Report by "sagkyndigt udvalg til bedømmelse af Helmuth Nyborgs forskningsprojekt vedrørende kønsforskelle i intelligens"

The committee was appointed by the Dean of Social Sciences, Aarhus University, by letter of December 7, 2005, and by letter of January 6, 2006.

The committee consists of

Professor Jan-Eric Gustafsson, Gothenburg University,

Professor Jens Ledet Jensen (chairman), Aarhus University, and

Professor Niels Keiding, Copenhagen University.

The committee has based its work on the 23 enclosures and the terms of reference of December 7, 2005. The terms of reference contains 14 requests and questions, which are reproduced in Appendix A and numbered I to XIV. Since the questions are interrelated the report is structured into subjects that each touch upon several questions. The committee has in its work also used *Vejledende Retningslinier for Forskningsetik i Samfundsvidenskaberne* from the Danish Social Science Research Council, November 2002 (referred to as [SSRC]).

The introductory part of the report presents a summary and conclusions. Then follow Sections 2–7 with the main text discussing different aspects of Nyborg's work. Supplementary details are given in appendices.

Aarhus March 16, 2006,

Jan-Eric Gustafsson   Jens Ledet Jensen   Niels Keiding

1 Summary and conclusions

1.1 Summary of observations

The questions asked in the frame of reference for the committee concern the following three main issues:

1. The nature and quality of the data sets
2. Documentation and access to reports and data
3. Procedures used in the analysis of data

A summary of the most important observations made by the committee is presented below.

1.1.1 The nature and quality of the data sets

When it comes to the questions about the nature and quality of the data sets analyzed by Nyborg the committee has reached the following conclusions:

- What has been referred to as two different sets of data by Nyborg (the $N = 52$ adult data set and the $N = 62$ adult data set) is for 15 variables one and the same data set, consisting of 62 persons. There is no explanation for the differences concerning the remaining 5 variables. The explanations that Nyborg has provided for the differences between the two data sets are not correct.
- The $N = 62$ adult data set consists of 31 persons from a cross sectional study and 31 persons remaining after dropout from 91 original participants in a longitudinal study. The large dropout makes the data of doubtful quality.
- The analysis of children's data analyzed for Nyborg's conference presentation in 2001 ($N = 325$) and referred to in the 2003 book chapter includes a large proportion of missing data, and repeated measurements for some of the children, the actual number of different children being 219. This has neither been adequately described, nor properly dealt with in the statistical analyses. For the 2005 article [22] Nyborg does not use these data, but data on a subset of $N = 119$ children with complete data and only one measurement. This latter set of data is reasonably well defined.

1.1.2 Documentation and access to reports and data

The committee has noted that Nyborg made preliminary results from the study public in an interview for Politiken and in the book chapter published in 2003, while also refusing access to documentation. Nyborg has previously been criticized by the "Committee for Good Scientific Practice"at Aarhus University for not agreeing with good scientific practice when refusing to make the contribution to the 2001 conference available.

The committee has made the following observations concerning documentation and access to data:

- The study design is incompletely described in the publications.
- The data sets are poorly described when it comes to repeated measurements and dropouts, and how missing data problems have been dealt with.

1.1.3 Procedures used in the analysis of data

On the basis of the review of the material the committee has identified several problems in the way that Nyborg has analyzed the data:

- An error has been noted in Nyborg's computations of the point biserial correlations (different from the one found by Nyborg himself).
- The p–value for testing the hypothesis that the point biserial correlation for the loading of sex on g is different from 0 has been computed with an incorrect formula.
- In his publications Nyborg has not described which specific choices were made in the application of the hierarchical factor analysis. In reanalyses of the data it has been possible to achieve similar results as those reported by Nyborg, but results that differ have also been found.
- The analytical procedures used by Nyborg suffer from an identification problem, which makes it impossible to estimate sex differences in the mean of the g–factor independently of sex differences in the primary factors.

1.2 Discussion and conclusions

In the assignment to evaluate Nyborg's work with the Skanderborg data and the reporting of the results the committee sees three levels of inquiry:

1. has Nyborg shown *due diligence*
   - in conducting the research
   - in publication
   - in relating with the academic community and the general public
2. has Nyborg committed *unintended mistakes*
   - indisputable technical mistakes
   - use of methods, which were well-established or at least not commonly rejected at the time of their use, but which on closer inspection show internal inconsistencies
3. has Nyborg committed *fraud* (intended mistakes), such as
   - consciously biased selection of samples
   - consciously biased selection of methods.

1.2.1 Due diligence

The first issue is due diligence in conducting the research. Here we have noticed the very serious attrition in the intended longitudinal study. This is, however, a common problem in longitudinal studies and does not in itself call for criticism, but it is a requirement that the attrition is documented. Nyborg's repeated reference to the fact that the study is still ongoing is not easily understandable for a project for which the data acquisition should have been concluded more than a decade ago.

Nyborg, as responsible investigator for the Skanderborg project, has not always taken the necessary care in the monitoring of the project activities, as is shown by the confusion over whether the analysis of adults was based on 52 or 62 persons. Nyborg apparently believed for several years that the analysis underlying a table in his conference presentation in 2001 was based on 52 persons, while our recalculations show that at least most of the results were already then based on the larger set of 62 individuals. The committee sees this as lack of due diligence.

We also ascribe to lack of due diligence the two mistakes in the computation of the so-called point biserial correlation. Both of these are elementary misreadings of a mathematical formula, and it so happens that the mistakes have little importance for the results. Nyborg discovered one of these mistakes himself. The committee finds that this point is of minor importance in the larger picture.

The relevant publications are three: the unpublished conference presentation from 2001, the contribution from 2003 to a book edited by Nyborg himself, and the journal publication from 2005. In the present context we restrict our comments to the latter, which is a formal publication in a peer-reviewed journal that commands respect in its field. We could not find sufficient information on the study design in the published information and had to rely on Nyborg's internal elaborations to understand what had been done. For example, the original design of the follow-up study is described in considerable detail, but there is no word about the massive drop-out problem and the consequent complicated composition of the unusually small sample of 62 for a study of this character. This lack of detail is not in our view satisfactory, although as already mentioned, we are here talking about a published peer-reviewed paper.

Nyborg gives unsatisfactorily few details about his concrete choices of methods in the highly complex statistical techniques of factor analysis. As we document in our main text, many other reasonable choices would have given different results, and due diligence would have required that these were adequately described, motivated and discussed. A special problem is that the central effect estimate turns out to be rather less impressive in all other choices than the one preferred by Nyborg. This would be serious if it were intentional. The committee has however no basis for claiming that Nyborg selected his version of the analysis so as to maximize the effect estimate.

Turning then to due diligence in the relation with the academic community and the general public, part of this matter has already been treated by the Committee for Good Scientific Practice of Aarhus University. That committee handled a complaint that Nyborg did not, on request, make available the basis of his 2001 conference presentation, which had formed the basis of extended media statements on these results by Nyborg. In short, that committee agreed that Nyborg should have made the results available on request, particularly in view of the extended time elapsed between the media coverage and the subsequent submission and publication of a proper journal article containing the findings. We find it inappropriate to comment on this earlier investigation but will allow ourselves to comment on what followed: Following the report, delivered in February 2004, the University did not explicitly request Nyborg to make his background material available. Requested by his department chair, Nyborg made various bits and pieces of the 2001 conference presentation available on his website over the coming months. As specified in our main text, we find this response to the public's justified interest to be too incomplete to qualify as due diligence.

1.2.2 Mistakes

Besides the misreadings of mathematical formulae discussed above and classified by us as lack of due diligence we have identified one indisputable technical mistake: when Nyborg computes significance values ('p-values') for the point biserial correlation, he uses what our recalculations has identified as the standard hypothesis test for correlation coefficients calculated from independent replications. This is unjustified. However, as far as we know, a correct calculation is not readily available in the literature. We therefore do not draw any wide-ranging conclusions from this mistake.

Nyborg's g–factor method, which is inspired by the American intelligence researcher A. Jensen, is mathematically quite involved. There is a lively current debate about the details and implications of such methods in the intelligence research literature, some of which has taken place after Nyborg did his analyses. The committee obviously cannot enter such a discussion, and it is equally obvious that we cannot blame Nyborg for not having taken note of methodological advances that were generated after his analyses were completed.

Nevertheless we want to put on record two clear mathematical facts that Nyborg does not seem to be aware of. First, there is an inherent unidentifiability in the g–factor method used by Nyborg, which makes it impossible to separately estimate sex differences in the mean of the g–factor from sex differences in the primary factors. Second, Nyborg claims that his version of the g–factor method avoids the problem that the conclusion will reflect gender bias in the composition of the test battery. However, we show in Appendix D that Nyborg's g–factor method too has this undesirable property.

1.2.3 Fraud

The prime candidates for fraud in a situation like this, disregarding blatant misreporting and conscious miscalculation, are:

- biased selection of data from the totality of data in order to achieve a desired result in a selected subsample that is unrepresentative for the population in general; and

- biased selection of methods of analysis so that the method is chosen which shows the strongest result in a desired direction.

We have found no evidence that Nyborg has acted fraudulently.

2 Adult $N = 52$ and $N = 62$ data sets

A number of questions concern the relation between the original Adult data set with $N = 52$ persons and the new Adult data set with $N = 62$ persons. Briefly, Nyborg claimed in his lecture [2], Table 4 (incorrectly termed Table 3) that the analysis of 16+ years old was based on N=52 persons. This is maintained in the book publication [3,p.209], but in the paper [22] there are 62 persons. The committee has been asked whether the submitted material allows clarification of the relation between these two sets of data, undoubtedly with a view to the representativity of the 52 persons within the larger sample of 62. In the published material [3] there is no information on the composition of the $N = 52$ data set, and in the paper [22] the $N = 62$ data set is referred to as the subset for which "WAIS and all the other test data were available". We therefore turn to other written material from Nyborg to get more information. Nyborg explains the difference by "updated and more complete data"([5]), in [7] both data sets are referred to as complete, in [10] it is said that the original data file for $N = 52$ adults were supplemented by new data to form the $N = 62$ adults, in [11] it is said that data for new individuals have been put directly into the data file making it impossible to reconstruct the original $N = 52$ data set, in [14] Nyborg says that there are data that still have not been put into the data file so that the number $N = 62$ will increase further and that persons are selected on the basis of complete data.

As explained in Appendix B there is no difference between the results in [2,Table 4, $N = 52$] for 15 of the 20 variables as compared to the similar table based on the $N = 62$ data set ([8]). Also for these 15 variables the degrees of freedom used to calculate the p-value correspond to the sample size $N = 62$ and not to the sample size $N = 52$. It is very difficult to guess an explanation as to why the two data sets $N = 52$ and $N = 62$ are identical for 15 variables and not for the remaining 5 variables, and it seems that Nyborg himself does not know the explanation. Since the $N = 52$ data set is identical to the $N = 62$ data set for 15 variables the question as to how the 52 persons are selected from the 62 persons is irrelevant. Also, since the two data sets are almost identical the calculations that Nyborg performs give almost the same results for the two data sets, for example the *point–biserial factor loading of sex* is given as 0.272 ($N = 52$, [2]) and 0.280 ($N = 62$, [6]). The reason that the p–value given by Nyborg is reduced is that the sample size $N = 52$ or $N = 62$ enters his calculation of the p–value. A minor, but slightly strange point, is that the calculation based on the $N = 62$ data set [6, Table 1] is not in complete accordance with the published numbers in [22, Table 1] (the numbers for effect size d are given in Appendix B), and the correct version is in [6] not in [22].

The committee sees no other explanation for these surprising findings than the following: Most of the data handling and probably also the basic calculations are performed by assistants, often mentioned in the correspondence but always anonymously. It seems that Nyborg does not monitor the activities of these assistants closely enough to be sure which data set is being analysed at each particular time. In particular the most obvious explanation behind the two different data sets is that the 62 persons have been analysed from the beginning, at least for most of the variables, so that there may never have been much if any analysis of a $N = 52$ data set. This borders on carelessness, but gives no indication of improper selection of a subset that particularly suits any preconceived structure in the data.

The committee concludes that the explanations given by Nyborg concerning the relation between the two data sets are not correct. The committee finds it criticizable that Nyborg as sole author on [3] does not know the true content of the data set used.

The above discussion relates to questions II, III, X, and XII.

3 The study design; the N=325 and N=119 data sets

The description of the study design in the published material made available to the committee is incomplete. It seems that the original intention was to create an age-stratified panel study, presumably with a view to study the individual development of intelligence with age. (The precise purpose of generating this panel study has not been made available to us.)

As explained in [3,p.208] and almost verbatim in [22,p.500], age-sex-stratified pseudo-random samples of children in Skanderborg were drawn from official Danish registers. Rather than letting the random selection mechanism take care of the representativity issue, each stratum was artificially balanced (somewhat similarly to the 'minimization' technique for medical studies) across 'all socio-economic and personal characteristics of the children', generating in principle a requirement for including these characteristics as covariates in a correct statistical analysis.

In addition to the panel study, some kind of cross-sectional study was carried out, but this has not been explained in the published texts.

A surprising issue is that while the age groups 8,10,12,14,16+ were recruited in 'the late 1970s' - and thus at the turn of the millennium were at least 28 years old - the study is nevertheless claimed to be still ongoing. We have not identified what this might mean.

It is indirectly clear that with the very serious drop-out rate from the panel study any initial ambition of achieving a proper follow-up study is unrealistic.

Because the published information is insufficient for a complete understanding of the study design, the committee has had to turn to other written material from Nyborg to get more information.

The figure of the study plan in [2] sets the cross sectional study at 1990 and in [15, enclosure 2] it is set at 08-12-88 to 22-02-95+. In [7] it is said that the database is continually being updated and that the cohort-sequential study is not completed. In [10] Nyborg speaks about the data being updated by assistants making it possible to make analyses with complete data for more persons. In [15] it is said that some data have not been inserted in the data files yet, and that new data are still being gathered. It is also said that phase 8, the cross sectional study, started in 1988 and is not completed yet and therefore appears as ending in 1995+.

Some of these statements are difficult to understand. If the phase 8 part is a truly cross sectional study it cannot run for a prolonged period. The $N = 119$ children data set are all from this cross sectional study and to make sense of the 8–14 years age group these data must have been collected around 1990. Also it is difficult to imagine that the data for the 31 adults from this cross sectional study have not been collected around 1990 (as already seen the $N = 62$ data set was available at the time of the calculations with the $N = 52$ data set). Thus these data have been around for a long time and it does not make much sense to say in [7] that the $N = 325$ data set has been redrawn because the $N = 119$ children data set is now available. Since the data for the two investigations ($N = 62$ adults and $N = 119$ children) most likely have been around since the early 1990s there has been ample time to insert data into data files (with the amount of data used it should be possible to write the data in roughly one day).

The composition of the $N = 325$ children data set appears from [9] and [15]. From group 0 29 children are measured in both phase 1 and 2, and 27 of these measured also in phase 3 (85 observations). From group 2 31 children are measured in phase 2 and 26 of these are measured again in phase 3 (57 observations). From group 3 30 children are measured in phase 2 and 24 of these are measured in phase 3 also (54 observations). These last measurements in phase 3 are contrary to what is written in [10] and what appears in [15, enclosure 2], but can be seen in [9]. There are two persons that have group number 4, but not mentioned anywhere (2 observations). There are 119+8 children from phase 8, of which complete measurements exist for 119 only (127 observations). This gives the $85+57+54+2+119+8 = 325$ observations. There exist complete measurements for the 119 children from phase 8 only. For 59 of the measurements 17 variables are available. For 113 of the measurements only 4 or 5 variables are available. For 24 of the measurements only 1 variable is available! The $N = 325$ measurements correspond to 219 different persons.

The massive occurrence of missing data in this N=325 data set would make any analysis strongly dependent on the exact method used for imputing the missing values. Nyborg apparently only mentions the most primitive of such methods, mean substitution, but more sophisticated imputation methods would also be questionable here. It therefore makes good sense to stop using this data set. The $N = 119$ children data set, being all children from the cross sectional study, seems a suitable study group. However, any analysis based on these data will be hampered by the fact that each age group (8,10,12, and 14 years) contains only around 30 persons.

The above discussion relates to questions I, II, V, VI, VII, and XI.

4 Adult $N = 62$ data set

In this section we consider the composition of the $N = 62$ adult data set. The only information we have from the published paper [22] is that the data set is the subset for which “WAIS and all the other test data were available”.

According to [9] and [15, enclosure 2] the $N = 62$ adult data set is composed as follows:

> 31 persons from cross sectional study in phase 8 (this seems to be all adults out of 150 persons, the remaining 119 constitutes the $N = 119$ children data set);

8 persons out of 9 persons left in phase 7 after dropout from original 30 persons in group 0;

14 persons out of 18 persons left in phase 7 after dropout from original 31 persons in group 2;

8 or 9 persons (there is a mysterious phase 4 person) out of 9 persons left in phase 7 after dropout from original 30 persons in group 3;

The data set thus consists of 31 persons from a cross sectional study and 31 persons that are left after dropout from 91 persons. This is a very problematic data set to use due to the high dropout. It is standard in investigations with dropouts that care must be taken to ensure that the dropout does not introduce bias. We have not registered any indication that Nyborg is aware of this basic problem, and the reader is never alerted to the mixed and unsatisfactory composition of the data set.

The sample size N=62 (or even N=52) is extremely small compared to usual standards for cross-sectional studies on relatively inexpensive endpoints such as these cognitive tests. The studies that Nyborg reviews in [3] have sample sizes 602, 10475, 636, 1201, 678, and 1369, respectively.

As mentioned in Section 3 above, the study design is never fully explained to the reader, who is therefore unable to capture the serious dropout problem. It is also impossible to see that "The particular selection procedure resulted in a total of 376 children and adults"in [3] refers to multiple testing of some of the children and not to 376 different persons. More precisely, it is not possible to see where in the study design the $N = 62$ adults come from . Furthermore, Nyborg does not mention in [3] that the use of the $N = 325$ data set for children involves mean substitution to such a degree as to make the results strongly questionable.

The test-batteries analyzed by Nyborg was composed of subtests from the Wechsler tests (12 subtests from the Wechsler Intelligence Scale for Children for the children, and 11 subtests from the Wechsler Adult Intelligence Scale for the adults) along with 9 additional variables. The additional tests were the Rod and Frame Test, Embedded Figures Test, Money left-right discrimination test, Mental Rotation, Tapping left hand and Tapping right hand, and Oral Fluency. Most of these tests measure different kinds of spatial abilities, for which a male performance advantage has been found in many studies. Furthermore, for the Rod and Frame Test Nyborg derived three scores (two dependence scores and one variability score). However, being derived from the same observations the relations among these variables are inflated. The two dependence scores were based on two different scorings of errors on the Rod and Frame Test, and for these a correlation of .94 was observed in the adult sample. It is not good practice to include both these variables in a factor analysis.

The committee finds it criticizable that Nyborg in [3] and [22] does not give a proper description of the study design, that he does not discuss the large dropout problem, and that he in [3] mentions neither the mean substitution nor the repeated measurements on some of the children.

We remark here that [SSRC] writes under point 5: “Forskeren bør redegøre for de anvendte metoder, materialer og analyser, således at kolleger og offentligheden i øvrigt har tilstrækkeligt grundlag for kritisk vurdering af arbejdet”.

The above discussion relates to questions I and II.

5 Public access

The committee has been asked to consider if Nyborg has followed scientific practice in what he has made public and in terms of data documentation. We first give the chronology of events.

December 2001 Nyborg presents his results at a conference

08-01-02 Nyborg makes his results public through an interview in Politiken. There is no public documentation at this time.

23-01-02 Department of Psychology chairman Jens Mammen receives [2] and is asked to keep this material confidential.

June 2003 The book [3] is published. In this book Nyborg states his results of the investigation, but does not give any documentation. In the book Nyborg refers to a paper Nyborg (2001) and a paper Nyborg (2003, submitted).

05-10-03 Pia Ankersen lodges a complaint to the Committee for Good Scientific Practice at Aarhus University for not being able to obtain Nyborg (2001) from Nyborg.

25-02-04 The Committee for Good Scientific Practice finishes a report on the complaint by Pia Ankersen. The president of the university agrees to the conclusion in the report, but no action is taken.

09-06-04 Parts of [2] relating to the $N = 52$ Adult data set are made available at Nyborg’s homepage at the request of Mammen.

10-06-04 Material is removed from homepage.

21-07-04 Official request from Mammen as to reestablishing the material at the homepage.

13-10-04 A new version of the homepage has been made. Results are given, but no data documentation.

05-12-04 At the request of Mammen the homepage is revised. In this material are now two correlation matrices, but not the raw data.

03-03-05 At the request of Mammen the data files in [8] are delivered.

08-03-05 Data protocols from the investigation are delivered from Nyborg.

August 2005 The article [22] is published.

Next we quote from [SSRC] point 5:

Forskeren skal ikke blot offentliggøre sine resultater men skal også gøre det muligt for eventuelle kritikere at efterprøve, om resultaterne er begrundet i det tilgrundliggende oplysningsmateriale. Dette bør derfor bevares i rimelig tid og gøres tilgængeligt for videnskabelige bedømmere. Hensynet til informantens anonymitet kan afskære udenforstående fra kontakt med informanter, f.eks. kan erhvervsforskning af virksomhedsinterne forhold ikke gøres til genstand for fri efterkontrol i vedkommende virksomheder. Offentliggørelse af konklusioner eller delresultater bør kun, hvis der foreligger særlige omstændigheder, ske før undersøgelsen er afsluttet og tilgængelig på den anførte måde.

Since the Committee for Good Scientific Practice at Aarhus University has considered the lack of access to Nyborg (2001) we will not comment on this aspect.

Nyborg has made public the conclusions of his investigation in two cases, i) through the interview in Politiken January 2002 and ii) through publishing [3] in June 2003. Both of these are prior to the publishing of an article describing the investigation in detail (the latter, [22], first appear in August 2005). After publishing [3] it is not possible for Pia Ankersen to obtain background material.

The Committee finds that Nyborg has not followed the rules laid out by [SSRC] saying that conclusions should not be made public, except in special circumstances, before the investigation is concluded and made available.

The above discussion relates to questions III and IV.

6 Reconstructing Nyborg's results

The committee has been asked if the results and tables produced by Nyborg can be reconstructed from the data files in [8].

The committee has spent considerable effort re-analysing Nyborg's data from [8]. We report this in the following way: below we first summarize Nyborg's own statements on the methods to be applied, and then give a short summary of our findings, with technical details in Appendix C. Finally we conclude that although we have identified one specific minor calculating mistake (much akin to another miscalculation discovered and corrected by Nyborg himself), and one misuse of a significance calculation, we generally get similar results as Nyborg, using a broad set of related methods.

In [22] Nyborg describes his analysis by "to calculate the point-biserial correlation (r_{pbs}), indicating the extent to which sex, as a dichotomous variable loads on the metric sex differences), and then to insert the (twenty) r_{pbs} in the subtest intercorrelation matrix, and factor analyse them", "The present study used an HFA analysis with the Schmid-Leiman (SL) transformation", and "The HFA/SL analysis permitted extraction of one second-order g factor and six first-order group factors". In [3] the hierarchical factor analysis (HFA) is described as "One thus begins the analytic process by first identifying the primary or group factors using PF (or in some special cases PC) analysis, and forcing an oblique rotation of factor axes to determine their correlations. Another step is to derive the second-order (or third order in the case of a large and varied test battery) g from correlations among the group factors at the primary level"and "The HFA analytic solution can be optimized by including a Schmid-Leiman (SL) transformation (Schmid & Leiman 1957). This procedure orthogonalizes all factors between and within all levels in the hierarchy". Nyborg does not refer to a computer package for doing the calculations.

The analysis thus has the following steps

i) Calculate the point biserial correlation r_{pbs} and insert this as an extra column in the correlation matrix.

ii) Perform a factor analysis to extract six primary factors.

iii) Perform an oblique rotation of the primary factors.

iv) Perform a factor analysis on the correlation matrix corresponding to the oblique factors and extract one second order factor.

v) Perform a Schmid-Leiman orthogonalization of the factors.

Concerning i) Nyborg uses a formula from Jensen (1998) written there as

$$d/2\sqrt{(d^2/4)+1}.$$

Nyborg misreads this and uses

$$r_{ny} = (d/2)\sqrt{(d^2/4)+1} \quad \text{instead of} \quad r_{pbs} = (d/2)/\sqrt{(d^2/4)+1}. \tag{1}$$

see Appendix C. This error is not corrected in any of the calculations by Nyborg, but this has no practical consequences. In [2], [3], and [4], Nyborg makes a further error and uses the formula

$$r_{old} = (d/2)\sqrt{(d^2/d)+1} \tag{2}$$

The two errors together have only minor influence on the results obtained by Nyborg. When using r_{ny} our calculation of effect size d and point biserial correlation is in accordance with the numbers reported in [5, Table 2] and [6, Table 1]. As previously mentioned there are some very small disagreements with the numbers given in [22, Table 1].

Concerning ii) there are a number of possibilities. In the early days of factor analysis the method used was that of *principal factor analysis*. The first step in this analysis is to choose preliminary communalities and two methods for doing this are often used. From the late 1960s onwards it has been possible instead to use a maximum likelihood approach, and this latter method also gives a possibility of judging the number of factors to include in the factor analysis. Nyborg does not state explicitly what he has done, but judging from the description in [3] he has used one of the principal factor analysis methods. In the example given in Jensen (1998, p. 80-81) it seems that maximum likelihood factor analysis is used.

The oblique rotation under point iii) is not specified. This is a problem since there are many possibilities around. One possibility is to use what is known as *oblimin*. For example it seems that Jensen (1998, p. 80-81) is using this (with a certain parameter set to zero). Another possibility is *promax*, which has a parameter m that is often set to $m = 4$.

In order to reconstruct Nyborg's findings we have tried a number of possibilities. The details can be seen in Appendix C. A summary in terms of *point-biserial factor loading on sex* for the $N = 62$ adult data set is as follows

method	Nyborg	MLE,obl	MLE,pro	PF1,obl	PF1,pro	PF2,pro
loading	0.280	0.140	0.233	0.210	0.277	0.235
distance		0.073	0.053	0.047	0.037	0.050

The entry *Nyborg* is from [5, Table 2, and 6, Table 1], *MLE* is maximum likelihood factor analysis, *PF1* and *PF2* are two methods of principal factor analysis, *obl* is using the oblimin ($\gamma = 0$) oblique rotation and *pro* is using the promax ($m = 4$) rotation. The last row gives a distance measure between Nyborg's g–loadings [5, Table 2] and the g–loadings obtained by the different methods. The similar table for the $N = 119$ children data set is

method	Nyborg	MLE,obl	MLE,pro	PF1,obl	PF1,pro	PF2,pro
loading	0.228	0.185	0.195	0.190	0.220	0.203
distance		0.114	0.080	0.085	0.071	0.070

These two tables indicate that Nyborg has used principal factor analysis (in accordance with the indirect statements in [3]) and the promax rotation. However, it is still puzzling that we do not get complete agreement. To investigate this further we have tried to guess the factor technique used independently of the rotation used. This is actually possible to investigate since the g–loadings obtained by the hierarchical analysis always ends up being a linear combination of the primary factor loadings. We can therefore try the different primary factors and ask if the g–loadings reported by Nyborg ([5]) can be written as a linear combination of the loadings of these factors. It turns out that this is not the case for the maximum likelihood factors nor for the two choices of principal factors, but if we use another method, namely factors from a principal component analysis, we get that Nyborg's g–loadings can be written exactly as a linear combination of the primary factor loadings. (There is a small possiblity that the PF2 method has been used for the $N = 62$ adult data, but for the $N = 119$ children data set the only possibility is principal component.) Documentation is given in Appendix C. That Nyborg uses principal component analysis for extracting the primary factors therefore gives extra meaning to the statement [3, p. 197] "One thus begins the analytic process by first identifying the primary or group factors using PF (or in some special cases PC) analysis".

We have once again tried the oblimin and promax rotations and either principal component or principal factors for extracting the second order factor. In none of these cases do we get Nyborg's numbers. A summary corresponding to the tables above is given below, first for the $N = 62$ adult data set and next for the $N = 119$ children data set.

method	Nyborg	PC,obl	PF1,obl	PF2,obl	PC,pro	PF1,pro	PF2,pro
loading	0.280	0.301	0.204	0.187	0.361	0.269	0.254
distance		0.101	0.069	0.089	0.111	0.046	0.062

method	Nyborg	PC,obl	PF1,obl	PF2,obl	PC,pro	PF1,pro	PF2,pro
loading	0.228	0.252	0.221	0.190	0.270	0.268	0.234
distance		0.099	0.075	0.082	0.093	0.085	0.062

Here PC signifies principal component analysis for extracting the second order factor, PF1 and PF2 are principal factor analysis for extracting the second order factor, obl is oblimin ($\gamma = 0$) and pro is promax ($m = 4$).

Thus, in spite of the fact that we have been able to figure out that Nyborg has used principal component analysis for finding the primary factors, we have not been able to guess what calculations Nyborg has done following the first step. We are able to make choices that give results that resemble those given by Nyborg [5 and 6].

Finally, in this section we comment on the possibility of reconstructing the p–value that Nyborg attaches to the *point-biserial factor loading of sex*. We have not been able to find a place where Nyborg gives information on the calculation of the p–value. Here is a small table with p–values given by Nyborg.

reference	loading	p–value
[3, p. 209] ($N = 52$)	0.272	0.026
[22, p. 502] ($N = 62$)	0.274	0.016
[6, p.5] ($N = 62$)	0.280	0.014
[6, p. 2] ($N = 119$)	0.231	0.006

In all of these cases we get exactly the same value by referring $t = r\sqrt{N-2}/\sqrt{1-r^2}$, with r the factor loading, to a t–distribution with $N - 2$ degrees of freedom. Thus it seems that Nyborg is using the classical test for a correlation being zero. However, this can not be used here. The classical test is for the situation where r is the empirical correlation from N independent observations, but the factor loading obtained through the hierarchical factor analysis is not an empirical correlation. It seems that a correct calculation of the p–value is not readily available in the literature. As a side remark we note that Nyborg himself in [2] highlights loadings greater than 0.300.

The committee has not, from the information provided by Nyborg, been able to reproduce Nyborg's calculations exactly, but we do obtain, qualitatively, the same numbers. During this process we have noted an error in the calculation of the point biserial correlation and the use of a unjustified p–value.

The above discussion relates to questions VIII and IX.

7 Are Nyborg's conclusions supported by the data?

It is outside the scope of this committee to discuss the use of hierarchical factor analysis for the study of a g–factor.

However, the way Nyborg uses the method for studying sex differences calls for some comments. The discussion below is split into two parts the first describing a basic problem of identifying a sex difference in the g-factor from the hierarchical factor model, and the second discussing the dependency of the chosen measure of the sex difference on the test battery used.

7.1 Non identifiability

Nyborg thinks of the *point-biserial factor loading on sex* as a correlation between sex and the g–factor. Just as the test for zero correlation based on an empirical correlation is equivalent to the test that there is no difference in the means of the two groups, a correlation ρ between the dichotomous variable *sex* and another variable can be written in terms of the standardized difference δ between the means for the two sexes,

$$\rho = \frac{\delta\sqrt{n_1 n_2/n^2}}{\sqrt{1+\delta^2 n_1 n_2/n^2}}. \tag{3}$$

When the sample sizes n_1 and n_2 are of the same order of magnitude and δ is not too large this expression reduces to the approximate relation $\delta \approx 2\rho$. This is for example used in Jensen (1998, p. 539, Table 13.1) for translating a loading into a sex difference in the g–factor. Now, the g–factor cannot be measured and the difference δ can only be seen through the implications it has on the differences in the means of the 20 test variables. What Nyborg (and Jensen, 1998) tries to do is to reverse this relation by inserting r_{pbs} into the correlation matrix before the factor analysis.

However, there is a fundamental non–identifiability problem in this approach. Just as a mean difference in the g-factor has implications for the differences in the means of the 20 test variables, so have the mean differences for the other factors. Since the g–loadings are a linear combination of the loadings of the primary factors the influence of the mean difference in the g-factor cannot be separated from the influence from the other factors. *More precisely, one cannot tell if the differences in the means of the 20 test variables are due to a difference for the g–factor or due to a difference among the primary factors.* Mathematical details of this argument are given in Appendix D.

The analysis with the point biserial correlation inserted into the correlation matrix before the factor analysis seems to have been introduced by Jensen (1998, p. 538). The committee has not had any material describing the properties of this method in detail. In Appendix D we have therefore given a mathematical reformulation of the method. It seems that this method lacks consistency in the following sense. If one imagines that there are no difference in the means of the primary factors, and the only difference is in the g–factor, then the method will not reproduce this mean difference exactly.

7.2 Dependence on test battery

Nyborg writes in [3, p. 195] about test selection bias and establishing a sex difference that "As mentioned previously, verbal and spatial tests typically benefit females and males differently; and their simultaneous presence in a test battery would tend to balance out the sex biasing effect", "In other words, summed or averaged subtest scores are no scientifically acceptable alternatives to measures of general ability based on inter–test correlations"and "However, in the process of establishing whether there is a sex difference, no IQ test results will suffice as a sole basis for deciding whether an observed difference in general ability g is real, or rather a mirror of test item composition".

Nyborg seems to think that when using the g-factor and somehow correlating this with the differences in the means of the test variables the above mentioned problems have been avoided. However, the method used by Nyborg based on the point biserial correlation inserted in the correlation matrix reduces approximately to a weighted sum of the effect sizes d (see Appendix D, formula (9)). Typically, these weights are almost all positive and, intuitively, what is measured is therefore the overweight of tests in the battery that favours one of the sexes. We are therefore precisely back to the problem that Nyborg purports to avoid. For the actual test battery that Nyborg uses we have a strong bias towards males with 16 or 17 out of the 20 tests giving a positive effect size. In the table below is an example of the weights entering the calculation of *point-biserial factor loading of sex*. The first column gives the g–loadings and the second column gives the weights multiplied by 4 (which makes the weights comparable to the loadings). The table is for the $N = 62$ adult data set using maximum likelihood factor analysis, the oblimin oblique rotation, and using the pooled variance estimate for the two sexes. In this case the *point-biserial factor loading of sex* is 0.112, whereas the approximation based on the weights below is 0.094.

```
 [1,]    0.43  0.23
 [2,]    0.46  0.37
 [3,]    0.48  0.22
 [4,]    0.44  0.53
 [5,]    0.53  0.35
 [6,]    0.34  0.22
 [7,]    0.11 -0.19
 [8,]    0.26  0.23
 [9,]    0.24  0.27
[10,]    0.42  0.10
[11,]    0.34  0.19
[12,]    0.39  0.21
[13,]    0.32  0.00
[14,]    0.14  0.09
[15,]    0.35  0.18
[16,]    0.10  0.12
[17,]    0.30  0.19
[18,]    0.64  0.76
[19,]    0.36  0.33
[20,]    0.68  1.21
```

The committee finds that a sex difference in the g–factor cannot be identified using the hierarchical factor model. The committee furthermore finds that the measure used by Nyborg is of the same nature as previous measures of general ability (a weighted average of effect sizes) and therefore flawed in the way described by Nyborg himself.

As mentioned in Section 6 the hierarchical factor analysis involves a number of choices. We have in Appendix D made a number of runs, varying the factor analysis method, varying the number of factors, and varying the oblique rotation. From the numbers in the appendix we get a *point-biserial factor loading on sex* that varies between 0.06 and 0.20 for the $N = 62$ adult data set, and varies between 0.14 and 0.23 for the $N = 119$ children data set. In these runs we have used a common correlation matrix for the two sexes taking into account that the two sexes have their own mean value, whereas Nyborg estimate the correlation matrix as though the two sexes have the same mean value.

In [SSRC] under point 5 is written "Forskeren bør ligeledes tage de nødvendige forbehold og redegøre for mulige fejlkilder og usikkerheder. Endvidere bør forskeren diskutere alternative fortolkninger af resultaterne.

The committee finds that since the hierarchical factor analysis involves a number of choices it is good scientific practice to discuss robustness of the results under the various choices.

The discussion in this section relates to questions I and XIV.

Appendix A Terms of reference

The terms of reference of December 7, 2005, contains 14 requests and questions:

I) Foretage en vurdering af de metoder, som Helmuth Nyborg har anvendt ved udvælgelsen og behandlingen af sine data, og dermed holdbarheden af de meddelte resultater.

II) Foretage en vurdering af overensstemmelsen mellem, hvad Helmuth Nyborg har rapporteret til offentligheden og forlag, og den faktiske udvælgelse og behandling af data.

III) Foretage en vurdering af Helmuth Nyborgs overholdelse af de krav, der efter udvalgets opfattelse gælder for sikring og fremlæggelse af datadokumentation for egne offentliggjorte videnskabelige resultater i en undersøgelse af denne karakter.

IV) Foretage en vurdering af Helmuth Nyborgs overholdelse af eventuelle andre krav til videnskabelig forskning, herunder offentliggørelse af dens resultater, som udvalget finder gældende for en undersøgelse af denne karakter.

V) Hvorledes forholder dette sig til det ifølge Helmuth Nyborg anvendte datamateriale ? ($N = 325$ data set)

VI) Findes denne metode ud fra sædvanlige standarder for forsvarlig anvendelse af statistiske metoder at være anvendelig på et datamateriale som det foreliggende ?

VII) Er det muligt – eventuelt med assistance fra Helmuth Nyborg – ved anvendelse af den af Helmuth Nyborg beskrevne beregningsmetode og den af ham i [7] beskrevne regnefejl at rekonstruere Tabel 3 og Figur 7 i [2] ud fra datamaterialet i [9, children 325] ?

VIII) Er det muligt – eventuelt med assistance fra Helmuth Nyborg – ved anvendelse af den af Helmuth Nyborg angivne beregningsmetode og nu uden den omtalte regnefejl, at rekonstruere tabelmaterialet i [5], Table 1 og Figure 1, og i [6], s. 2, 3 og 7, herunder den korte omtale af resultaterne i [22], p. 506, ud fra datamaterialet i [9, children 119] ?

IX) Er det helt tilsvarende muligt – eventuelt med assistance fra Helmuth Nyborg – ved anvendelse af Helmuth Nyborgs angivne beregningsmetode at rekonstruere tabelmaterialet i [5], Table 2 og Figure 1, i [6], s. 5, 6 og 7, og i [22], Table 1, Table 2 og Fig. 1, ud fra datamaterialet i [9, adult 62], jf. file *Adult complete data set* $N = 62$ *with id and sex.sta* på disketten [8] ?

X) Kan det ud fra denne oplysning vurderes, hvorvidt de 52 voksne er et tilfældigt udvalgt subsample af de 62 voksne ?

XI) Lever denne metode og den måde hvorpå den ifølge [10] og referaterne af måder med Helmuth Nyborg den 18. marts 2005 [11] og den 26. april 2005 [14] samt efterfølgende korrespondance [15,16,17,19,20] administreres, op til kravene om objektiv og neutral udvælgelse af subsamples og til de sædvanlige forventninger om gennemsigitghed ved udvælgelsen af subsamples ?

XII) Kan det antages, at man ved sædvanlig omhu i omgangen med protokolmateriale [8a] og datamateriale vil kunne komme i en situation som den af Helmuth Nyborg i referaterne [11,14] og efterfølgende korrespondance [15,16,17,19,20] beskrevne, når han forklarer, hvorledes dataene for de oprindelige 52 undersøgte voksne personer ikke kan udsondres fra de nu foreliggende data for 62 undersøgte personer og heller ikke på anden vis kan gendannes ?

XIII) Er det i overensstemmelse med sædvanlig videnskabelig praksis, når en regnefejl i offentliggjorte resultater skal korrigeres, at man – som det synes i Helmuth Nyborgs tilfælde – ikke gennemfører og offentliggør en korrekt beregning ud fra det oprindelige datamateriale ?

XIV) Udvalget bedes vurdere, hvilket grundlag for de udsagn der fremsættes i bogen [3] afsnit 4.2.10, pp. 208-209 og i manuskriptet [4], der er i det materiale, Helmuth Nyborg har stillet til rådighed, herunder i de oplysninger, han giver i referaterne [11,14] og i den efterfølgende korrespondance [15,16,17,19,20].

Appendix B Adult $N = 52$ and $N = 62$ data sets

The table below parallels Table 4 in [2], but here based on the $N = 62$ adult data set. The order of the rows are as in [2]. The columns included are in Nyborg's notation: Mean Boys (md), Mean Girls (mg), Difference Boys-Girls (dif), t–test p (ptt), Effect Size d (efs), Point-biserial correlation (pbs: correct version, pbsny: Nyborg's wrong expression (2)), sd Boys (sdb), and sd Girls (sdg).

```
         mb     mg   dif  ptt   efs   pbs pbsny   sdb   sdg
 [1,]  -1.63  -2.74  1.11 0.12  0.40  0.20  0.23  2.27  3.26
 [2,]  -1.44  -1.90  0.45 0.12  0.40  0.20  0.24  0.77  1.39
 [3,]  -2.56  -3.49  0.94 0.16  0.36  0.18  0.21  1.88  3.12
 [4,] -34.19 -40.39  6.20 0.42  0.21  0.10  0.11 29.16 30.57
 [5,]  -1.39  -3.45  2.06 0.04  0.53  0.26  0.33  2.96  4.61
 [6,]   7.32   6.10  1.23 0.11  0.41  0.21  0.25  3.05  2.88
 [7,] 313.29 289.00 24.29 0.02  0.60  0.29  0.38 38.11 42.88
 [8,] 353.68 341.35 12.32 0.24  0.30  0.15  0.17 40.06 41.72
 [9,]  14.23  14.61 -0.39 0.76 -0.08 -0.04 -0.04  5.19  4.51
[10,]  10.90  10.00  0.90 0.10  0.43  0.21  0.25  2.01  2.22
[11,]  10.10  10.71 -0.61 0.39 -0.22 -0.11 -0.10  2.71  2.85
[12,]  11.13  10.84  0.29 0.62  0.13  0.06  0.07  2.29  2.31
[13,]  13.19  12.23  0.97 0.18  0.34  0.17  0.20  2.65  2.95
[14,]   9.94   9.55  0.39 0.53  0.16  0.08  0.09  2.48  2.38
[15,]  11.58  10.90  0.68 0.17  0.35  0.18  0.20  1.63  2.18
[16,]  11.13  12.61 -1.48 0.03 -0.56 -0.27 -0.19  2.68  2.60
[17,]  12.97  12.00  0.97 0.10  0.43  0.21  0.26  2.14  2.38
[18,]  14.29  14.06  0.23 0.74  0.08  0.04  0.04  2.60  2.76
[19,]  11.94  11.84  0.10 0.88  0.04  0.02  0.02  2.50  2.41
[20,]  13.52  13.68 -0.16 0.81 -0.06 -0.03 -0.03  2.64  2.56
```

The interesting observation is that for rows 5 and 7–20 there is complete agreement with the numbers given in [2, Table 4] based on the $N = 52$ data set. This is highly puzzling since [2] is made before any changes to the original $N = 52$ data set is supposedly made. To further highlight this puzzle we next show that Nyborg appears to have used the number $N = 62$ in the calculation of the variance p–value in the last column of [2, Table 4]. We first explain that Nyborg has used the *Variance F ratio* column to calculate the variance p–value. We can see this because Nyborg has an error in row 6 so that the F ratio is not in accordance with the stated standard deviations (the F ratio is wrong with a factor 2), however the p–value is in accordance with the stated F–ratio. Thus we take the *Variance F ratio* column and calculate the corresponding p–value from an $F(25, 25)$–distribution, corresponding to $N = 52$ or from an $F(30, 30)$–distribution, corresponding to $N = 62$. We also have to take into consideration that Nyborg most likely has rounded the numbers in the *Variance F ratio* column after calculating the p–value. So, as an example, where Nyborg in row 5 writes the number 2.42 we calculate p–values for this number as well as for 2.415 and 2.425. The p–values are listed below for the rows where Nyborg seems to have used the $N = 62$ data set. Columns marked + are where we add 0.005 to the number and columns marked - are where we subtract 0.005 from the number. We see that in all cases we have agreement using degrees of freedom corresponding to $N = 62$, and in 11 of the cases we cannot get agreement using degrees of freedom corresponding to $N = 52$.

```
            N=52 N=52 N=52    N=62 N=62 N=62
      Nyborg    +  p25    -       +  p30    -
 [5,]   0.02 0.03 0.03 0.03    0.02 0.02 0.02
 [7,]   0.52 0.55 0.55 0.56    0.51 0.52 0.52
 [8,]   0.83 0.84 0.85 0.86    0.82 0.83 0.84
 [9,]    0.45   0.49 0.49 0.50   0.45 0.45 0.46
[10,]    0.58   0.60 0.61 0.62   0.57 0.57 0.58
[11,]    0.78   0.79 0.80 0.80   0.77 0.78 0.79
[12,]    0.96   0.95 0.96 0.97   0.95 0.96 0.97
[13,]    0.56   0.59 0.59 0.60   0.55 0.56 0.57
[14,]    0.83   0.84 0.85 0.86   0.82 0.83 0.84
[15,]    0.11   0.15 0.15 0.15   0.11 0.12 0.12
[16,]    0.87   0.88 0.89 0.89   0.86 0.87 0.88
[17,]    0.56   0.59 0.59 0.60   0.55 0.56 0.57
[18,]    0.75   0.75 0.76 0.77   0.73 0.74 0.75
[19,]    0.84   0.84 0.85 0.86   0.82 0.83 0.84
[20,]    0.86   0.86 0.87 0.88   0.84 0.85 0.86
```

The above observations show that for 15 of the variables The $N = 52$ data set is not a subset of the $N = 62$ data set, but instead identical to the $N = 62$ data set, and that the number $N = 62$ is used in the calculations.

Having turned to the $N = 62$ data set Nyborg states the effect size d in [6, Table 1] and in [22, Table 1]. We reproduce these numbers below together with the numbers (efs) that we obtain from the data in [8] for the $N = 62$ data set. We see that we get complete agreement with [6, Table 1: d06] and that, for some unknown reason, the published numbers, in [22, Table 1: d22] contain some minor errors.

```
        efs   d06   d22
 [1,]  0.40  0.40  0.39
 [2,]  0.40  0.40  0.40
 [3,]  0.36  0.36  0.36
 [4,]  0.21  0.21  0.21
 [5,]  0.53  0.53  0.52
 [6,]  0.41  0.41  0.41
 [7,]  0.60  0.60  0.58
 [8,]  0.30  0.30  0.30
 [9,] -0.08 -0.08 -0.08
[10,]  0.43  0.43  0.42
[11,] -0.22 -0.22 -0.22
[12,]  0.13  0.13  0.13
[13,]  0.34  0.34  0.34
[14,]  0.16  0.16  0.16
[15,]  0.35  0.35  0.35
[16,] -0.56 -0.56 -0.54
[17,]  0.43  0.43  0.42
[18,]  0.08  0.08  0.08
[19,]  0.04  0.04  0.04
[20,] -0.06 -0.06 -0.06
```

Appendix C Reconstructing Nyborg's results

The point biserial correlation is the Pearson correlation between sex (coded as 1 for men and 0 for women) and a continuous variable. The exact formula is

$$\frac{\sum_i y_i(x_i-\bar{x})}{\sqrt{\sum_i(y_i-\bar{y})^2\sum_i(x_i-\bar{x})^2}} = \frac{d\sqrt{\frac{n_1n_2}{n(n-2)}}}{\sqrt{1+d^2\frac{n_1n_2}{n(n-2)}}}. \tag{4}$$

Nyborg uses the approximation r_{pbs} in (1) from Jensen (1998), but misreads this and uses instead r_{ny} from (1). This is illustrated below for the $N = 62$ data set. In the Table the columns are exact (formula (4)), Jensen (formula r_{pbs} in (1)), wrong (formula r_{ny} in (1)), Nyborg [6, Table 1], and old from (2). The last column is included to show that Nyborg's two errors together do not influence the results dramatically.

```
      exact Jensen wrong Nyborg   old
 [1,]  0.20   0.19  0.20   0.20  0.23
 [2,]  0.20   0.20  0.21   0.21  0.24
 [3,]  0.18   0.18  0.19   0.19  0.21
 [4,]  0.10   0.10  0.10   0.10  0.11
 [5,]  0.26   0.26  0.28   0.28  0.33
 [6,]  0.21   0.20  0.21   0.21  0.25
 [7,]  0.29   0.29  0.31   0.31  0.38
 [8,]  0.15   0.15  0.15   0.15  0.17
 [9,] -0.04  -0.04 -0.04  -0.04 -0.04
[10,]  0.21   0.21  0.22   0.22  0.25
[11,] -0.11  -0.11 -0.11  -0.11 -0.10
[12,]  0.06   0.06  0.06   0.06  0.07
[13,]  0.17   0.17  0.18   0.18  0.20
[14,]  0.08   0.08  0.08   0.08  0.09
[15,]  0.18   0.17  0.18   0.18  0.20
[16,] -0.27  -0.27 -0.29  -0.29 -0.19
[17,]  0.21   0.21  0.22   0.22  0.26
[18,]  0.04   0.04  0.04   0.04  0.04
[19,]  0.02   0.02  0.02   0.02  0.02
[20,] -0.03  -0.03 -0.03  -0.03 -0.03
```

The next table shows the g–loadings for different choices of the method for finding the factors and different oblique rotations. The column Nyborg is taken from [6, Table 1]. MLE signifies maximum likelihood factor analysis, PF1 and PF2 are two principal factor analyses. The principal factor analysis (PF1) and promax rotation give the best agreement with Nyborg's results. The PF1 method uses the largest correlation between the ith variable and one of the other variables as preliminary estimates of the communalities, whereas the PF2 method uses the square of the multiple correlation coefficient of the ith variable with all the other variables as preliminary estimates of the communalities.

```
N=62 adult data set:
      Nyborg MLE,obl MLE,pro PF1,obl PF1,pro PF2,pro
 [1,]   0.37    0.42    0.46    0.38    0.43    0.46
 [2,]   0.46    0.46    0.53    0.44    0.48    0.50
 [3,]   0.41    0.47    0.52    0.42    0.47    0.51
 [4,]   0.53    0.45    0.51    0.45    0.48    0.45
 [5,]   0.61    0.55    0.66    0.52    0.58    0.57
 [6,]   0.46    0.40    0.48    0.40    0.48    0.45
 [7,]   0.31    0.17    0.30    0.26    0.34    0.30
 [8,]   0.35    0.25    0.33    0.29    0.35    0.31
 [9,]   0.23    0.21    0.20    0.23    0.24    0.22
[10,]   0.55    0.48    0.55    0.50    0.58    0.53
[11,]   0.39    0.34    0.35    0.35    0.35    0.32
[12,]   0.46    0.44    0.45    0.40    0.39    0.37
[13,]   0.46    0.37    0.42    0.43    0.51    0.44
[14,]   0.23    0.14    0.15    0.20    0.18    0.17
[15,]   0.47    0.41    0.46    0.43    0.49    0.43
[16,]   0.00    0.07    0.05    0.03    0.02    0.03
[17,]   0.40    0.32    0.39    0.35    0.41    0.39
[18,]   0.60    0.64    0.68    0.56    0.58    0.59
[19,]   0.35    0.35    0.39    0.31    0.37    0.36
[20,]   0.46    0.53    0.55    0.44    0.45    0.47
[21,]   0.28    0.14    0.23    0.21    0.28    0.24
-----------------------------------------------------
dist            0.073   0.053   0.047   0.037   0.050

N=119 chldren data set:
      Nyborg MLE,obl MLE,pro PF1,obl PF1,pro PF2,pro
 [1,]   0.55    0.39    0.39    0.45    0.45    0.45
 [2,]   0.29    0.38    0.19    0.33    0.24    0.26
 [3,]   0.58    0.53    0.44    0.54    0.50    0.51
 [4,]   0.53    0.70    0.56    0.60    0.54    0.57
 [5,]   0.39    0.49    0.37    0.45    0.39    0.40
 [6,]   0.59    0.63    0.59    0.60    0.60    0.59
 [7,]   0.32    0.47    0.25    0.42    0.32    0.34
 [8,]   0.27    0.44    0.19    0.39    0.27    0.30
 [9,]   0.28    0.45    0.33    0.42    0.37    0.37
[10,]   0.53    0.50    0.65    0.55    0.64    0.60
[11,]   0.48    0.47    0.54    0.52    0.59    0.55
[12,]   0.42    0.34    0.51    0.39    0.48    0.45
[13,]   0.47    0.44    0.58    0.50    0.60    0.56
[14,]   0.48    0.40    0.59    0.48    0.59    0.55
[15,]   0.35    0.29    0.39    0.33    0.39    0.37
[16,]   0.46    0.25    0.33    0.28    0.34    0.31
[17,]   0.38    0.27    0.35    0.31    0.34    0.32
[18,]   0.68    0.60    0.68    0.58    0.62    0.61
[19,]   0.56    0.35    0.47    0.39    0.44    0.42
```

```
[20,]   0.25    0.30    0.25    0.28    0.23    0.25
[21,]   0.38    0.24    0.33    0.23    0.30    0.28
[22,]   0.23    0.19    0.20    0.19    0.22    0.20
-----------------------------------------------------
dist            0.114   0.080   0.085   0.071   0.070
```

The following table gives the sum of the squared differences between the g–loadings given in [5] and the best fit from a linear combination of six primary factors. The methods included in the table is the maximum likelihood factor analysis (M-LE), two implementations of principal factor analysis (PF1 and PF2), and principal component analysis.

method	MLE	PF1	PF2	PC
$N = 62$	0.013	0.0011	0.0000037	0.0000014
$N = 119$	0.012	0.0044	0.0023	0.0000059

Our conclusion from this table is that Nyborg has used PC for extracting the primary factors, although with a small possibility that PF2 has been used for the $N = 62$ data set.

The two following tables show the result of using principal component for finding the primary factors and one of principal component, principal factor 1 or 2 for finding the general factor, and either oblimin or promax for the oblique rotation. None of the methods reproduces Nyborg's numbers exactly.

```
N=62 adult data set:
      Nyborg PC,obl PF1,obl PF2,obl PC,pro PF1,pro PF2,pro
 [1,]   0.37   0.51    0.40    0.33   0.54    0.42    0.42
 [2,]   0.46   0.59    0.45    0.38   0.58    0.47    0.47
 [3,]   0.41   0.57    0.43    0.37   0.60    0.46    0.46
 [4,]   0.53   0.60    0.43    0.40   0.55    0.49    0.49
 [5,]   0.61   0.69    0.51    0.46   0.71    0.60    0.60
 [6,]   0.46   0.56    0.39    0.36   0.63    0.51    0.51
 [7,]   0.31   0.40    0.24    0.23   0.42    0.33    0.33
 [8,]   0.35   0.43    0.28    0.26   0.39    0.33    0.33
 [9,]   0.23   0.34    0.20    0.21   0.22    0.20    0.20
[10,]   0.55   0.68    0.45    0.44   0.66    0.55    0.55
[11,]   0.39   0.45    0.31    0.31   0.34    0.32    0.32
[12,]   0.46   0.48    0.35    0.34   0.39    0.37    0.37
[13,]   0.46   0.59    0.37    0.38   0.56    0.47    0.47
[14,]   0.23   0.23    0.19    0.16   0.12    0.12    0.12
[15,]   0.47   0.56    0.37    0.37   0.56    0.47    0.47
[16,]   0.00   0.08    0.04    0.04   0.02    0.04    0.04
[17,]   0.40   0.52    0.36    0.33   0.53    0.43    0.43
[18,]   0.60   0.72    0.53    0.48   0.70    0.60    0.60
[19,]   0.35   0.45    0.32    0.30   0.54    0.44    0.44
[20,]   0.46   0.58    0.44    0.39   0.56    0.48    0.48
[21,]   0.28   0.30    0.20    0.19   0.36    0.27    0.27
-----------------------------------------------------------
```

```
dist           0.101   0.069   0.089  0.111   0.046   0.062

N=119 children data set:
      Nyborg PC,obl PF1,obl PF2,obl PC,pro PF1,pro PF2,pro
 [1,]   0.55   0.57    0.41    0.41   0.56    0.43    0.43
 [2,]   0.29   0.31    0.19    0.23   0.28    0.15    0.15
 [3,]   0.58   0.61    0.43    0.45   0.59    0.43    0.43
 [4,]   0.53   0.64    0.43    0.46   0.60    0.44    0.44
 [5,]   0.39   0.45    0.29    0.32   0.42    0.30    0.30
 [6,]   0.59   0.68    0.50    0.49   0.67    0.55    0.55
 [7,]   0.32   0.45    0.28    0.32   0.39    0.25    0.25
 [8,]   0.27   0.39    0.22    0.28   0.32    0.16    0.16
 [9,]   0.28   0.40    0.25    0.30   0.36    0.23    0.23
[10,]   0.53   0.69    0.55    0.51   0.71    0.64    0.64
[11,]   0.48   0.63    0.49    0.47   0.63    0.55    0.55
[12,]   0.42   0.52    0.42    0.38   0.54    0.52    0.52
[13,]   0.47   0.61    0.49    0.46   0.63    0.58    0.58
[14,]   0.48   0.65    0.53    0.48   0.66    0.62    0.62
[15,]   0.35   0.41    0.31    0.30   0.43    0.38    0.38
[16,]   0.46   0.45    0.37    0.31   0.47    0.45    0.45
[17,]   0.38   0.52    0.41    0.36   0.50    0.46    0.46
[18,]   0.68   0.76    0.58    0.53   0.76    0.67    0.67
[19,]   0.56   0.62    0.48    0.43   0.62    0.54    0.54
[20,]   0.25   0.35    0.20    0.23   0.30    0.19    0.19
[21,]   0.38   0.33    0.29    0.23   0.38    0.39    0.39
[22,]   0.23   0.25    0.22    0.19   0.27    0.27    0.27
-----------------------------------------------------------
dist           0.099   0.075   0.082  0.093   0.085   0.062
```

Appendix D Are Nyborg's conclusions supported by the data?

We start by describing the hierarchical factor model mathematically.

Let x be an m–dimensional vector with the results of the test battery for a single person. For consistency in the notation in the formulae below we let the coordinates of x be scaled by the standard deviation for each test. A k–factor model can then be written in the form

$$x = \mu + \Lambda f + u,$$

where μ is the mean, Λ is the $m \times k$ matrix of factor loadings, f is the vector with the k factors with mean zero, u is the vector of unique factors with mean zero, and we have $V(f) = I$, $Cov(f, u) = 0$, and $V(u)$ is a diagonal matrix. Let T be an oblique rotation and define $\tilde{f} = Tf$ and $\tilde{\Lambda} = \Lambda T^{-1}$. The second level of the hierarchical factor analysis consists of a factor model for $\tilde{f}$ with one factor only,

$$\tilde{f} = \tilde{g} z + \tilde{v},$$

where $\tilde{g}$ is the k–dimensional vector of loadings, z is the g–factor, $\tilde{v}$ is the vector of unique factors, and $v(z) = 1$, $Cov(z, \tilde{v}) = 0$, and $V(\tilde{v}) = D(\tilde{g})$ is a diagonal matrix with entries $1 - \tilde{g}_i^2$. Combining the two factor models we have

$$\begin{aligned} x &= \mu + \Lambda f + u = \mu + \tilde{\Lambda}\tilde{f} + u \\ &= \mu + (\tilde{\Lambda}\tilde{g})z + \tilde{\Lambda}D(\tilde{g})^{1/2}D(\tilde{g})^{-1/2}\tilde{v} + u \\ &= \mu + gz + \bar{\Lambda}v + u, \end{aligned}$$

where $g = \tilde{\Lambda}\tilde{g}$ is the vector of loading for the g–factor, and $\bar{\Lambda} = \tilde{\Lambda}D(\tilde{g})^{1/2}$ are the primary rotated and scaled factor loading, scaled so as the factors $v = D(\tilde{g})^{-1/2}\tilde{v}$ have variance one. The scaling by $D(\tilde{g})^{-1/2}$ is what Nyborg refers to as the Schmid–Leiman transformation.

Now consider two populations corresponding to the two sexes. Imagine that the factor structure is the same for the two sexes. This means that g and $\bar{\Lambda}$ are identical for the two sexes. The difference appears in the means only, $\psi = \mu_m - \mu_w$, where μ_m is the mean for men and μ_w is the mean for women. Since we from the outset have scaled the variables by their standard deviations ψ represents the theoretical effect sizes. We now imagine that ψ can be written as

$$\psi = g\delta + \bar{\Lambda}\beta, \tag{5}$$

where δ is a difference in the mean of the g–factor for the two sexes, and β is similarly the difference in the means of the primary factors in v (this structure of ψ is sometimes referred to as Spearman's weak hypothesis). An alternative way of formulating this structure for ψ is that the unique factors do not contribute to the sex difference. Nyborg does not mention this explicitly, but without this assumption there is no hope of identifying the difference δ for the g–factor. The important thing to realize in the structure (5) is that knowing ψ does not allow us to find δ. In particular we can write

$$\psi = g\delta + \bar{\Lambda}\beta = g(\delta - c) + \bar{\lambda}[\beta + D(\tilde{g})^{-1/2}\tilde{g}c], \tag{6}$$

so that by changing the difference in the means of the primary factors we obtain the same ψ, but now being able to choose an arbitrary value for the difference in the mean of the g–factor. In particular we can take $c = \delta$ so that the g–factor has the same mean for the two sexes. This situation is known as non–identifiability of the parameters, δ and β cannot be identified independently of one another. This is a fundamental property of the hierarchical factor model because the vector g of loadings of the g–factor is a linear combinations of the loadings in $\bar{\Lambda}$.

Let us now speak of δ and β as "true values". Then any procedure that tries to estimate δ really estimates $\delta - c$ for some implicitly chosen and unknown value of c. We illustrate this by a number of estimates. Let h be a vector and imagine that we want to "correlate"h with the vector of measured effect sizes d to obtain an estimate of δ. Because of the structure in (6) the natural estimate is

$$\hat{\delta}_h = \frac{h^* d}{h^* g}. \tag{7}$$

However, using (6), what we estimate in this case is not δ, but the composition

$$\frac{h^*\psi}{h^*g} = \delta + \frac{h^*\bar{\Lambda}\beta}{h^*g}.$$

The method of correlated vectors that Nyborg refers to [3, p. 197; 22 p. 501], Jensen (1998, p. 589-591), roughly corresponds to taking $h = g - \bar{g}$ in the above estimate. (The p–value that Nyborg and Jensen use is not correct, since the latter assumes that the coordinates of d are independent.)

We next consider in detail the estimate Nyborg obtains through the *point-biserial factor loading on sex*. Thus in this method the point biserial correlations are inserted in the correlation matrix before the factor analysis. (Nyborg uses the estimated correlation matrix obtained by treating the two sexes as one group, whereas, if there is a difference in the mean for the two sexes, it seems more appropriate to use a common correlation matrix treating the two sexes as separate groups.) Nyborg writes [22, p.501]: "It is empirically established that the inclusion of the sex r_{pbs} in the correlation matrix has no effect on the factor structure and only negligible effects on the subtests' g loadings (Jensen, 1998, p. 542, note 9)". Assuming this to be correct we can formulate the statement mathematically by saying that $\tilde{g}$ is not changed and $\tilde{\Lambda}$ is changed to

$$\ddot{\Lambda} = \begin{pmatrix} \tilde{\Lambda} \\ b^* \end{pmatrix},$$

for some vector b, when the point biserial correlations are inserted. Defining $\ddot{g} = \ddot{\Lambda}\tilde{g}$ the correlation matrix corresponding to this factor structure is

$$\ddot{g}\ddot{g}^* + \ddot{\Lambda}D(\tilde{g})\ddot{\Lambda}^* + D(u),$$

where $D(u)$ is the diagonal matrix with the unique variances for u. The last column, not including the last element, in this matrix is

$$\begin{aligned} &\tilde{\Lambda}\tilde{g}b^*\tilde{g} + \tilde{\Lambda}D(\tilde{g})b \\ &\quad = \tilde{\Lambda}\tilde{g}\tilde{g}^*b + \tilde{\Lambda}D(\tilde{g})b \\ &\quad = \tilde{\Lambda}[\tilde{g}\tilde{g}^* + D(\tilde{g})]b \\ &\quad = \tilde{\Lambda}\bar{b}, \end{aligned}$$

with $\bar{b} = [\tilde{g}\tilde{g}^* + D(\tilde{g})]b$. We want the vector $\tilde{\Lambda}\bar{b}$ to resemble the vector of point biserial correlations r_{pbs}, and we therefore take

$$\bar{b} = (\tilde{\Lambda}^*\tilde{\Lambda})^{-1}\tilde{\Lambda}^*r_{pbs}.$$

Finally, from this we find the *point-biserial factor loading on sex* ρ to be

$$\begin{aligned} \rho &= \tilde{g}^*b = \tilde{g}^*[\tilde{g}\tilde{g}^* + D(\tilde{g})]^{-1}\bar{b} \\ &= \tilde{g}^*[\tilde{g}\tilde{g}^* + D(\tilde{g})]^{-1}(\tilde{\Lambda}^*\tilde{\Lambda})^{-1}\tilde{\Lambda}^*r_{pbs} \\ &= G^*r_{pbs}, \end{aligned} \tag{8}$$

where G is defined implicitly through the expression in the second line. Since $r_{pbs} \approx \frac{1}{2}d$ and $\rho \approx \frac{1}{2}\delta$ (see 3), we get the estimate

$$\hat{\delta}_{pbs} \approx G^*d. \tag{9}$$

We thus see that the method used by Nyborg produces an estimate of the same form as described generally above, and with the same inherent problems due to the non–identifiability problem. Nyborg's estimate does not contain the standardization used in (7), where we divide by h^*g, that is, Nyborg estimate corresponds to G^*d instead of G^*d/G^*g. Replacing d by the true mean we thus get

$$G^*\psi = G^*g\delta + G^*\bar{\lambda}\beta,$$

and we thus do not get δ in the optimal situation with $\beta = 0$ unless $G^*g = 1$. Actually, G^*g can be expressed as

$$\begin{aligned} G^*g &= \tilde{g}^*[\tilde{g}\tilde{g}^* + D(\tilde{g})]^{-1}\tilde{g} \\ &= \frac{\tilde{g}^*D(\tilde{g})\tilde{g}}{1+\tilde{g}^*D(\tilde{g})\tilde{g}}, \end{aligned}$$

which is always less than 1. However,

$$\tilde{g}^*D(\tilde{g})\tilde{g} = \sum_i \frac{\tilde{g}_i^2}{1-\tilde{g}_i^2},$$

and this expression will be large if one of $\tilde{g}_i^2$ is close to one.

It is difficult to see why Nyborg (and Jensen, 1998) finds the use of G^*d through the *point-biserial factor loading on sex* better than the use of g^*d/g^*g.

To summarise: it is not possible to identify the difference δ in the mean of the g–factor for the two sexes within the hierarchical factor model. Any estimate of δ involves a hidden assumption as to what is put into δ and what is put into β, the difference in the means of the primary factors. Furthermore, the method base on the point-biserial factor loading on sex *reduces to a weighted sum (9) of effect sizes.*

Appendix D.1 Robustness

Since we did not manage to reproduce Nyborg's calculations exactly we quote here the results obtained by different parameter settings. Thus we try maximum likelihood factor analysis (M), principal factor analysis PF1 (P), we try 4 and 6 primary factors, and we try the oblimin ($\gamma = 0$, o) and the promax ($m = 4$, p) oblique rotations. The correlation matrix we use is the common correlation estimate allowing for different means in the two groups (thus we do not consider the two sexes as one group in this estimation, which is contrary to Nyborg). We use the exact point biserial correlation (4) and insert this in the correlation matrix before the factor analysis. We thus imitate Nyborg's analysis and give below the *point-biserial factor loading on sex* as well as the maximal correlation of the primary factors after the oblique rotation.

		M,o,4	M,o,6	M,p,4	M,p,6	P,o,4	P,o,6	P,p,4	P,p,6
N=62	loading	0.12	0.11	0.15	0.20	0.17	0.19	0.06	0.20
	max corr	0.34	0.37	0.54	0.47	0.33	0.33	0.59	0.52
N=119	loading	0.14	0.16	0.20	0.16	0.15	0.18	0.21	0.23
	max corr	0.44	0.41	0.56	0.63	0.33	0.41	0.55	0.60

For the $N = 62$ adult data set *point-biserial factor loading on sex* is generally smaller than the one obtained by Nyborg, maximum likelihood generally give smaller values than principal factor analysis, 4 factors generally gives a smaller value than 6 factors, and the promax oblique rotation generally gives a higher value than the oblimin rotation. The promax rotation makes the primary rotated factors more correlated than does oblimin. We emphasise here that the *point-biserial factor loading on sex* does not give a proper evaluation of the difference in the two sexes for the g–factor due to the reasons given above.