Manuscript Number: JADD-07-239R2

Title: Exploratory and Confirmatory Factor Analysis of the Autism Diagnostic Interview-Revised

Article Type: Article

Section/Category:

Keywords: Autism; Autism Diagnostic Interview-Revised; exploratory factor analysis; and multi-group confirmatory factor analysis

Corresponding Author: Dr. Thomas W. Frazier, Ph.D.

Corresponding Author's Institution: The Cleveland Clinic

First Author: Thomas W. Frazier, Ph.D.

Order of Authors: Thomas W. Frazier, Ph.D.; Eric A Youngstrom, Ph.D.; Cynthia Kubu, Ph.D.; Leslie Sinclair, CC/SLP, BCBA; Ali Rezai, MD

Manuscript Region of Origin:

Abstract: The factor structure of the Autism Diagnostic Interview-Revised (ADI-R) algorithm items was examined using exploratory (EFA) and confirmatory (CFA) factor methods. The ADI-R was completed for 1170 youths and adults (ages 2-46). Results of EFAs indicated strong support for two factor structure, with social communication and stereotyped behavior factors. CFAs computed in a holdout sub-sample indicated roughly equal support for the above described two-factor model and a three factor model separating peer relationships and play from other social and communicative behaviors. Multi-group CFAs suggested that both two and three factor models showed good stability across age, with only slight changes in factor relationships. These findings indicate that the current ADI-R structure be revised to more accurately reflect the relationships between subscales.

Response to Reviewers:

June 7, 2007

Dr. Gary B. Mesibov
Dept. of Psychiatry
University of North Carolina School of Medicine
Chapel Hill, NC

Dear Dr. Mesibov and reviewers:

I am submitting the revised manuscript entitled, “Exploratory and Confirmatory Factor Analysis of the Autism Diagnostic Interview-Revised” 07-239R1 for re-review by the Journal of Autism and Developmental Disorders. The manuscript continues to be 25 pages long (including title, abstract, references, and tables). We thank the editor and reviewers for their thoughtful and detailed comments and believe the manuscript is not significantly improved. Below is a point-by-point response to the reviewers comments.

Comments for the Author:

Reviewer #1:

1)Overall, this is a well-done, well-written study, taking advantage of the AGRE dataset. I was puzzled why the nonverbal children were only included in a paragraph as a footnote. Given that the paper is not particularly long, it would seem worthwhile to go ahead and include their data and address them to the degree possible (basically expand the footnote into real text).

We thank you for the positive comments. As to the non-verbal subjects, we did not include their data in the primary analyses because these individuals do not have scores for two of the twelve sub-domains. However, we did impute data for these individuals and re-examined the structure of the ADI-R subdomains. When we imputed data for these items on these individuals the structure was very similar. However, we stick to the notion of only reporting these in the footnote because of the inherent uncertainty of imputing values for these participants. If the editor prefers we would be a happy to report these analyses as well, but for brevity and clarity we continue to leave the brief description in the footnote. We would also note that the exclusion of non-verbal subjects does not appear to have hurt the range of the variables as the full range is well represented on each variable but that it likely did improve the skew and kurtosis of the variables an important consideration when computing covariances.

2) The primary limitation of the study has to do with the sample. It is not any more a limitation than any other sample, but just that this is a unique sample. The verbal children do particularly well on the Raven's

and PPVT and so it is not quite clear how representative a sample this is of children with autism. As the authors carefully note repeatedly, this is also a sample primarily derived from families with two or more children with autism; thus, the number of individuals without autism is low and there may be something different (although so far no one has been able to really find it) about children with autism from multiplex families than simplex families. More attention to the uniqueness of the sample in terms of psychometric properties would be worthwhile. Also, clinical diagnoses are not available for this sample and so some acknowledgement or simply discussion about how this sample might be different from more typical research samples where there is more cross validation in terms of diagnosis would be useful. One could certainly argue in some ways that this might even be a more accurate representation of people who use "autism" to refer to their children rather than those with formal diagnoses in clinics.

We now include in the discussion further acknowledgement of the special nature of this sample (pg. 15). We attempt to give attention to the psychometric properties of the sample by noting the range of scores (Table 2), the reliability of sub-domains (pg. 7), the inter-rater reliability needed for the raters (pg. 6), and the mean's and standard deviations for each sub-domain. If the editor would prefer some other psychometric analyses to further clarify the nature of the sample and the properties of these scores we would be happy to include this in future revisions.

3) The data in the ADI-R are not ratio-data but are actually ordinal data. It was not clear if this fact was taken into account, but it would seem crucial to do so. It was clear to me why the approach was taken to analyze the data on the subdomains, but it was not clear to me how it might have been different had individual items been used. Using the subdomains assumes that the items in those domains are more strongly related to each other than to other items, which may not be the case, given the very limited psychometric data used to determine the initial ADI algorithms. Thus, one wonders if the authors had used individual items how it might have changed the results. Again, given that the analyses are set up and that this is not a very long paper, it might be worth exploring this possibility, or providing a more detailed argument as to why that would not be appropriate.

We now include a further discussion of how future work would benefit from examining individual items and their covariances within and across sub-domains (pg. 13). We agree that examining individual items is an interesting possibility. We did not include these analyses because the purpose of the present paper was to evaluate the diagnostic algorithm items as they are typically used clinically and not to explore the larger notion of the true structure of autism symptoms. This notion has been explored previously and we cite some of this work in the introduction. Because the DSM-IV based algorithm is used clinically we felt it was important to focus on these sub-domains.

As to the ordinal data issue, we performed data transformations to attempt to normalize the distribution of scores and minimize the impact of skew or kurtosis on the resulting correlation and covariance matrices. However, given the reviewers comments we also went back and re-ran all of the primary analyses using polychoric correlations and asymptotic covariance matrices. The results are highly similar. In fact, the polychoric correlation matrix is extremely similar to the pearson-product moment correlation matrix. Therefore, we continue to present the initial analyses. We thank the reviewer for pointing out this issue and helping us to make sure that are findings are not specific to the particular analysis. This issue is now discussed on pg. 9. As can be seen, given the multiple iterations of analyses now conducted, careful attention has been paid to the scaling of the data to ensure that results were robust across analytic approaches.

As stated previously, we chose to examine the sub-domains because they are the scores that are used in the diagnostic algorithm and these scores generally have better range. We agree that examining individual items would be interesting, but acknowledge that even though individual items tend to have greater specificity many of the items still tap heterogeneous behaviors. For example, there are many different types of self-stimulatory behaviors in autism but these tend to be captured by only one or two items (items 69 and 71 most frequently). So the exact gain in specificity is unknown.

4) Given that the RMSEAs are higher than one would want, exploring avenues that might yield factors that account for more of the variance would seem appropriate, given the work that the authors have already gone to. Inclusion of non-algorithm items also might be appropriate, again given that this is a far larger dataset and much more is known about the ADI-R now than when the algorithms were proposed. This requires more work on the part of the authors, but might add significantly to the contribution that it can make.

We agree that examining non-algorithm items may be very useful in future work attempting to specific the true structure of autism symptoms and for refining phenotypic heterogeneity. However, this was not the primary purpose of the present study and therefore these analyses were not conducted. We appreciate the reviewers' consideration that this requires a very different and intensive body of work.

In terms of the fit statistics, we agree that the RMSEAs are not optimal and we tried to look at several other plausible a priori models. However, we did not engage in post-hoc model fitting because this often results in sample dependent solutions that appear useful initially but are difficult to replicate. We would also note that model misspecification can occur at multiple levels and the lack of optimal fit may suggest the need to trim sub-domains, collapse sub-domains, or include other sub-domains not currently included. Since the scope of the paper was to evaluate the existing sub-domains we did not go further in trying to examine all of these possibilities.

5) Similarly, the argument used that "3" codes were included in the analyses is understandable, but it is not clear that the administrators of the ADI-R made reliable distinctions between "2"s and "3"s, so one could argue that this is not necessarily appropriate. More comments about the degree to which the analyses would have been changed had 3s not been used (particularly because other investigators have typically not used 3s unless they have made a specific attempt to be reliable on the 2/3 distinction).

We included the 3 codes to increase the range of scores. However, we acknowledge that coders may not have made reliable distinctions. In most cases 4-point scales will be more useful than 3-point scales and the ADI-R includes descriptions for each item and scale point. This likely involves some level of distinction being made and the descriptions of 2 and 3 point responses are fairly clear. Therefore we believe it is likely that using 4-point scales probably is useful by increasing the range and true score variability of items. However, even in the event that the 2- 3- distinction was not made reliably, the results are unlikely to have been substantially altered because the covariances would be highly similar under both scenarios. In fact, initial analyses were performed with transformed data and the results were highly similar. We now note that in footnote 2.

6)Overall, the findings from the study fall in line with recent studies and proposals by several other authors (e.g. Szatmari's group) that a two-factor structure (social/communication and repetitive behaviors, including stereotyped speech) more accurately reflect the nature of autism than the current DSM-IV three-factor structure used in the ADI-R. It is a bit surprising that the authors did not find distinctions between the two proposed areas within stereotyped behavior of insistence on sameness and sensory motor behaviors. It is possible that this might have been because collapsing items within the subdomains and limiting the factor analysis to algorithm items reduced the number of items contributing to these results. This is a place where further investigation might be worthwhile.

The failure to find a distinction between insistence on sameness and sensory motor behaviors is consistent with previous factor analytic work examining the sub-domains of the diagnostic algorithm and we agree with the reviewer that this is likely due, at least in part, to the collapsing of items. However, we also believe it is likely due to the small number of items tapping these areas on the ADI-R and the lack of specificity in item content. Other studies using greater item specificity have found distinctions and that is why we did not feel the need to go further in this area. We feel it simply points out a limitation of this diagnostic instrument.

7) Overall, there is a great deal that is worthwhile in this paper. It is very likely to make a substantial contribution as researchers are beginning to consider organizing DSM-V. It adds to an existing series of papers calling for reconsideration of an independent "communication" factor from social, and the question of the specific placement of stereotype language within communication as opposed to as a repetitive behavior. It would be stronger if the very large number of nonverbal cases were directly addressed rather than just in a

footnote. It would also be helpful to clarify in the method section the ranges of ages (how young were the youngest children and how old were the oldest) as well as ranges of scores on the Raven's and PPVT. Even if one removed all the nonverbal children, it seems surprising that the Raven's and PPVT scores were so high. Just a comment about this would be helpful. Distribution across ethnicity should be reported as well as some comment about the lack of a clinical diagnosis (at least a standardized clinical diagnosis).

It is not clear to us how not having clinical diagnoses is a limitation (or strength; please see Reviewer 3's comments on the issue of diagnoses). This contention is based upon two pieces of information. First, all participants had one of the gold standard research measures, the ADI-R, and many also had another very good observation scale, the ADOS. Second, the present study was focused on the factor structure and therefore the emphasis was on capturing a range of functioning, not pinning down the most accurate diagnostic description. If we are missing something we would appreciate the editor's guidance and would be happy to add this caveat.

As stated above, we continue to address the non-verbal cases in the footnote (please see reasoning above).

The range of ages is now included on page 5. The range of PPVT and Raven's scores is now included with table 1 on pg. 22.

Absolutely minor comments:

1. Page 5. Which Institutional Review Boards approved the procedures of the study?

We now include the name of the IRB on pg. 5.

2. Bottom of Page 8. It would be really helpful to tabulate the different two and three factor models. This is a particularly useful summary of what has been done in the past. My preference would be to actually name the different approaches instead of using 2a and 3b, to give them a name, but this is minor.

We found it difficult to name each without being misleading or reifying constructs so we stuck with the number letter convention. This is consistent with numerous other recent factor analytic studies. We would be happy to supplement what is already included with table 3 if the editor prefers. However, basic information about the models is already included in the text and at the bottom of table 3. We are hesitant to include an additional table just for this information for space reasons but we would be happy to do so if the editor would prefer. Please give us some guidance on this issue.

3. In the discussion, some comment about what might have been different had the distribution of this sample been different. That is, there were very few individuals without an ADI-R or ADOS diagnosis. That

does not mean that the sample is not excellent for the purposes intended, but it is quite a different undertaking than, for example, analyses such as those by John Constantino with a full population, which seemed to yield some different kinds of factors.

We would note that the excellent work by John Constantino is difficult to compare to the present work because of the differences in population and the differences in indicators used for the analyses. On pg. 13 we now note that it is possible that future work using more heterogeneous populations may find a more similar structure to DSM. However, beyond that possibility, we would prefer not to speculate as to what we think would happen since this is a very open question and we would have no real guidance as to what would be the most appropriate things to speculate.

4. Conceptually, the most important things would seem to be to make it clear what assumptions were made in the organization of the data (e.g., leaving them in subdomains, only analyzing algorithm items) and what effects this particular sample might have on these findings, because the findings themselves are quite strong.

We hope that with the reviewer’s guidance we have now been clear conceptually about the organization of the data and the effects of the sample on the findings. However, if the editor feels that this has not occurred satisfactorily then we are very open to trying to be even more explicit.

Reviewer #3:

1)As the authors have noted in their literature review, there have been a number of important factor analysis studies published in the past several years in which an attempt has been made to better define the diagnostic characteristics for the autism spectrum, especially in terms of which symptoms are the most important in defining the core social communication problems. This study adds to this literature in a number of important ways. The evaluation and diagnosis of the subjects is, in my opinion, the best of any of the factor analysis studies.

We thank the reviewer for this positive comment and note that this is actually part of our argument (see response to reviewer 1 major comment #7) for not including comments about the limitations of not having clinical diagnoses.

2) It also has a very large sample size, far more than that of the other studies, which has permitted the authors to do exploratory and confirmatory factor analyses. The statistics which they use for the latter are more sophisticated than those used in most previous factor analytic studies. They have also looked at two age groups, although I have always been somewhat skeptical of the "age 4-5" values when the children are

above 9 - that they found the results for each age group to be highly correlated may be due more to a "halo" effect (wherein parents of older children tend to give the same answer for the 4-5 age group as they do for the "present" age group). They might wish to discuss whether they think this may be so.
I share the reviewer's general concern about retrospective reporting. However, several key features of the autism population should be noted. In many caregivers, careful account of development has now become a part of their lives and thus this population may be more accurate than we give them credit for. Also, we would note that we compared age groups from 2-6 and 7 and above. If the latter group had significant reporting biases this is likely to have shown up in factorial or measurement invariance, but it did not. We think this argues against a significant difficulty with obtaining these reports.

3) It is unfortunate that the non-verbal subjects were excluded, though I understand why this was done and agree that given that the investigators were using the ADI-R summary scores it was necessary. I was pleased to read in Footnote 1 that they did go back and do an analysis of the excluded subjects in which they imputed the most severe rating to the verbal items. As it is, their initial study populations may have been representative of subjects with Asperger's rather than autism.

We thank the reviewer in understanding our reasoning for this and note that the reviewer shares our perspective arguing against directly reporting the non-verbal subjects (see note to reviewer 1, major comment #1). We doubt that this population was more representative of subjects with Asperger's though given the level of communication deficit still observed in a large portion of the sample. See table 1 for support for this perspective.

4) The authors have chosen not to replace "3" scores with "2" as the authors of the ADI-R have suggested. While it is true that this increased the overall variance, there are some who might frown on this and say that the increased variance is spurious in that "3" and "2" are not reliable differences. My colleagues and I have done factor analyses on moderately large data sets in which the replacement was both done and not done. We have not found that replacement made much of a difference to the final factor analysis - though we were using individual variable values rather than the summary values used in the present investigation.

We agree with the authors that this transformation does not make much of a difference as our preliminary analyses also found very little difference. We think the psychometric argument for reporting untransformed data still stands but we acknowledge that it is an open empirical issue.

5) A comparison of models of various factor structures indicates that a two-factor and a three-factor model best express the structure of the data. The latter (3-factor model) is fairly similar to what others have found. The value of the authors findings is that they came at the question from a different direction than did previous investigators, which serves to reinforce previous findings. The authors are correct in stating that

studies such as the present one could be important in determining how the diagnostic algorithms for autism in DSM-V or ICD-11 could be improved.

We thank the reviewer for the positive comments and agree with the sentiments.

We look forward to the next steps in this process and appreciate the editor’s guidance for several of the issues discussed above.

Sincerely,

Thomas W. Frazier, PhD
Associate Staff
Behavioral Medicine CR11
The Cleveland Clinic
2801 Martin Luther King Jr. Drive
Cleveland, OH 44104

June 7, 2007

Dr. Gary B. Mesibov
Dept. of Psychiatry
University of North Carolina School of Medicine
Chapel Hill, NC

Dear Dr. Mesibov:

I am submitting the revised manuscript entitled, "Exploratory and Confirmatory Factor Analysis of the Autism Diagnostic Interview-Revised" 07-239R1 for re-review by the *Journal of Autism and Developmental Disorders*. The manuscript continues to be 24 pages long (including title, abstract, references, and tables). Below is a point-by-point response to the reviewers comments.

Comments for the Author:

Reviewer #1:

1)Overall, this is a well-done, well-written study, taking advantage of the AGRE dataset. I was puzzled why the nonverbal children were only included in a paragraph as a footnote. Given that the paper is not particularly long, it would seem worthwhile to go ahead and include their data and address them to the degree possible (basically expand the footnote into real text).

We thank you for the positive comments. As to the non-verbal subjects, we did not include their data in the primary analyses because these individuals do not have scores for two of the twelve sub-domains. However, we did impute data for these individuals and re-examined the structure of the ADI-R subdomains. When we imputed data for these items on these individuals the structure was very similar. However, we stick to the notion of only reporting these in the footnote because of the inherent uncertainty of imputing values for these participants. If the editor prefers we would be a happy to report these analyses as well, but for brevity and clarity we continue to leave the brief description in the footnote. We would also note that the exclusion of non-verbal subjects does not appear to have hurt the range of the variables as the full range is well represented on each variable but that it likely did improve the skew and kurtosis of the variables an important consideration when computing covariances.

2) The primary limitation of the study has to do with the sample. It is not any more a limitation than any other sample, but just that this is a unique sample. The verbal children do particularly well on the Raven's and PPVT and so it is not quite clear how representative a sample this is of children with autism. As the authors carefully note repeatedly, this is also a sample primarily derived from families with two or more children with autism; thus, the number of individuals without autism is low and there may be something different (although so far no one has been able to really find it) about children with autism from multiplex families than simplex families. More attention to the uniqueness of the sample in terms of psychometric properties would be worthwhile. Also, clinical diagnoses are not available for this sample and so some acknowledgement or simply discussion about how this sample might be different from more typical research samples where there is more cross validation in terms of diagnosis would be *useful. One could certainly argue in some ways that this might even be a more accurate representation of people who use "autism" to refer to their children rather than those with formal diagnoses in clinics.*

We now include in the discussion further acknowledgement of the special nature of this sample (pg. 15). We attempt to give attention to the psychometric properties of the sample by noting the range of scores (Table 2), the reliability of sub-domains (pg. 7), the inter-rater reliability needed for the raters (pg. 6), and the mean's and standard deviations for each sub-domain. If the editor would prefer some other psychometric analyses to further clarify the nature of the sample and the properties of these scores we would be happy to include this in future revisions.

3) The data in the ADI-R are not ratio-data but are actually ordinal data. It was not clear if this fact was taken into account, but it would seem crucial to do so. It was clear to me why the approach was taken to analyze the data on the subdomains, but it was not clear to me how it might have been different had individual items been used. Using the subdomains assumes that the items in those domains are more strongly related to each other than to other items, which may not be the case, given the very limited psychometric data used to determine the initial ADI algorithms. Thus, one wonders if the authors had used individual items how it might have changed the results. Again, given that the analyses are set up and that this is not a very long paper, it might be worth exploring this possibility, or providing a more detailed argument as to why that would not be appropriate.

We now include a further discussion of how future work would benefit from examining individual items and their covariances within and across sub-domains (pg. 13). We agree that examining individual items is an interesting possibility. We did not include these analyses because the purpose of the present paper was to evaluate the diagnostic algorithm items as they are typically used clinically and not to explore the larger notion of the true structure of autism symptoms. This notion has been explored previously and we cite some of this work in the introduction. Because the DSM-IV based algorithm is used clinically we felt it was important to focus on these sub-domains.

As to the ordinal data issue, we performed data transformations to attempt to normalize the distribution of scores and minimize the impact of skew or kurtosis on the resulting correlation and covariance matrices. However, given the reviewers comments we also went back and re-ran all of the primary analyses using polychoric correlations and asymptotic covariance matrices. The results are highly similar. In fact, the polychoric correlation matrix is extremely similar to the pearson-product moment correlation matrix. Therefore, we continue to present the initial analyses. We thank the reviewer for pointing out this issue and helping us to make sure that are findings are not specific to the particular analysis. This issue is now discussed on pg. 9. As can be seen, given the multiple iterations of analyses now conducted, careful attention has been paid to the scaling of the data to ensure that results were robust across analytic approaches.

As stated previously, we chose to examine the sub-domains because they are the scores that are used in the diagnostic algorithm and these scores generally have better range. We agree that examining individual items would be interesting, but acknowledge that even though individual items tend to have greater specificity many of the items still tap heterogeneous behaviors. For example, there are many different types of self-stimulatory behaviors in autism but these tend to be captured by only one or two items (items 69 and 71 most frequently). So the exact gain in specificity is unknown.

4) Given that the RMSEAs are higher than one would want, exploring avenues that might yield factors that account for more of the variance would seem appropriate, given the work that the authors have already gone to. Inclusion of non-algorithm items also might be appropriate, again given that this is a far larger dataset and much more is known about the ADI-R now than when the algorithms were proposed. This requires more work on the part of the authors, but might add significantly to the contribution that it can make.

We agree that examining non-algorithm items may be very useful in future work attempting to specific the true structure of autism symptoms and for refining phenotypic heterogeneity. However, this was not the primary purpose of the present study and therefore these analyses were not conducted. We appreciate the reviewers' consideration that this requires a very different and intensive body of work.

In terms of the fit statistics, we agree that the RMSEAs are not optimal and we tried to look at several other plausible a priori models. However, we did not engage in post-hoc model fitting because this often results in sample dependent solutions that appear useful initially but are difficult to replicate. We would also note that model misspecification can occur at multiple levels and the lack of optimal fit may suggest the need to trim sub-domains, collapse sub-domains, or include other sub-domains not currently included. Since the scope of the paper was to evaluate the existing sub-domains we did not go further in trying to examine all of these possibilities.

5) Similarly, the argument used that "3" codes were included in the analyses is understandable, but it is not clear that the administrators of the ADI-R made reliable distinctions between "2"s and "3"s, so one could argue that this is not necessarily appropriate. More comments about the degree to which the analyses would have been changed had 3s not been used (particularly because other investigators have typically not used 3s unless they have made a specific attempt to be reliable on the 2/3 distinction).

We included the 3 codes to increase the range of scores. However, we acknowledge that coders may not have made reliable distinctions. In most cases 4-point scales will be more useful than 3-point scales and the ADI-R includes descriptions for each item and scale point. This likely involves some level of distinction being made and the descriptions of 2 and 3 point responses are fairly clear. Therefore we believe it is likely that using 4-point scales probably is useful by increasing the range and true score variability of items. However, even in the event that the 2- 3-distinction was not made reliably, the results are unlikely to have been substantially altered because the covariances would be highly similar under both scenarios. In fact, initial analyses were performed with transformed data and the results were highly similar. We now note that in footnote 2.

6)*Overall, the findings from the study fall in line with recent studies and proposals by several other authors (e.g. Szatmari's group) that a two-factor structure (social/communication and repetitive behaviors, including stereotyped speech) more accurately reflect the nature of autism than the current DSM-IV three-factor structure used in the ADI-R. It is a bit surprising that the authors did not find distinctions between the two proposed areas within stereotyped behavior of insistence on sameness and sensory motor behaviors. It is possible that this might have been* *because collapsing items within the subdomains and limiting the factor analysis to algorithm items reduced the number of items contributing to these results. This is a place where further investigation might be worthwhile.*

The failure to find a distinction between insistence on sameness and sensory motor behaviors is consistent with previous factor analytic work examining the sub-domains of the diagnostic algorithm and we agree with the reviewer that this is likely due, at least in part, to the collapsing of items. However, we also believe it is likely due to the small number of items tapping these areas on the ADI-R and the lack of specificity in item content. Other studies using greater item specificity have found distinctions and that is why we did not feel the need to go further in this area. We feel it simply points out a limitation of this diagnostic instrument.

7) Overall, there is a great deal that is worthwhile in this paper. It is very likely to make a substantial contribution as researchers are beginning to consider organizing DSM-V. It adds to an existing series of papers calling for reconsideration of an independent "communication" factor from social, and the question of the specific placement of stereotype language within communication as opposed to as a repetitive behavior. It would be stronger if the very large number of nonverbal cases were directly addressed rather than just in a footnote. It would also be helpful to clarify in the method section the ranges of ages (how young were the youngest children and how old were the oldest) as well as ranges of scores on the Raven's and PPVT. Even if one removed all the nonverbal children, it seems surprising that the Raven's and PPVT scores were so high. Just a comment about this would be helpful. Distribution across ethnicity should be reported as well as some comment about the lack of a clinical diagnosis (at least a standardized clinical diagnosis).

It is not clear to us how not having clinical diagnoses is a limitation (or strength; please see Reviewer 3's comments on the issue of diagnoses). This contention is based upon two pieces of information. First, all participants had one of the gold standard research measures, the ADI-R, and many also had another very good observation scale, the ADOS. Second, the present study was focused on the factor structure and therefore the emphasis was on capturing a range of functioning, not pinning down the most accurate diagnostic description. If we are missing something we would appreciate the editor's guidance and would be happy to add this caveat.

As stated above, we continue to address the non-verbal cases in the footnote (please see reasoning above).

The range of ages is now included on page 5. The range of PPVT and Raven's scores is now included with table 1 on pg. 22.

Absolutely minor comments:
1. Page 5. Which Institutional Review Boards approved the procedures of the study?

We now include the name of the IRB on pg. 5.

2. Bottom of Page 8. It would be really helpful to tabulate the different two and three factor models. This is a particularly useful summary of what has been done in the past. My preference would be to actually name the different approaches instead of using 2a and 3b, to give them a name, but this is minor.

We found it difficult to name each without being misleading or reifying constructs so we stuck with the number letter convention. This is consistent with numerous other recent factor analytic studies. We would be happy to supplement what is already included with table 3 if the editor prefers. However, basic information about the models is already included in the text and at the bottom of table 3. We are hesitant to include an additional table just for this information for space reasons but we would be happy to do so if the editor would prefer. Please give us some guidance on this issue.

3. In the discussion, some comment about what might have been different had the distribution of this sample been different. That is, there were very few individuals without an ADI-R or ADOS diagnosis. That does not mean that the sample is not excellent for the purposes intended, but it is quite a different undertaking than, for example, analyses such as those by John Constantino with a full population, which seemed to yield some different kinds of factors.

We would note that the excellent work by John Constantino is difficult to compare to the present work because of the differences in population and the differences in indicators used for the analyses. On pg. 13 we now note that it is possible that future work using more heterogeneous populations may find a more similar structure to DSM. However, beyond that possibility, we would prefer not to speculate as to what we think would happen since this is a very open question and we would have no real guidance as to what would be the most appropriate things to speculate.

4. Conceptually, the most important things would seem to be to make it clear what assumptions were made in the organization of the data (e.g., leaving them in subdomains, only analyzing algorithm items) and what effects this particular sample might have on these findings, because the findings themselves are quite strong.

We hope that with the reviewer's guidance we have now been clear conceptually about the organization of the data and the effects of the sample on the findings. However, if the editor feels that this has not occurred satisfactorily then we are very open to trying to be even more explicit.

Reviewer #3:

1)As the authors have noted in their literature review, there have been a number of important factor analysis studies published in the past several years in which an attempt has been made to better define the diagnostic characteristics for the autism spectrum, especially in terms of which symptoms are the most important in defining the core social communication problems. This study adds to this literature in a number of important ways. The evaluation and diagnosis of the subjects is, in my opinion, the best of any of the factor analysis studies.

We thank the reviewer for this positive comment and note that this is actually part of our argument (see response to reviewer 1 major comment #7) for not including comments about the limitations of not having clinical diagnoses.

2) It also has a very large sample size, far more than that of the other studies, which has permitted the authors to do exploratory and confirmatory factor analyses. The statistics which they use for the latter are more sophisticated than those used in most previous factor analytic

studies. They have also looked at two age groups, although I have always been somewhat skeptical of the "age 4-5" values when the children are above 9 - that they found the results for each age group to be highly correlated may be due more to a "halo" effect (wherein parents of older children tend to give the same answer for the 4-5 age group as they do for the "present" age group). They might wish to discuss whether they think this may be so.
I share the reviewer's general concern about retrospective reporting. However, several key features of the autism population should be noted. In many caregivers, careful account of development has now become a part of their lives and thus this population may be more accurate than we give them credit for. Also, we would note that we compared age groups from 2-6 and 7 and above. If the latter group had significant reporting biases this is likely to have shown up in factorial or measurement invariance, but it did not. We think this argues against a significant difficulty with obtaining these reports.

3) It is unfortunate that the non-verbal subjects were excluded, though I understand why this was done and agree that given that the investigators were using the ADI-R summary scores it was necessary. I was pleased to read in Footnote 1 that they did go back and do an analysis of the excluded subjects in which they imputed the most severe rating to the verbal items. As it is, their initial study populations may have been representative of subjects with Asperger's rather than autism.

We thank the reviewer in understanding our reasoning for this and note that the reviewer shares our perspective arguing against directly reporting the non-verbal subjects (see note to reviewer 1, major comment #1). We doubt that this population was more representative of subjects with Asperger's though given the level of communication deficit still observed in a large portion of the sample. See table 1 for support for this perspective.

4) The authors have chosen not to replace "3" scores with "2" as the authors of the ADI-R have suggested. While it is true that this increased the overall variance, there are some who might frown on this and say that the increased variance is spurious in that "3" and "2" are not reliable differences. My colleagues and I have done factor analyses on moderately large data sets in which the replacement was both done and not done. We have not found that replacement made much of a difference to the final factor analysis - though we were using individual variable values rather than the summary values used in the present investigation.

We agree with the authors that this transformation does not make much of a difference as our preliminary analyses also found very little difference. We think the psychometric argument for reporting untransformed data still stands but we acknowledge that it is an open empirical issue.

5) A comparison of models of various factor structures indicates that a two-factor and a three-factor model best express the structure of the data. The latter (3-factor model) is fairly similar to what others have found. The value of the authors findings is that they came at the question from a different direction than did previous investigators, which serves to reinforce previous findings. The authors are correct in stating that studies such as the present one could be important in determining how the diagnostic algorithms for autism in DSM-V or ICD-11 could be improved.

We thank the reviewer for the positive comments and agree with the sentiments.

We look forward to the next steps in this process and appreciate the editor's guidance for several of the issues discussed above.

Sincerely,

Thomas W. Frazier, PhD
Associate Staff
Behavioral Medicine CR11
The Cleveland Clinic
2801 Martin Luther King Jr. Drive
Cleveland, OH 44104

Running Head: Factor Analysis of ADI-R

Exploratory and Confirmatory Factor Analysis of the Autism Diagnostic Interview-Revised

Thomas W. Frazier,[1] Eric A. Youngstrom,[2] Cynthia S. Kubu,[3]
Leslie Sinclair,[4] and Ali Rezai[5]

[1]Section of Behavioral Medicine, Cleveland Clinic

[2] Department of Psychology, University of North Carolina at Chapel Hill

[3]Section of Neuropsychology, Cleveland Clinic

[4]Center for Autism, Cleveland Clinic

[5]Center for Neurological Restoration, Cleveland Clinic

Abstract

The factor structure of the Autism Diagnostic Interview-Revised (ADI-R) algorithm items was examined using exploratory (EFA) and confirmatory (CFA) factor methods. The ADI-R was completed for 1170 youths and adults (ages 2-46). Results of EFAs indicated strong support for two factor structure, with social communication and stereotyped behavior factors. CFAs computed in a holdout sub-sample indicated roughly equal support for the above described two-factor model and a three factor model separating peer relationships and play from other social and communicative behaviors. Multi-group CFAs suggested that both two and three factor models showed good stability across age, with only slight changes in factor relationships. These findings indicate that the current ADI-R structure be revised to more accurately reflect the relationships between subscales.

Keywords: Autism, Autism Diagnostic Interview-Revised, exploratory factor analysis, and multi-group confirmatory factor analysis

Exploratory and Confirmatory Factor Analysis of the Autism Diagnostic Interview-Revised

The DSM-IV-TR describes three major domains of autism symptoms. These include qualitative impairments in social interaction, communication, and stereotyped behaviors and restricted interests. The Autism Diagnostic Interview-Revised (ADI-R) is a commonly-used diagnostic measure designed to evaluate these core symptom domains of autism. This measure includes three full scales assessing the core domains posited in the DSM. Each full scale has four subscales evaluating specific symptoms within each domain. A diagnostic algorithm is used to determine autism spectrum diagnoses.

Several studies have examined the factor structure of autism symptoms (Constantino et al., 2004; Szatmari et al., 2002; Tadevosyan-Leyfer et al., 2003). However, previous studies have frequently included ADI-R items that are not included in the diagnostic algorithm. These studies are quite useful for specifying constructs frequently associated with autism, but do not directly evaluate psychometric issues relevant to the use of the ADI-R as a diagnostic instrument. The present authors are only aware of two studies that have directly examined the factor structure of the ADI-R items or sub-scales used in the diagnostic algorithm. Lecavalier and colleagues (2006) examined individual items and reported support for a three factor model that closely resembled the DSM-IV-TR diagnostic symptom domains with the exception that non-verbal communication items loaded with social interaction items. In contrast, van Lang and colleagues (2006) examined ADI-R sub-scales using confirmatory factor analytic methods. They reported that models based upon the subscales did not fit well and that an alternative three factor model provided the best fit. This model posited a peer interaction and imaginative play factor separate from other indicators of social interaction and communication as well as a stereotyped behavior/restricted interests factor that included the stereotyped language subscale.

These two studies have some methodological differences and limitations. First, the study by Lecavalier and colleagues examined all 36 items in a modest sample of 226 people. This modest sample size for factor analytic work, coupled with the lower reliability and difficulties with examining item data with limited range, make drawing conclusions from these analyses tenuous. Second, the study by van Lang and colleagues did not directly evaluate the ADI-R algorithm. The algorithm states specifically which ratings (current, ever, or most abnormal manifestation between the ages of 4 and 5) are used to generate diagnoses. Third, these studies did not attempt internal or external replication of the purported structure. Fourth, previous analyses did not examine the stability of factor structure across age. Examination of the stability of the factor structure across age is particularly important because the instrument is frequently used in both young children for whom a diagnosis has not yet been established and for older children and adolescents who are participating in research to clarify the autism phenotype (Hus, Pickles, Cook, Risi, & Lord, 2007).

The present study examined the factor structure of the ADI-R diagnostic algorithm subscales in a large sample of individuals using both exploratory and confirmatory factor analytic methods. Analyses were conducted on two sub-samples to provide a preliminary examination of factor structure replicability. The present study also evaluated the stability of ADI-R factor structure across two age groups (ages 2-6 and 7 and older). The latter aim is important because conceptualizations of symptoms change over time and caregiver's memory for symptoms at an earlier point in life may be greatly influenced by the current age and functioning of the person being evaluated.

Method

Participants

Data for the present study were obtained from a publicly available database provided by the Autism Genetic Resource Exchange (AGRE) program (Geschwind et al., 2001). The Autism Genetic Resource Exchange preferentially selects for multiple-incidence autism families. As part of the initial assessment for this program, the ADI-R is administered to confirm the autism diagnosis. In the initial database, the ADI-R was conducted with caregivers from 914 families yielding 1950 total participants. However, non-verbal cases were excluded from the sample (N=673) because these cases do not have observations for two ADI-R sub-scales.[1] Also, when more than two ADI-Rs were completed per family (N=107, i.e. 3 or more siblings suspected of having an autism spectrum disorder), interview data from only two of the family members suspected of having an autism spectrum disorder were randomly chosen for inclusion. After excluding cases according to the procedures described above, the study sample consisted of 415 caregivers who completed ADI-R interviews for two family (N=830) members suspected of having an autism spectrum disorder and 340 families for whom only one ADI-R interview was available. Consequently, the final dataset contained 1170 total participants (age range 2-46). The institutional review board at John Carroll University approved the procedures of this study.

Measures

The ADI-R (Lord, Rutter, & LeCouteur, 1994) is a standardized, semi-structured clinical interview for caregivers of children and adults. In all cases, ADI-R interviews were completed by individuals who were trained by a certified doctoral level practitioner and completed ongoing inter-rater reliability checks in order to maintain acceptable levels of reliability. The first author of the ADI-R served as a consultant for administration and scoring of the instrument. All raters were trained and checked to ensure good inter-rater reliability. For the present study, items were scored as specified by the manual and high scores (3's) were not transformed because such a transformation reduces the variability of scores, decreases the information provided by the item, and lowers the reliability of the item and any sub-scale scores to which the item contributes.[2] The twelve ADI-R subscales were computed using the scoring algorithm specified in the manual (Rutter, Le Couteur, & Lord, 2003). Subscales include four social interaction sub-scales: (S1) failure to use non-verbal behavior to regulate social interaction, (S2) failure to develop peer relationships, (S3) lack of shared enjoyment, and (S4) lack of socioemotional reciprocity; four communication sub-scales: (C1) delay in spoken language and failure to compensate through gesture, (C2) relative failure to initiate or sustain conversational interchange, (C3) stereotyped, repetitive, or idiosyncratic speech, and (C4) lack of spontaneous make-believe or social imitative play; and four restricted, repetitive, or stereotyped patterns of behavior indictors: (R1) encompassing preoccupation or circumscribed interests, (R2) apparently compulsive adherence to nonfunctional routines or rituals, (R3) stereotyped and repetitive motor mannerisms, and (R4) preoccupation with parts of objects or nonfunctional elements of material. Each ADI-R rater needed to achieve inter-rater reliability K>/=.90 on the diagnostic algorithm items and other ADI-R items across three assessments prior to being certified. Ongoing reliability checks were conducted to ensure that inter-rater reliability remained high (K>/=.90). The internal consistent reliability of ADI-R sub-scales generally ranged from fair to excellent (α=.37-.78), particularly considering the very brief, two to five item, nature of these scales. These reliability estimates further underscore the importance of aggregating items when examining the factor structure of the instrument, as individual items contain large amounts of error variance.

Analyses

The twelve indicators were submitted to factor analyses to provide the most realistic appraisal of the factor structure of the instrument as it is used clinically. Prior to evaluating ADI-R factor structure, the sample was split into two sub-samples (N=558 and 612). Sub-sample creation was guided by the desire to balance between cross-validation and the management of nesting within families. Sub-samples were created by randomly selecting one family member from each family. When only one ADI-R interview was available but another family member had been evaluated and excluded for the reasons described above, participants were randomly assigned to sub-sample 1 or sub-sample 2. All other cases in which only one family member had a completed ADI-R were assigned to sub-sample 2. This resulted in a slightly larger sample size for sub-sample 2. The sub-sample assignment procedure increases differences between sub-samples in order to provide a more stringent test of factor structure replicability. The assignment procedure also avoids having multiple children from the same family in the same sub-sample.

Analyses examining sub-sample differences were computed assuming independent groups. While this assumption is clearly not met due to the inclusion of family members in separate sub-samples, this approach provides a conservative, but more realistic, test of sub-sample differences. Comparisons between sub-samples were performed for age at the time of evaluation, gender, ADI-R diagnosis, Autism Diagnostic Observation Scale (ADOS) diagnosis, estimated non-verbal IQ from the Raven's progressive matrices (Raven, 1956), estimated verbal IQ from the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1997), ADI-R subscale scores, and ADI-R full scale scores.

To examine ADI-R factor structure, a principal components analysis (PCA) was computed on the first sub-sample. Horn's parallel analysis with Glorfeld's modification (GHPA; Glorfeld, 1995) was used to determine the number of components to retain by comparing the results of GHPA to the results of PCA. Then, principal axis factoring with promax rotation was used to examine the composition of the factors retained. Oblique rotation was used since psychopathology constructs are likely to be substantially correlated (Fabrigar, Wegener, MacCallum, & Strahan, 1999).

The replicability of the resulting factor structure from analyses in the first sub-sample was examined in two ways. First, principal axis factoring with promax rotation was recomputed in the second sub-sample. Tucker coefficients compared factor structures of the first and second sub-samples (Tucker, 1951). These coefficients were computed for 2, 3, and 4 factor solutions to gauge stability in relation to the number of factors retained. Second, a series of confirmatory factor analyses (CFAs) were computed in the second sub-sample comparing the model identified in the first sub-sample with several other plausible models. The models evaluated using CFA were: a one-factor model with all twelve indicators loading on a single factor (model 1), a two-factor model

(model 2a) based upon the results of exploratory analyses in the first sub-sample with all social interaction items (S1 thru S4) and three of the four communication items (C1, C2, and C4) on one factor and all remaining items loading on a second factor (C3 and R1 thru R4), a two-factor model (model 2b) with all social and communication items loading on the first factor (S1 thru S4 and C1 thru C4) and all repetitive behavior and restricted interests items (R1 to R4) loading on a second factor, a three factor model (model 3a) representing the DSM criteria organization, and a three-factor model (model 3b) based upon the previously described CFA study (van Lang et al., 2006). The single factor model was computed to examine the possibility that a solution representing autism symptom severity provides the most parsimonious fit. Model 3b posits a peer interaction and imaginative play factor (S2 and C4), a combined social interaction and communication factor (S1, S3, S4, C1, and C2), and a stereotyped behavior/restricted interests factor (C3 and R1 thru R4). A four-factor model (model 4) was also computed. Model 4 is similar to model 3a except that restricted interests and repetitive behavior are specified as separate factors, based upon previous research supporting this distinction (Cuccaro et al., 2003; South, Ozonoff, & McMahon, 2005).

Amos 5.0 (Arbuckle, 2003) was used to compute all confirmatory factor analyses. To examine the impact of the ordinal nature of the data on the results, analyses were conducted using standard product-moment correlations and covariance matrices as well as polychoric correlations and asymptotic covariance matrices. Results indicated very similar structure for all analyses, therefore only analyses based on product-moment and standard covariance matrices are presented. The present study used multiple indices to evaluate model fit based upon recommendations in the literature (Bollen, 1989; Hu &

Bentler, 1995): Chi-square (X^2), Chi-square/degrees of freedom (X^2/df, <3.0), the Comparative Fit Index (CFI, >.90), the Tucker-Lewis Index (TLI, >.90), the root mean square error of approximation (RMSEA, <.08), the Parsimony Goodness of Fit Index (PGFI), Parsimony Comparative Fit Index (PCFI), and the Bayesian Information Criterion (BIC). The latter three indices do not have interpretive guidelines; however, they can be useful for comparing non-nested models.

Analyses examining factorial invariance across age groups were performed using two age groupings (ages 2-6 and 7 and above). First, these analyses compared the unconstrained model to a model with factor loadings constrained across age groups, then the model with factor loadings constrained was compared to a model with both factor loadings and factor covariances constrained (Byrne, Shavelson, & Muthen, 1989; Byrne, 2001). The chi-square difference test and changes in CFI <-.01 were used to evaluate factorial invariance at each step (Cheung & Rensvold, 2002).

Results

Table 1 presents demographic, diagnostic, and cognitive test information separately by sub-sample and for the total sample. There were no statistically significant differences between sub-samples for the distribution of age, gender, ADI-R diagnoses, ADOS diagnoses, or estimated IQ from the Raven's or PPVT. Table 2 presents score ranges, means, and standard deviations for ADI-R subscales and full scales, separately by sub-sample and for the total sample. Results indicated no significant differences between sub-samples for any ADI-R full or sub-scale. The difference between sub-samples for R2: "apparently compulsive adherence to non-functional routines or rituals" approached significance.

[place table 1 about here]

[place table 2 about here]

Generating and comparing models. Principal components analysis in the first sub-sample indicated two large components (4.72 and 1.50) with the remaining components being substantially smaller. Comparison with the results of GHPA indicated two components. The observed eigenvalue for the third principal component (.986) was substantially smaller than the 95% confidence interval of the third eigenvalues generated using GHPA (95% CI=1.11-1.13). Principal axis factoring with promax rotation showed stronger congruence between the first and second sub-samples for the two-factor solution (.97) relative to the three- (.95) or four-factor solutions (.79). The two-factor solution had high loadings (>.60) from S1 thru S4 and C1, C2, and C4 on the first factor. This factor appeared to measure general social interaction and communication difficulties. The second factor had high loadings (>.43) from C3 and R1 thru R2. R3 and R4 had minor loadings on this factor (.21 and .24). This factor assesses stereotyped/repetitive behaviors and restricted interests.

Table 3 presents fit indices from CFAs for each of the models evaluated in the second sub-sample. The data met the assumptions of maximum likelihood estimation. Comparisons between models on absolute fit indices indicated that only two models, 2a and 3b, had acceptable absolute fit. None of the models met the stringent X^2/df, <3.0, however, both of the above models had RMSEA values <.08. Relative fit indices indicated that models 2a and 3b showed the best fit. For the parsimony fit indices, model 2a had the best fit, followed by model 1 and model 2b. Three and four factor models had noticeably weaker parsimony fit as a result of penalization for greater model complexity. The BIC suggested roughly equivalent fit for model 2a and 3b.

[place table 3 about here]

Measurement invariance and age-differences in factor structure. CFAs examining factorial invariance across age were performed using models 2a and 3b, because these models had the greatest support in model comparison analyses. For both models 2a and 3b, constraining factor loadings did not significantly decrease model fit (model 2a $\Delta X^2(10)=6.18$, p=.800, ΔCFI=.001; model 3b $\Delta X^2(9)=6.00$, p=.740, ΔCFI=.002), indicating no significant age-differences in factor loadings. Constraining factor covariances resulted in significantly decreased model fit for both models 2a and 3b, however only trivial decreases in CFI were observed (model 2a $\Delta X^2(3)=10.63$, p=.014, ΔCFI=-.001; model 3b $\Delta X^2(6)=16.79$, p=.010, ΔCFI=-.003), indicating minimal changes in factor covariances.

Discussion

The present data indicate that the factor structure of the ADI-R subscales used in generating clinical diagnoses of autism is likely different than the currently published and recommended structure. Instead, this study found support for a two-factor structure that separates stereotyped behaviors from impairments in social interaction and communication. Specifically, stereotyped language and restricted, repetitive, and stereotyped behavior loaded together on one factor and impairments in social interaction and communication loaded together on a second factor. Some support was also garnered for further splitting the latter factor into general social and communicative impairments and impairments in peer relationships and imaginative play. However, the latter factor contains only two items, making this a relatively minor factor. Future revisions of the ADI-R may wish to bolster this additional factor by including a larger number of indicators evaluating peer interaction and play. Factor structures fit equally well in the younger and older age groups, supporting the use of the ADI-R for evaluation of children and adolescents in both age ranges. The relatively minor changes in factor covariances indicate that the constructs measured change little across development.

A few caveats to these general conclusions are worth noting. First, even the best fitting models did not fully satisfy absolute fit criteria. This suggests that future revisions of the ADI-R diagnostic algorithm would benefit from a more careful examination of item and subscale covariances, both within and across sub-domains. Second, the present replication occurred within the same sample with a specific recruitment strategy heavily loaded toward individuals who meet diagnostic criteria for autism. Future studies examining measurement invariance, factor structure replicability, and factor structure stability should also be performed in more heterogeneous samples with a greater proportion of individuals who do not meet diagnostic criteria for pervasive developmental disorder. It is possible that studies using more heterogeneous populations will find more similar structure to the DSM organization.

The present findings have direct relevance for the structure of DSM-IV-TR symptom criteria because the ADI-R attempts to measure these symptoms. The present data indicate that autism symptom domains may need to be restructured to more accurately reflect the strong relationship between social and communication impairments and their separation from stereotyped and repetitive behaviors. Although less parsimonious, specifying separate diagnostic criteria examining play and peer relationships from other social and communicative impairments has the potential to provide incremental validity if additional criteria evaluating this construct are included in future diagnostic revisions. Analyses examining the age-invariance of two- and three-factor structures support the continuity of diagnostic criteria across age. However, this will need to be re-evaluated if additional peer relationship/social approach items are added to the diagnostic criteria, as these types of symptoms and their relationship with other autism symptoms may change across age.

Although not the focus of this study, the present results did not support a distinction between circumscribed interests or unusual preoccupations and stereotyped or repetitive behavior. This may be due to the relatively small number of indicators of these constructs in the present study. Future studies should examine a larger number of these symptoms to better delineate whether these items actually measure separate constructs. Some data does suggest that these are different constructs and may occur at different rates in subgroups of individuals with autism (Lam & Aman, in press; South et al., 2005). It may be that repetitive/stereotyped behaviors are best represented by a hierarchical structure with a general factor that is most useful for diagnostic distinctions (typical development vs. PDD spectrum) and sub-factors that are more useful for refining phenotypic heterogeneity within the autism spectrum. Regardless, the present data are consistent with previous findings (van Lang et al., 2006) indicating that the ADI-R, as it is currently framed, should include stereotyped language within the restricted, repetitive, stereotyped symptom cluster and not the communication cluster.

The present findings concerning the structure of autism symptoms have relevance for future studies examining autism endophenotypes submitted to genetic analyses.

Recent genetics studies have focused on reducing heterogeneity by separating individuals according to presumed autism subtype differences, such as history of regression (Molloy, Keddache, & Martin, 2005) or broad symptom groups, such as delayed speech and language (Spence et al., 2006). These studies may wish to define endophenotypes based upon the factor structures supported by the present study rather than the typical algorithmic scoring of the ADI-R or individual items, which are likely to have weaker reliability than factor scores. For example, using the present data, unit weighting of the items loading highly on each of the three factors from model 3b produced measures with adequate to excellent reliability (ICC 3,7-17 = .67-.89; Shrout & Fleiss, 1979). However, the estimated reliability of individual items was quite poor (ICC 3,1 = .14-.37). Future studies examining the latent taxonomic structure of the autism spectrum may also benefit from using empirically-supported ADI-R factors, rather than the clusters of items currently proposed.

Lastly, the present data are limited by the sample employed. Although this study is the first to comprehensively examine the factor structure of the ADI-R sub-scales in a large sample of individuals, the sample almost exclusively recruited families with multiple members carrying autism or autism spectrum diagnoses. This did not cause any statistical estimation problems, as the data generally had good range.[4] However, most of the participants included met diagnostic criteria for autism. Sampling of the lower end of the autistic spectrum was not representative of the epidemiology of pervasive developmental disorder diagnoses. Additionally, the exclusion of non-verbal participants likely further limited the representativeness of the sample to the larger autism spectrum. However, additional analyses using imputed data for non-verbal subjects suggested little change in factor structure. Future confirmation of the present results in a more heterogenous sample of individuals with and without autism spectrum diagnoses is needed.

References

Arbuckle, J. L. (2003). Amos (Version 5.0). Chicago: Smallwaters.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Byrne, B. M., Shavelson, R. J. & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Byrne, B. N. (2001). *Structural equation modeling with AMOS*. Rahwah, NJ: Lawrence Erlbaum Associates.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling, 9*, 233-255.

Constantino, J. N., Gruber, C. P., Davis, S., Hayes, S., Passanante, N. & Przybeck, T. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry, 45*, 719-726.

Cuccaro, M. L., Shao, Y., Grubber, J., Slifer, M., Wolpert, C. M., Donnelly, S. L., Abramson, R. K., Ravan, S. A., Wright, H. H., DeLong, G. R. & Pericak-Vance, M. A. (2003). Factor analysis of restricted and repetitive behaviors in autism using the Autism Diagnostic Interview-R. *Child Psychiatry and Human Development, 34*, 3-17.

Dunn, L. M. & Dunn, L. M. (1997). *Examiner's manual for the Peabody Picture Vocabulary Test--Third Edition*. Circle Pines, MN: American Guidance Service.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Geschwind, D. H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P., Ducat, L. & Spence, S. (2001). The autism genetic resource exchange: A resource for the study of autism and related neuropsychiatric conditions. *American Journal of Human Genetics, 69*, 463-466.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55*, 377-393.

Hu, L. & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications.

Hus, V., Pickles, A., Cook, E. H., Risi, S. & Lord, C. (2007). Using the Autism Diagnostic Interview - Revised to increase the phenotypic homogeneity in genetic studies of autism. *Biological Psychiatry, 61*, 438-448.

Lam, K. S. L. & Aman, M. G. (in press). The Repetitive Behavior Scale-Revised: Independent validation in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*.

Lecavalier, L., Aman, M. G., Scahill, L., McDougle, C. J., McCracken, J. T., Vitiello, B., Tierney, E., Arnold, L. E., Ghuman, J. K., Loftin, R. L., Cronin, P., Koenig, K., Posey, D. J., Martin, A., Hollway, J., Lee, L. S. & Kau, A. S. M. (2006). Validity of the Autism Diagnostic Interview-Revised. *American Journal of Mental Retardation, 111*, 199-215.

Lord, C., Rutter, M. & LeCouteur, A. (1994). ADI-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*, 569-685.

Molloy, C. A., Keddache, M. & Martin, L. J. (2005). Evidence for linkage on 21q and 7q in a subset of autism characterized by developmental regression. *Molecular Psychiatry, 10*, 741-746.

Raven, J. C. (1956). *Coloured progressive matices*. Los Angeles, CA: Western Psychological Services.

Rutter, M., Le Couteur, A. & Lord, C. (2003). *ADI-R: Autism Diagnostic Interview-Revised*. Los Angeles, CA: Western Psychological Services.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

South, M., Ozonoff, S. & McMahon, W. M. (2005). Repetitive behavior profiles in asperger syndrome and high-functioning autism. *Journal of Autism and Developmental Disorders, 35*, 145-158.

Spence, S. J., Cantor, R. M., Chung, L., Kim, S., Geschwind, D. H. & Alarcon, M. (2006). Stratification based on language-related endophenotypes in autism: Attempt to replicate a reported linkage. *American Journal of Medical Genetics, 141B*, 591-598.

Szatmari, P., Merette, C., Bryson, S. E., Thivierge, J., Roy, M., Cayer, M. & Maziade, M. (2002). Quantifying dimensions in autism: a factor-analytic study. *Journal of the American Academy of Child and Adolescent Psychiatry, 41*, 467-474.

Tadevosyan-Leyfer, O., Dowd, M., Mankoski, R., Winklosky, B., Putnam, S., McGrath, L., Tager-Flusberg, H. & Folstein, S. E. (2003). A principal components analysis of the Autism Diagnostic Interview-Revised. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*, 864-872.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies (Personnel Research Section Report No. 984)*. Washington, DC: Department of the Army.

van Lang, N. D. J., Boomsma, A., Sytema, S., de Bildt, A. A., Kraijer, D. W., Ketelaars, C. & Minderaa, R. B. (2006). Structural equation analysis of a hypothesized symptom model in the autism spectrum. *Journal of Child Psychology and Psychiatry, 47*, 37-44.

Author Note

Thomas W. Frazier, Section of Behavioral Medicine, The Cleveland Clinic, Cleveland, OH, USA; Eric A. Youngstrom, Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; Cynthia S. Kubu, Section of Neuropsychology, The Cleveland Clinic, Cleveland, OH, USA; Leslie Sinclair, Center for Autism, The Cleveland Clinic, Cleveland, OH, USA; Ali Rezai, Center for Neurological Restoration, The Cleveland Clinic, Cleveland, OH, USA.

Correspondence concerning this article should be addressed to the first author using the following email address: fraziet2@ccf.org or by mail at: Children's Hospital for Rehabilitation, Cleveland Clinic-Shaker Campus CR11, 2801 Martin Luther King Jr. Drive, Cleveland, OH 44104.

We gratefully acknowledge the resources provided by the Autism Genetic Resource Exchange (AGRE) Consortium and the participating AGRE families. The Autism Genetic Resource Exchange is a program of Cure Autism Now and is supported, in part, by grant MH64547 from the National Institute of Mental Health to Daniel H. Geschwind (PI).

Footnotes

[1] Non-verbal participants were excluded from analyses because these individuals do not have observations on the verbal communication items (C2 and C3). However, in the interests of examining the effects of this exclusion on findings, we imputed the most severe ratings for these individuals on these items and re-computed exploratory factor analyses. Results indicated a highly similar pattern of findings, with the exception that only two sub-scales (C3 and R1) showed significant loadings on the second factor and other repetitive behavior sub-scales loaded more heavily on the first factor. Thus, inclusion of these participants tends to shift the interpretation of the first factor from a social communication factor to a less specific autism severity factor.

[2] Preliminary analyses indicated similar factor structure transforming and not transforming “3” codes to “2” codes. Therefore we presented analyses using untransformed “3” codes.

[3] Analyses were re-computed after applying transformations to sub-scales with limited range to attempt to normalize the distributions for these variables. Results of analyses for transformed data were highly consistent with raw data.

Table 1. Demographic, diagnostic, and cognitive test data for each sub-sample and the total sample.

	Sub-Sample 1	Sub-Sample 2	Total Sample	S1 vs. S2
	M (*sd*)	*M* (*sd*)	*M* (*sd*)	*t*,X^2 (df, *p*)
N	558	612	1170	
Age	8.94 (4.84)	9.05 (4.92)	9.00 (4.88)	-.38 (1168, .701)
Gender (% male)	78.0	76.6	77.3%	.29 (1, .590)
ADI-R dx. N (%)				.42 (2, .810)
Autism	82.6	83.0	82.8	
PDD NOS	15.2	15.4	15.3	
No Dx.	2.2	1.6	1.9	
ADOS dx. %				2.90 (2, .234)
Autism	72.7	74.3	73.5	
PDD NOS	18.6	19.8	19.3	
No Dx.	8.7	5.9	7.2	
Raven's IQ	101.54 (21.97)	100.80 (22.22)	101.16 (22.09)	0.45 (720, .650)
PPVT IQ	89.17 (22.60)	86.20 (25.76)	87.63 (24.32)	1.65 (730, .099)

Note. PDD NOS= Pervasive Developmental Disorder - Not Otherwise Specified. No Dx. = did not meet diagnostic cutoffs for an autism spectrum diagnosis. Autism Diagnostic Observation Scale (ADOS) data, Raven's Matrices IQ, and Peabody Picture Vocabulary Test (PPVT) IQ scores were only available for a subsets of participants (ADOS N=998; Raven's N=722, range 45-140; PPVT N=732, range 40-142).

Table 2. Score ranges, means, and standard deviations for ADI-R subscales and full scales separately by sub-sample and for the total sample.

		Sub-sample 1 *M* (*sd*)	Sub-sample 2 *M* (*sd*)	Total Sample M (sd, range)	S1 vs. S2 *t* (*p*)
Social Interaction Scale		19.69 (7.11)	19.92 (7.36)	19.81 (7.24, 0-30)	-0.53 (.596)
	S1: non-verbal	3.66 (1.87)	3.71 (1.89)	3.69 (1.88, 0-6)	-0.42 (.670)
	S2: peer relations	5.73 (2.22)	5.77 (2.27)	5.75 (2.24, 0-8)	-0.34 (.731)
	S3: shared enjoyment	4.38 (1.83)	4.34 (1.88)	4.36 (1.85, 0-6)	0.35 (.724)
	S4: socio-emotional reciprocity	5.92 (2.61)	6.10 (2.61)	6.01 (2.61, 0-10)	-1.12 (.261)
Communication Scale		15.98 (5.21)	16.17 (5.14)	16.08 (5.17, 0-26)	-0.61 (.539)
	C1: gesture	4.73 (2.62)	4.86 (2.67)	4.80 (2.65, 0-8)	-0.86 (.389)
	C2: sustain conversation	2.83 (1.34)	2.80 (1.34)	2.81 (1.34, 0-4)	0.02 (.983)
	C3: stereotyped speech	3.87 (2.03)	3.96 (1.94)	3.91 (1.98, 0-8)	0.46 (.649)
	C4: imitative play	4.56 (1.61)	4.56 (1.59)	4.56 (1.60, 0-6)	-0.78 (.437)
Repetitive Behavior Scale		5.85 (2.69)	6.00 (2.74)	5.93 (2.72, 0-12)	-0.94 (.350)
	R1: circumscribed interests	1.49 (1.26)	1.48 (1.29)	1.49 (1.28, 0-4)	0.10 (.923)
	R2: compulsive routines	1.48 (1.40)	1.63 (1.40)	1.56 (1.40, 0-4)	-1.84 (.067)
	R3: stereotyped motor mannerisms	1.32 (0.86)	1.35 (0.83)	1.34 (0.85, 0-2)	0.61 (.540)
	R4: preoccupation with objects	1.56 (0.69)	1.53 (0.71)	1.54 (0.70, 0-2)	0.60 (.548)

Note. df=1168 for all analyses.

Table 3. Fit indices from confirmatory factor analyses for each model.

Model	X^2	DF	X^2/df	RMSEA	CFI	TLI	PGFI	PCFI	BIC
One Factor (1)	324.82	54	6.02	0.09	0.89	0.87	0.63	0.73	478.82
Two Factor (2a)	216.24	53	4.08	*0.07*	*0.94*	*0.92*	*0.64*	*0.75*	*376.66*
Two Factor (2b)	272.03	53	5.13	0.08	0.91	0.89	0.63	0.73	432.45
Three Factor (3a)	265.17	51	5.20	0.08	0.92	0.89	0.61	0.71	438.42
Three Factor (3b)	203.71	51	*3.99*	*0.07*	*0.94*	*0.92*	0.62	0.73	376.96
Four Factor (4)	240.57	48	5.01	0.08	0.92	0.90	0.58	0.67	433.07

Note. Italics denotes best-fitting model, separately for each fit statistic.
Model 2a= model based on exploratory factor analytic results with combined social interaction and communication impairments and stereotyped speech sub-scale loading on the second factor with other stereotyped behavior sub-scales. Model 2b= DSM model with social interaction and communication impairments combined. Model 3a=DSM model with three separate factors. Model 3b=Same as model 2a but with peer relationships and play sub-scales as a separate third factor. Model 4a=DSM model with separation of restricted interests and stereotyped, repetitive behavior sub-scales.