of the participants ($N=367$), taken without replacement. All 1000 trees had autism diagnosis as the first split. All 1000 trees included sex as splitting the non-autism group, and 332 trees included sex as splitting the autism group. Age split the autism group into 114 trees, and in no trees in the non-autism group. Education split the autism group into four trees. In sum, the split of the autism group by age in the main analysis was unstable, as minor changes in the data removed this split, or replaced it by a split by sex. Stable results were the split between autism and non-autism groups, the split of the non-autism group into men and women, and the lack of inclusion of education.

Traditional Rasch methods

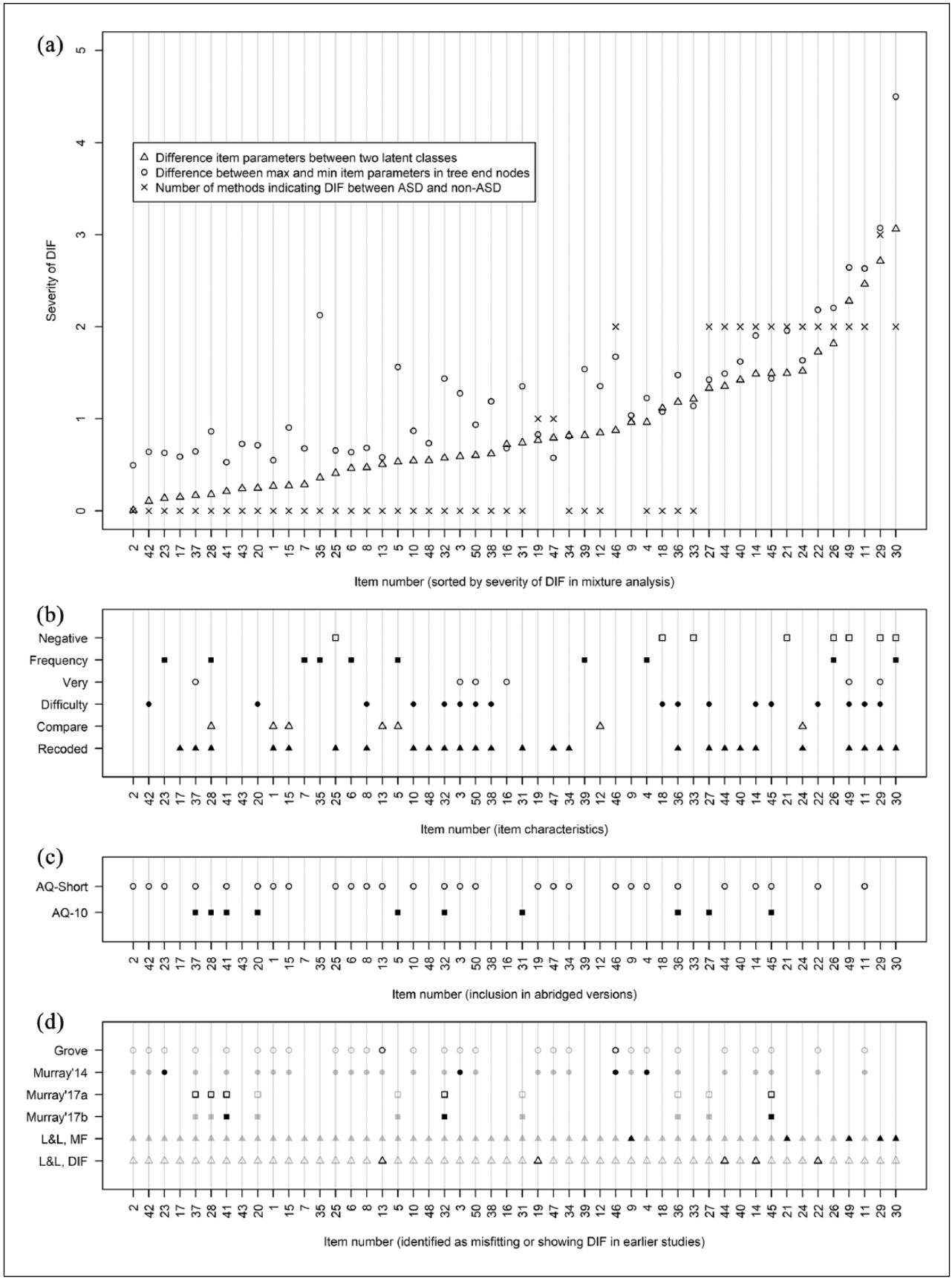
Finally, traditional methods for detecting DIF were used, with autism and non-autism groups as the predefined groups to compare. Autism diagnosis was chosen because it is theoretically relevant to know whether those with and without autism interpret items differently, and whether separate questionnaires may be necessary (McConachie et al., 2017). For three out of five traditional methods, the item purification algorithm converged. These three methods were logistic regression (Swaminathan & Rogers, 1990), Lord's (1980) chi-square test, and Raju's (1988) area method. Specifics on these methods are provided by Magis et al. (2010).

The logistic regression method showed the fewest significant results, with only four items identified as displaying DIF (with moderate effect sizes for three, and negligible effect size for one). Two of these items were not identified as displaying DIF by Lord's chi-square test and Raju's area method. Lord's chi-square test and Raju's area method both identified the same 14 items (with 14 large effect sizes, as identified by both Lord's method and Raju's method). For all 50 items, Item Characteristic Curves are plotted in the Supplemental Material (Figure M), so the direction and size of the effect can be interpreted visually. For most items that show DIF, like items 29, 30, and 49, participants without autism were more likely than participants with autism to give the autism-typical response, given equal autistic traits. For some items that show DIF, like item 11, participants with autism were more likely to give the autism-typical response, given equal autistic traits.

The number of methods that indicate significant DIF is displayed in Figure 1(a) for each item, designated by the crosses. One item was identified as showing DIF by three out of three methods, item 29. The items that were identified by two out three methods as showing DIF were 11, 14, 21, 22, 24, 26, 27, 30, 40, 44, 45, 46, and 49 (see Table 3).

Item characteristics

There are a number of items that show DIF in all three analyses, of which the top eight in Table 3 are the most


(a)
Severity of DIF
Difference item parameters between two latent classes
Difference between max and min item parameters in tree end nodes
Number of methods indicating DIF between ASD and non-ASD
Item number (sorted by severity of DIF in mixture analysis)
(b)
Negative
Frequency
Very
Difficulty
Compare
Recoded
Item number (item characteristics)
(c)
AQ-Short
AQ-10
Item number (inclusion in abridged versions)
(d)
Grove
Murray'14
Murray'17a
Murray'17b
L&L, MF
L&L, DIF
Item number (identified as misfitting or showing DIF in earlier studies)

(Continued)

Figure 1. Multi-paneled figure, with different kinds of information about the items on the *y*-axis, and the item numbers on the *x*-axis. The *x*-axis is ordered by the results of the Rasch mixture model analysis, with the smallest DIF on the left (item 2), and the largest DIF on the right (item 30). (a) Severity of DIF, where a higher DIF indicates larger differences between groups, as measured in three analyses: Rasch mixture models (triangles), Rasch regression trees (circles), and traditional methods (crosses). (b) Item characteristics, such as, whether the item was negatively phrased (Negative, in empty squares), contained a word denoting frequency (Frequency, in filled squares), contained the word "very" (in empty circles), contained a word denoting difficulty (Difficulty, in filled circles), contained a comparison (Compare, in empty triangles), and/or was a reversely coded item (Recoded, in filled triangles). (c) Whether the items were included in AQ-28 and AQ-10. (d) Whether the items were identified as misfitting, or as showing DIF, in a previous study comparing groups. The items with faint coloring were included in the versions used in these studies. Lundqvist and Lindner (2017) used all 50 items.
Grove: Grove et al., (2017); Murray'14: Murray et al. (2014); Murray'17a: Murray, Allison, et al. (2017); Murray'17b: Murray, Booth, et al. (2017); L&L: Lundqvist and Lindner (2017); MF: misfit; DIF: differential item functioning; AQ: Autism Spectrum Quotient.

Table 3. Item content for the items of the AQ that display the most differential item functioning in the different analyses, ordered by severity and the number of analyses that identified these items.

Item no.	Item content	DIF in analysis
30	I don't usually notice small changes in a situation, or a person's appearance.	RM RT TM
29	I am not very good at remembering phone numbers	RM RT TM
11	I find social situations easy.	RM RT TM
49	I am not very good at remembering people's date of birth.	RM RT TM
26	I frequently find that I don't know how to keep a conversation going.	RM RT TM
22	I find it hard to make new friends.	RM RT TM
21	I don't particularly enjoy reading fiction.	RM RT TM
14	I find making up stories easy.	RM RT TM
24	I would rather go to the theater than a museum.	RM TM
45	I find it difficult to work out people's intentions.	RM TM
46	New situations make me anxious.	RT TM
35	I am often the last to understand the point of a joke.	RT
27	I find it easy to "read between the lines" when someone is talking to me.	TM
40	When I was young, I used to enjoy playing games involving pretending with other children.	TM
44	I enjoy social occasions.	TM

DIF: differential item functioning; RM: Rasch mixture model (analysis 1); RT: Rasch regression tree (analysis 2); TM: traditional methods (analysis 3).

noticeable. There are a number of characteristics these items have in common. For example, five out of eight are negatively phrased. However, this may be true for many items, including those that do not show DIF. Therefore, we displayed item characteristics in Figure 1(b).

The first characteristic we examined was negation ("I don't," "It does not," "It isn't," "I am not"). Negation occurs almost exclusively in items that display DIF. The second was whether words denoting frequency ("usually," "frequently," "often," "all the time") were included. The third was the inclusion of the word "very." The fourth was whether words denoting difficulty ("easy," "difficult," "hard," "easily," "good at") were included. These characteristics do not seem to be correlated to DIF, as the whole range of DIF severity is covered by items with these characteristics. The fifth characteristic was whether a comparison is made ("I would rather," "more strongly … than"). Items with such comparisons showed relatively little DIF. The sixth characteristic was whether the item was reverse coded ("slightly disagree" and "definitely disagree" are coded as typical of autism). Although coding was reversed for the items that displayed the most DIF, and was not reversed for the items that displayed the least DIF, there does not seem to be a pattern.

Inclusion in short versions

Next, we considered whether the items were included in the various short forms. This is displayed in Figure 1(c), where inclusion is plotted for the AQ-28 (Hoekstra et al., 2011) and the AQ-10 (Allison et al., 2012). Items 21, 26, 29, 30, and 49, all negatively phrased, were not included in the AQ-28 or AQ-10, so the results from these shortened versions are not affected by the DIF displayed by these items. Items 11, 14, and 22 were included in the AQ-28, so these items may be problematic within this shortened version as well.

Relation to results in the literature

If we compare the results with those in the literature, there is a disparity in what items were identified as showing

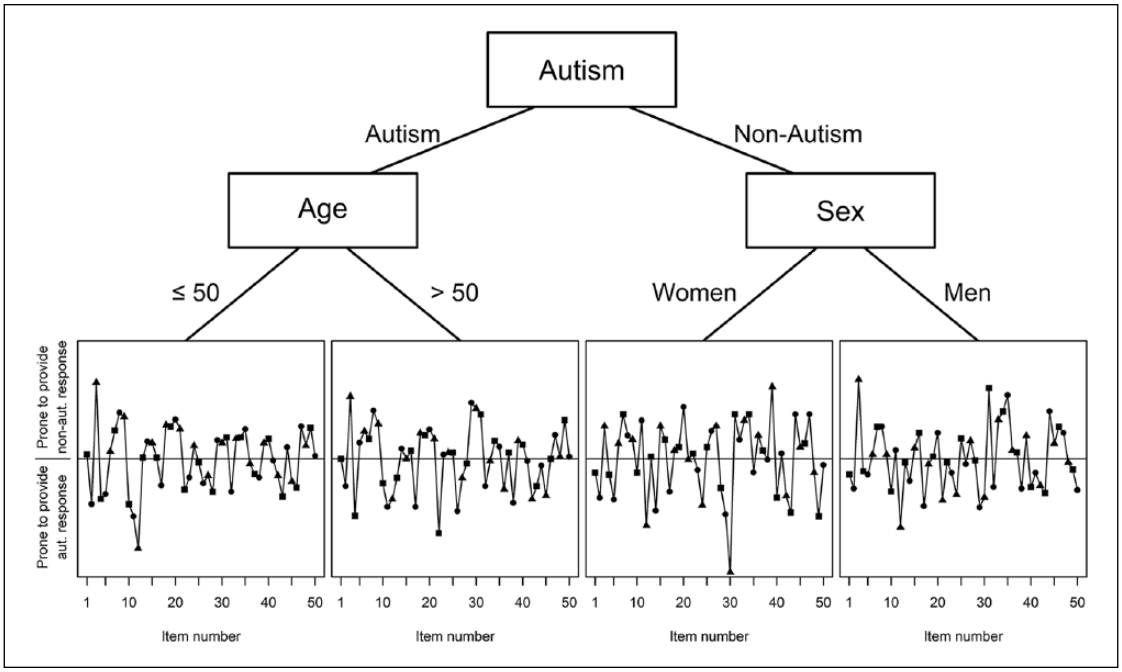


Figure 2. Rasch regression tree. In the bottom plots, the normalized item difficulties are plotted for the different groups. The higher a particular point is, the more prone a person within that subgroup is to provide a response that is not typical of autism; the lower a particular point is, the more prone a person within that subgroup is to provide a response that is typical of autism, given equal values on autistic traits. The different symbols (square, circle, and triangle) do not have separate meanings, but were chosen so the points can be easily distinguished and the position can be compared between the four different subgroups.

DIF. However, most other studies looked at the shortened versions, which do not include the items that showed the largest DIF effects in the present analysis. Which items were identified in the different studies is plotted in Figure 1(d). Grove et al. (2017), using the AQ-28, established DIF between men and women in items 13 and 46. Murray, Allison et al. (2017), using the AQ-10, established DIF between men and women in items 28, 32, 37, and 41. Murray, Booth, et al. (2017) again using the AQ-10, established DIF between men and women in items 32, 41, and 45, of which only item 41 remained significant after a Bonferroni correction. Murray et al. (2014) used the AQ-28, and fitted a model for four response options. They identified items 3, 4, 23, and 46 as showing differences between autism and non-autism groups. Across these four studies, only items 41 and 46 stand out, which were not identified in our analyses of the full AQ.

The study by Lundqvist and Lindner (2017) is notable because they also examined the full AQ. They identified items 9, 21, 29, 30, and 49 as showing misfit. The latter three correlated negatively with the latent trait, as was noted by Baron-Cohen et al. (2001) when the AQ was initially constructed. Lundqvist and Lindner (2017) also examined DIF, and identified items 13, 22, 44, 14, and 19 as showing DIF between autism and non-autism groups. Of these 5 items, items 22 and 14 were identified as showing DIF in all three of our analyses.

Multidimensionality

To check our DIF results, we investigated how tenable the assumption of unidimensionality is. Then, we repeated the main analyses for the two dimensions "social interaction" and "attention to detail" from the two-factor confirmatory model (Hoekstra et al., 2008) and the four dimensions "socialness," "patterns," "understanding others/communication," and "imagination" from a four-factor exploratory model (Stewart & Austin, 2009) that describes the structure of 43 out of 50 items. Versions of Figures 1 and 2 from the main analysis are provided in the Supplemental Material for each analysis (Figures A–L).

Testing unidimensionality. To test unidimensionality, we performed an exploratory factor analysis and a confirmatory factor analysis, following the same procedure as Murray, Booth, et al. (2017), using the *psych* (Revelle, 2018) and *lavaan* (Rosseel, 2012) R-packages. A scree plot showed that either two or three factors are required. Parallel analysis (Hayton, Allen, & Scarpello, 2004) and the minimum average partial criterion showed that four factors are required. The ratio of the first and second eigenvalues was 5.7. The fit of the unidimensional confirmatory factor model was bad (comparative fit index (CFI) = 0.934) to acceptable (root mean square error approximation (RMSEA) = 0.52, standardized root mean

square residual (SRMR) = 0.092, $\chi^2(1175) = 2471.2$, $\chi^2/df = 2.1$, Schermelleh-Engel & Moosbrugger, 2003). In short, the exploratory factor analysis indicated that including two to four dimensions is advised, and the confirmatory factor analysis indicated that unidimensionality is more or less acceptable.

Two-factor model. We repeated all three analyses for those 40 items that contribute to the latent "social interaction" construct and those 10 items that contribute to the latent "attention to detail" construct (Hoekstra et al., 2008). In the Rasch mixture analysis, a model with two groups, broadly corresponding to autism diagnosis, was found to fit best for "social interaction", and a model with three groups was found to fit best for "attention to detail". In the regression tree analysis, autism diagnosis provided the first split for both "social interaction" and "attention to detail" models, and in both models, the non-autism group was split by sex. In the "social interaction" model, the autism group was also split by sex. Traditional analyses were also performed for both models. All three analyses indicate DIF within items that measure "social interaction" and "attention to detail," with a large degree of overlap with the main analyses in which items show DIF: There was DIF in items 11, 21, and 26 of the "social interaction" scale, and items 29, 30, and 49 of the "attention to detail" scale (see Figures A–D in the Supplementary Material).

Four-factor model. We also repeated all three analyses for those 12 items that contribute to the "socialness" construct, 8 items that contribute to the "patterns" construct, 16 items that contribute to the "understanding others/communication" construct, and 7 items that contribute to the "imagination" construct (Stewart & Austin, 2009). In the Rasch mixture analysis, a model with respectively three, four, two and three groups fitted best, indicating that for every subscale, multiple latent groups that show DIF were recovered. For the regression trees, the most important split was by autism diagnosis for all four subscales. For "socialness" there were no further splits. The autism group was split by age (at 51 years) for the "patterns" construct. For "understanding others" both autism and non-autism groups were split by sex, and for "imagination," only the non-autism group was split by sex. Traditional methods were also applied for these subscales. Again, the results from the main analysis were confirmed, and many of the same items displayed DIF: items 11, 22, and 26 from the "socialness" subscale, item 29 from the "patterns" subscale, items 30 and 21 from the "understanding others" subscale, and items 14 and 40 of the "imagination" subscale.

Discrimination parameters

Our discussion has focused on the Rasch model, also called the one-parameter logistic model, which allows assessment of bias in difficulty of items. A two-parameter logistic model allows for the assessment of bias in discrimination of items as well, also known as nonuniform DIF (Magis et al., 2010). Discrimination refers to how well a particular item distinguishes between persons with different values on the latent trait. Murray, Allison, et al. (2017) did not find large differences in discrimination parameters between men and women. One example of nonuniform DIF would be that the same item may be highly informative for men, showing distinct responses in men with high autistic traits and men with moderate autistic traits, but may be uninformative in women, showing random responses in women with high and low autistic traits.

To examine whether there are differences in discrimination parameters between groups with and without autism, we used the *difR* package. To our knowledge, the Rasch mixture models and Rasch regression trees have not been extended yet to models with discrimination parameters. For three out of five methods available in the *difR* package, the purification algorithm converged and the results could be interpreted. Raju's method indicates that all but seven items show nonuniform DIF. The logistic regression method indicates that items 21, 31, 47, and 50 show nonuniform DIF. Lord's method indicates that only item 21, "I don't particularly enjoy reading fiction," shows non uniform DIF, as it does not discriminate at all between those with low and high autistic traits in the autism group, while it does somewhat in the non-autism group.

Discussion

In this article, we studied in a sample of 408 participants whether there are groups of participants for whom the items of the AQ show DIF. This means that we looked for items to which members of different groups are likely to provide different answers, even if they are equal in autistic traits. From the different analyses we conducted, a set of eight items emerged that showed DIF consistently across analyses, and these items were generally negatively phrased. We conclude that group comparisons, between groups with and without autism, and between sexes, may be compromised by these items. The worst performing items are not included in the short forms AQ-28 and AQ-10, which suggests that it is better to use either of these when comparing groups.

Three main analyses were conducted. First, a Rasch mixture model showed that differences in item parameters were largest if we split participants into two latent groups. Earlier studies have not looked at latent groups that could show DIF. The definition of the two groups is almost equivalent to autism diagnosis, although it also somewhat corresponds to sex. Therefore, we can conclude that there is DIF within the AQ, and that autism and non-autism groups are the main groups of interest for the detection of DIF, as there does not seem to be an unmeasured division into subgroups that is more informative.

Second, a Rasch regression tree model showed that if we split the data on demographic characteristics, there were four discernible groups that respond differently to items, even given equal autistic traits in the participants. These were men without autism, women without autism, people 50 years and younger with autism, and people older than 50 years with autism. Earlier studies only looked at a single factor (sex or autism) at a time, which means these interactions between sex, autism, and age could not be detected before. The split between young and old was unstable: Small changes in the sample could remove the split, or split the autism group by sex instead. Therefore, it remains unclear whether we can use the AQ to compare old and young groups with autism, or men and women with autism. However, we can conclude that comparisons between men without autism and women without autism may be compromised by DIF, as well as comparisons between all participants with and without autism, when the full AQ is used.

Third, traditional DIF analyses were performed, splitting the group into autism and non-autism groups. Three traditional methods for detecting DIF were used, and each showed that there are items within the AQ that show DIF between autism and non-autism groups. This corroborates the results from the previous two analyses.

To understand what may cause certain items to show DIF, a number of item characteristics were examined. The items that show DIF contain qualifiers that may make the items more difficult to understand. However, these qualifiers are also used in items that do not show DIF. The only characteristic that seems to be unique to DIF items seems to be negative phrasing. This suggests that when developing new autistic trait questionnaires it is better to refrain from negative phrasing, especially as many side effects of negative phrasing have been noted in the literature (Barnette, 2000).

To our knowledge, this was the first analysis of the AQ that also studied DIF between different levels of education, and different age groups. The results showed that different educational groups did not show DIF, but this could be due to the limited range of educational levels we had in our sample. As noted before, stability analyses showed that DIF between age groups with autism may not be stable. Also, age groups did not differ in the two additional subscale analyses. Our initial analyses showed that young and old participants within the group with autism did differ in how items functioned, with the cutoff point between the young and old determined to be at 50 years. Although this result is not unequivocal, some care should be taken when drawing conclusions about differences on the AQ between young and old groups.

One limitation of this study is that the proportions of men and women are not the same between autism and non-autism samples. Therefore, DIF analyses between autism and non-autism are confounded by sex, and vice versa. However, the sex imbalance between the autism and non-autism groups was small, especially in light of the difference in prevalence of autism between men and women that is reported (Lai, Lombardo, & Baron-Cohen, 2014). Also, in the regression tree analysis, sex and autism were included separately, and these showed separate DIF effects. Finally, it is important to note that the autism group in this study may be dissimilar from some other autism groups in the literature. First, they were mostly diagnosed late in life, and it is difficult to know whether the results generalize to participants who age with an early life autism diagnosis. Second, they were highly educated, while it is known that many people with autism have difficulties completing educational programs (Lai et al., 2014). For these reasons, they might form a special subgroup of people with autism with increased coping mechanisms.

Interestingly, the three items that showed the most convincing DIF in our analyses were items that showed misfit in the analysis by Lundqvist and Lindner (2017). Both analyses indicate that these items are not measuring what they are supposed to be measuring: Women without autism are more prone to provide the response that is typical of autism than other groups for item 30 (regarding noticing small changes), while Lundqvist and Lindner note that on item 30, those with low autistic traits on average provide the response that is typical of autism (which they call misfit). Therefore, our finding for these items can be considered a replication of their study, and this particular item should perhaps be the first to be adapted if there is ever an update of the full AQ.

Our current recommendation is to avoid a number of items that have been consistently shown in our analyses to be biased. There are alternatives to dropping items completely. For example, it has been shown for factor analytic models that if at least two items are unbiased, that is, partial measurement invariance is demonstrated, latent mean differences between groups can be interpreted (Steenkamp & Baumgartner, 1998). Many items in our analyses showed no bias, which provides support for this strategy. However, this strategy requires that a model is fitted to estimate latent means and compare these between groups, as even with partial measurement invariance, sum scores have been shown to be biased in group comparisons (Steinmetz, 2013). We consider removing the biased items to be the simpler solution, as it allows for the interpretation of the sum score as it is used in practice, rather than having to fit a model to arrive at a latent factor score for every individual.

In this article, the conventional dichotomous scoring rule for the AQ was used, taking the "definitely" and "slightly" options together. Dichotomous scoring allowed us to apply the item response theory framework to examine parameters for individual items. Recently, analyses using the four-point scoring showed that this scoring may be better, because the resolution is higher (Austin, 2005;

Stevenson & Hart, 2017). However, the two methods that do not make a priori assumptions on the compared groups, that is, the Rasch mixture and Rasch regression tree methods, have to our knowledge not been extended to more than two response options. Although the same items that show DIF in a two-point scoring method would most probably show DIF with a different scoring method, we do not have the statistical framework to examine this in the same way.

In the main analyses, unidimensionality was assumed, which could potentially lead to inappropriate findings of DIF if this assumption is violated. Our confirmatory analysis of a unidimensional model showed that the fit was acceptable. However, our exploratory analyses showed that two to four dimensions were appropriate for describing the structure, suggesting that unidimensionality is violated. Therefore, although the DIF results of the main analyses should not be affected, it is reasonable to also assess DIF within separate dimensions. We reran the main analyses for six subscales from the literature: two from a confirmatory model (Hoekstra et al., 2008) and four from an exploratory model (Stewart & Austin, 2009). Items that were identified as showing DIF in the main analyses were also identified as showing DIF within these subscales. So even though the literature indicates that the assumption of unidimensionality is violated, which could cause DIF, the DIF findings of our main analyses seem robust.

In the main analyses, we focused on the Rasch model, also called the one-parameter logistic model, which allowed us to study biases in item difficulties. However, bias may also be present in how well items discriminate between participants with varying levels of autistic traits. Therefore, we also fitted a two-parameter logistic model, to investigate differences in discrimination parameters between groups with and without autism. Different methods indicated that either very many or very few items showed this bias, making it difficult to draw conclusions. Item 21, "I don't particularly enjoy reading fiction," did not discriminate well between those with low and high autistic traits in the autism group.

With the results of this study, we think the full AQ can be improved to provide a more reliable and informative measure of autistic traits. Furthermore, the effects of negative phrasing should be taken into account in the construction of questionnaires for use in populations with autism. For now, our advice when comparing groups on the AQ, such as people with and without autism, men and women, or young and old participants, is to use the AQ-28, as it does not contain many of the items that proved dysfunctional in our analysis.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by Innovational Research Incentives Scheme VICI (NWO-MagW) awarded to Hilde M. Geurts (Grant Number 453-16-006).

Notes

1. Throughout this article, we use the term autism, as this is the preferred term across community groups (Kenny et al., 2016). "Non-autism" or "without autism" is used to refer to groups without a diagnosis of autism (as in e.g. Leekam et al., 2007), as our non-autism sample contains participants with diagnoses other than autism (Lever & Geurts, 2016).
2. A sixth study (Stewart, Allison, Baron-Cohen, & Watson, 2015) also looked at AQ items and groups, but the method was too dissimilar to directly compare the results with those of the other five studies. They used Mokken scaling to find hierarchies of items, within a university student sample, and an autism sample. They found that a single scale can be formed for the university student sample, and three scales for the autism sample.

Supplemental material

Supplemental Material for this article is available online.

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.

Austin, E. J. (2005). Personality correlates of the broader autism phenotype as assessed by the Autism Spectrum Quotient (AQ). *Personality and Individual Differences*, *38*, 451–460.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, *60*, 361–370.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17.

Freeth, M., Sheppard, E., Ramachandran, R., & Milne, E. (2013). A cross-cultural comparison of autistic traits in the UK, India and Malaysia. *Journal of Autism and Developmental Disorders*, *43*, 2569–2583.

Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch mixture models with package psychomix. *Journal of Statistical Software*, *48*(7), 1–25.

Grove, R., Hoekstra, R. A., Wierda, M., & Begeer, S. (2017). Exploring sex differences in autistic traits: A factor analytic study of adults with autism. *Autism*, *21*, 760–768.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205.

Hoekstra, R. A., Bartels, M., Cath, D. C., & Boomsma, D. I. (2008). Factor structure, reliability and criterion validity

of the Autism-Spectrum Quotient (AQ): A study in Dutch population and patient groups. *Journal of Autism and Developmental Disorders*, *38*, 1555–1566.

Hoekstra, R. A., Vinkhuyzen, A. A., Wheelwright, S., Bartels, M., Boomsma, D. I., Baron-Cohen, S., … van der Sluis, S. (2011). The construction and validation of an abridged version of the Autism-Spectrum Quotient (AQ-Short). *Journal of Autism and Developmental Disorders*, *41*, 589–596.

Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., & Pellicano, E. (2016). Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism*, *20*, 442–462.

Kloosterman, P. H., Keefer, K. V., Kelley, E. A., Summerfeldt, L. J., & Parker, J. D. (2011). Evaluation of the factor structure of the Autism-Spectrum Quotient. *Personality and Individual Differences*, *50*, 310–314.

Lai, M.-C., Lombardo, M. V., & Baron-Cohen, S. (2014). Autism. *The Lancet*, *383*, 896–910.

Lau, W. Y. P., Gau, S. S. F., Chiu, Y. N., Wu, Y. Y., Chou, W. J., Liu, S. K., & Chou, M. C. (2013). Psychometric properties of the Chinese version of the Autism Spectrum Quotient (AQ). *Research in Developmental Disabilities*, *34*, 294–305.

Lau, W. Y. P., Kelly, A. B., & Peterson, C. C. (2013). Further evidence on the factorial structure of the Autism Spectrum Quotient (AQ) for adults with and without a clinical diagnosis of autism. *Journal of Autism and Developmental Disorders*, *43*, 2807–2815.

Leekam, S. R., Nieto, C., Libby, S. J., Wing, L., & Gould, J. (2007). Describing the sensory abnormalities of children and adults with autism. *Journal of Autism and Developmental Disorders*, *37*, 894–910.

Lever, A. G., & Geurts, H. M. (2016). Psychiatric co-occurring symptoms and disorders in young, middle-aged, and older adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *46*, 1916–1930.

Lever, A. G., & Geurts, H. M. (2018). Is older age associated with higher self-and other-rated ASD characteristics? *Journal of Autism and Developmental Disorders*, *48*, 2038–2051.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lundqvist, L. O., & Lindner, H. (2017). Is the autism-spectrum quotient a valid measure of traits associated with the autism spectrum? A Rasch validation in adults with and without autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *47*, 2080–2091.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847–862.

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement*, *22*, 357–367.

McConachie, H., Mason, D., Parr, J. R., Garland, D., Wilson, C., & Rodgers, J. (2017). Enhancing the validity of a quality of life measure for autistic people. *Journal of Autism and Developmental Disorders*, *48*, 1596–1611.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

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.

Murray, A. L., Booth, T., Auyeung, B., McKenzie, K., & Kuenssberg, R. (2017). Investigating sex bias in the AQ-10: A replication study. *Assessment*. Advance online publication. doi:10.1177/1073191117733548

Murray, A. L., Booth, T., McKenzie, K., & Kuenssberg, R. (2016). What range of trait levels can the Autism-spectrum Quotient (AQ) measure reliably? An item response theory analysis. *Psychological Assessment*, *28*, 673–683.

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.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.

Revelle, W. (2018). psych: Procedures for personality and psychological research (Computer software manual, R package version 1.8.4). Retrieved from http://cran.r-project.org/web/packages/psych/

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*(2), 1–36.

Russell-Smith, S. N., Maybery, M. T., & Bayliss, D. M. (2011). Relationships between autistic-like and schizotypy traits: An analysis using the Autism Spectrum Quotient and Oxford-Liverpool Inventory of Feelings and Experiences. *Personality and Individual Differences*, *51*, 128–132.

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: A systematic review of the Autism-spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular Autism*, *6*(1), Article 2.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

Steinmetz, H. (2013). Analyzing observed composite differences across groups. Is partial measurement invariance enough? *Methodology*, *9*, 1–12.

Stevenson, J. L., & Hart, K. R. (2017). Psychometric properties of the Autism-spectrum Quotient for assessing low and high levels of autistic traits in college students. *Journal of Autism and Developmental Disorders*, *47*, 1838–1853.

Stewart, M. E., Allison, C., Baron-Cohen, S., & Watson, R. (2015). Investigating the structure of the autism-spectrum quotient using Mokken scaling. *Psychological Assessment*, *27*, 596–604.

Stewart, M. E., & Austin, E. J. (2009). The structure of the Autism-Spectrum Quotient (AQ): Evidence from a student

sample in Scotland. *Personality and Individual Differences*, *47*, 224–228.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*, 289–316.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*, 323–348.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

Totsika, V., Felce, D., Kerr, M., & Hastings, R. P. (2010). Behavior problems, psychiatric symptoms, and quality of life for older adults with intellectual disability with and without autism. *Journal of Autism and Developmental Disorders*, *40*, 1171–1178.

Verhage, F. (1964). *Intelligentie en leeftijd onderzoek bij Nederlanders van twaalf tot zevenenzeventig jaar* [Intelligence and age research with Dutch people aged twelve to seventy seven years] (Doctoral thesis). Van Gorcum Prakke en Prakke, Assen, The Netherlands.