PROCEEDINGS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Group differences in the heritability of items and test scores

Jelte M. Wicherts and Wendy Johnson

Proc. R. Soc. B 2009 **276**, 2675-2683 first published online 29 April 2009
doi: 10.1098/rspb.2009.0238

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2009) **276**, 2675–2683
doi:10.1098/rspb.2009.0238
Published online 29 April 2009

Group differences in the heritability of items and test scores

Jelte M. Wicherts[1,*] and Wendy Johnson[2]

[1]*Department of Psychology, Psychological Methods, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands*
[2]*Department of Psychology, University of Edinburgh, Edinburgh EH8 9AD, UK*

It is important to understand potential sources of group differences in the heritability of intelligence test scores. On the basis of a basic item response model we argue that heritabilities which are based on dichotomous item scores normally do not generalize from one sample to the next. If groups differ in mean ability, the functioning of items at different ability levels may result in group differences in the heritability of items, even when these items function equivalently across groups and the heritability of the underlying ability is equal across groups. We illustrate this graphically, by computer simulation, and by focusing on several problems associated with a recent study by Rushton *et al*. who argued that the heritability estimates of items of Raven's Progressive Matrices test in North-American twin samples generalized to other population groups, and hence that the population group differences on this test of general mental ability (or intelligence) had a substantial genetic component. Our results show that item heritabilities are strongly dependent on the group on which the heritabilities were based. Rushton *et al*.'s results were artefactual and do not speak to the nature of population group differences in intelligence test performance.

Keywords: behaviour genetics; heritability; intelligence; psychometrics; nature versus nurture

1. INTRODUCTION

Heritability is the proportion of phenotypic trait variance that can be attributed to genetic variation. Though quantitative genetics textbooks are clear that heritability estimates depend on the properties of the trait in question, the population in which it is estimated, the environmental circumstances particular to that population and the way the trait is measured, these conditions are rarely given the recognition they deserve. For example, in a study of the heritability of age of menarche, Anderson *et al*. (2007) summarized the differences in results from prior studies in varying national groups as potentially owing to increasing inaccuracy of report of age of menarche with age of report and/or to censoring when age of report was young enough that not all participants had experienced menarche. But they did not raise the possibility that results might have varied because heritability might vary with population.

Yet it is clear such population differences in heritability do exist. For example, Heath *et al*. (1985) estimated the heritability of educational attainment in Norway based on pairs of male and female twins born between either 1915 and 1939 or 1940 and 1960. For females, heritabilities were basically stable in the two periods at approximately 45 per cent. In addition, more than 40 per cent of the variance was accounted for by shared environmental influences in both periods. By contrast, in males born between 1915 and 1939, heritability and shared environmental influences were similar to those for females, but for males born between 1940 and 1960, heritability was approximately 70 per cent and shared environmental influences accounted for approximately 10 per cent of the variance. Heath *et al*. interpreted this as evidence that increasing equalization of opportunity for education for males but not for females had led to educational attainment based on innate ability.

This example points to differences in environmental circumstances as the reason for differences in heritability across population groups. But group differences may also exist owing to genetic differences (i.e. allele frequencies), interactions between genetic and environmental effects, and lack of measurement invariance across groups resulting in differing genes contributing to what is termed, but is not really the same trait in different populations (Lubke *et al*. 2004). In this paper, we argue that, because of basic measurement properties, group differences in trait mean levels may also result in group differences in heritability of item scores. This occurs even if the heritability of the underlying trait is equal across groups, and even if the items are measurement invariant across groups, and so function equivalently across groups.

For example, if we measure the heritability of alcoholism using only questions such as, 'Have you ever had an alcoholic drink?', we will obtain much lower heritability estimates than if we also include questions such as, 'Do you sometimes share a bottle of wine with someone?' simply because questions of the first kind will show only a small variability in response within many populations, but also because heritability will be more readily apparent the more variability there is in the measure of a trait. The heritability of the measure of a trait depends not only on the characteristics of a measurement instrument, but also on the distribution of the trait in a particular population in relation to this instrument. For instance, an item such as 'Do you often drink enough to pass out?' may show higher heritability in a sample of alcoholics than in a sample of

* Author for correspondence (j.m.wicherts@uva.nl).

Received 11 February 2009
Accepted 7 April 2009

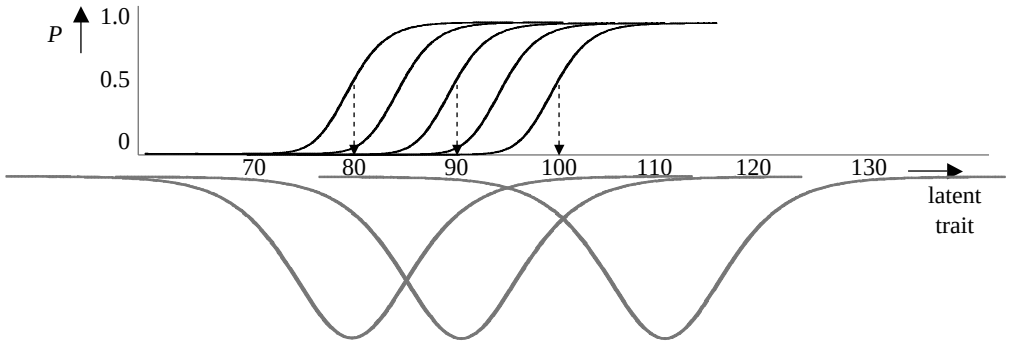

Figure 1. Five ICCs and the ability distribution in three groups.

non-alcoholics because it provides more information on differences in alcoholism among the former than among the latter population.

Failure to recognize the potential reasons for population differences in heritability may lead to the development and report of erroneous conclusions about the reasons for group differences in trait mean levels. For example, Rushton *et al.* (2007) recently aimed to show that population group differences on a test of general mental ability (or intelligence) are genetic rather than environmental in origin. Specifically, they computed heritability estimates of the intelligence test items on the basis of two North American twin samples and related these item heritabilities to differences in performance on these intelligence test items across a variety of ethnic groups in which heritability itself was not measured. Because they observed positive correlations, they concluded that group differences in the underlying trait must have genetic origins. Because environmental influences on the items in the North American groups were not correlated with the group differences, they concluded that the group differences in the underlying trait could not have environmental origins. These conclusions were unwarranted, however, as there is no reason to expect that the correlations between item statistics should be attributed to group differences in the intelligence test scores. In fact, the correlations between item statistics could be accounted for by inherent statistical properties of the data that said nothing about the sources of the group differences in the trait.

The purpose of this paper is to demonstrate how inherent statistical properties of the data that say nothing about the sources of the group differences in the trait could account for those group differences. We do this by pointing to several general properties of heritabilities based on item scores (henceforth item heritabilities), and, by extension, heritabilities based on test scores. These properties have not been generally recognized. We focus on dichotomous item scores (i.e. scored as either correct or incorrect) and test scores based on summations of such item scores. We develop our argument in a non-technical manner. In particular, we show that item heritabilities are not intrinsic properties of test items, as Rushton *et al.* appeared to suggest, but that item heritabilities, similar to all heritabilities, depend on the distribution of the trait in the population in which they are measured. This means that heritabilities do not necessarily generalize across population groups nor can they be used to explain the existences and magnitudes of group differences as Rushton *et al.* claimed.

2. FUNCTIONING OF ITEMS AT DIFFERENT ABILITY LEVELS

Regardless of the trait or how the trait is measured, heritability requires phenotypic variance. The estimation of item heritabilities as in the Rushton *et al.* study makes the effect of phenotypic variance particularly clear. Therefore, we use the data from Rushton *et al.*'s adult twin sample as an illustration. Rushton *et al.* made use of Ravens' Standard Progressive Matrices (Raven 1941), a well known and regarded intelligence test often chosen for population group comparisons because it does not rely explicitly on verbal cultural knowledge (though see Flynn 2007, for discussion of some of the issues involved in this assessment). It consists of 60[1] non-verbal matrix reasoning items that follow the same format and are scored as either correct or incorrect.

If the test is psychometrically sound (and Raven's is; see for example Raven *et al.* 1996; van der Ven & Ellis 2000), the probability of answering an easy item correctly should be higher than the probability of answering a hard item correctly, whatever the level of ability of the test taker. At the same time, a test taker with higher ability should have a greater probability of answering any item correctly than a test taker with lower ability. These properties are often depicted using item characteristic curves (ICCs; see, e.g. Embretson & Reise 2000) that show how the probabilities of correct response vary for each item with level of ability. Figure 1 presents examples of these ICCs for five hypothetical Raven's items. The x-axis of the graph represents test takers' latent ability expressed in IQ units, and the y-axis represents the probability that a test taker will answer an item correctly. Each item has its own curve, and each curve increases from left to right, showing the increasing probability of answering the item correctly with increasing ability. If the items reflect ability well, i.e. discriminate well among test takers of different ability levels, the curves are relatively steep, indicating that the range over which there is variability in the probability of answering the item correctly is narrow: test takers with greater ability will almost all get the item correct, and test takers with lower ability will almost all get it wrong. The location of the centre of each curve, which represents the point of ability at which there is a 50 per cent probability that a test taker will get the item correct, reflects the *item difficulty*: items further to the right of the graph require higher ability to answer correctly and hence have higher item difficulties. The easiest item in figure 1 has a difficulty of 80 on the latent trait scale, which means that test takers with an IQ of 80 have a 50 per cent of answering the item correctly. The Raven's items were

Table 1. CTT item statistics for three hypothetical groups and five hypothetical items. (*Note*. Item difficulties and item discrimination parameters are identical across groups.)

	proportion correct or *p*-value			scale reliability			item-total correlation		
item number	low	middle	high	low	middle	high	low	middle	high
1	0.50	0.84	1.00	0.879	0.955	0.657	0.79	0.66	0.16
2	0.30	0.69	0.99				0.76	0.75	0.21
3	0.16	0.50	0.98				0.66	0.80	0.35
4	0.07	0.30	0.93				0.52	0.77	0.52
5	0.02	0.16	0.83				0.37	0.66	0.66

specifically designed to differ in item difficulty but not in steepness of the slope, i.e. the *item discrimination parameters* are equal across the five items (Raven 2008). Tests similar to the Raven that have this property (van der Ven & Ellis 2000) fit the so-called Rasch model, which is a basic and well-known model in item response theory (IRT; see, e.g. Embretson & Reise (2000) for an excellent introduction to these modern psychometric techniques).

Within the framework of classical test theory (CTT; Lord 1980), item parameters are not linked to the latent trait scale, but rather to the observed performance of a group of examinees. The degree to which items discriminate between different trait levels and 'difficulty' of items is defined differently in IRT than in CTT. In CTT, the difficulty of an item is the percentage of people in a sample that answer the item correctly, so it can best be denoted by the proportion correct, or the *p*-value (i.e. because of potential confusion with the notion of item difficulty from IRT). Proportion correct of an item will differ between samples that differ in mean ability. Also, in CTT, the correlation between the dichotomous item scores and the total test score, or *item-total correlation* (ITC) will differ from one sample to the next, unless the ability distribution is equal in both samples. The ITC may be seen as an approximation to the square root of the reliability of an item (Bechger *et al*. 2003), so the same applies to item reliability. By contrast, in IRT, the item difficulty and discrimination parameter of items are defined with respect to the latent trait scale (Mellenbergh 1996). In modern IRT, the item parameters are theoretically *independent* of group ability distribution and are linked directly to the latent trait, while in CTT, item parameters are confounded by the ability distribution of the group (Lord 1980).

Figure 1 also displays the normal distributions of the trait in three groups that differ in mean, but not in variance. If the ICCs of the five items apply equally to the groups with these different distributions, the test is said to be *measurement invariant* across groups (Mellenbergh 1989). The existence of measurement invariance means that the probability of answering an item correctly depends only on the ability level of the test taker and not on his or her group membership.

Table 1 contains classical item statistics for the five hypothetical items in the three hypothetical ability groups. The ITCs in this example are based on the group-based correlations between the item scores and the latent trait. The ITC may thus be seen as the degree to which the item score correlates with the trait, *within* a group. Because the groups have identical ICCs as shown in figure 1, the relations between the latent traits and the item scores are identical across groups even though the groups differ in mean ability. Nonetheless, the ITCs and the items' *p*-values are clearly group-dependent. In fact, the ITCs of the low and the high groups across the items are almost perfectly negatively correlated despite the fact that the relation between the items and the underlying trait is identical across groups for all five items. The *p*-values of the items correlate highly across groups, although this relation is nonlinear when groups differ in mean ability (Lord 1980). The classical item statistics differ strongly for the groups, because the items function differently over different trait levels and the groups differ in mean trait level.

In IRT, the steepness of the slope of the ICC at a particular level of the trait indicates the amount of information that the item provides on differences in the trait. This so-called *item information* differs over ability levels and is at its maximum at the level of the trait at which the expected probability of answering the item correctly is 0.50 (i.e. at the item difficulty). The item information approaches zero at levels of the trait at which the expected probabilities of answering correctly approach zero or one. Although the items in figure 1 have equal discrimination parameters, the items differ in item difficulty (i.e. location on the latent trait scale), which results in differences over ability levels in how precisely the items measure ability. For instance, the easiest item is most precise around an IQ level of 80, while the most difficult item is most precise at an IQ level of 100. At higher levels of ability the easiest item provides little information on differences in the trait, because it is too easy and will be answered correctly by almost all test takers at the higher ability level. Similarly, the most difficult item is quite informative for higher levels of ability, but less informative for the lower ability levels. If we compare groups that differ in mean ability, the groups differ in the degree to which the items differentiate between ability levels *within* the groups, which may be expressed in terms of the ITCs within each group.

Now suppose that the underlying trait is heritable to the same degree across the three groups and we establish the heritabilities of the five items in each of the three groups. Even if the heritabilities of the trait are identical across groups, this will not show up at the item level, because the item reliabilities and thus their heritabilities depend on the degree to which each item differentiates between ability levels *within* each group, as reflected in group differences in the ITCs. The easiest item will show a very low heritability for the high-scoring group, because

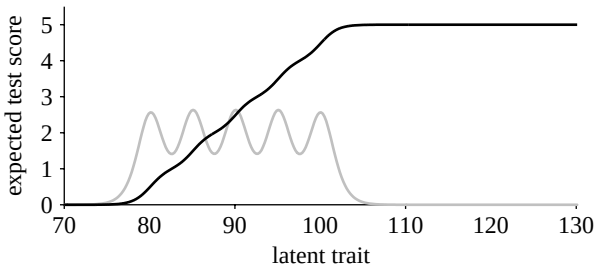

Figure 2. Test characteristic curve and test information of five item test in figure 1.

this item hardly measures the trait in this group (i.e. its ITC is near zero and both monozygotic and dizygotic twins in this group all tend to get it right). At the same time, the easiest item will have a relatively high heritability in the low-scoring group because this item is most informative for differences in ability within this group, thus showing the highest ITC (and monozygotic and dizygotic twins in this group get it right differentially, depending on genetically influenced similarity of ability). In this way, the heritability of an item depends on the item difficulty parameter in relation to the ability distribution of the group in which the heritability is determined. Therefore, the within-group heritability of an item does not necessarily generalize to other groups that differ in mean ability, even if the item functions equivalently (i.e. IRT item parameters are group-invariant) across groups and the groups have equal trait variance *and* trait heritability. Note that the item heritability can be expected to be at a maximum when its probability of being answered correctly is approximately 0.50 because the item variance and item information are at the maximum at this point. If item heritability is due only to heritability of underlying trait then item heritability is the heritability of the underlying trait multiplied by the item-test correlation, which in turn depends in a lawful way on both item difficulty and the distribution of the trait within the population in which heritability is measured. In this perfect hypothetical world, in which all items have equal discrimination parameters, difficulties of items are invariant across groups, and the extent to which they are imperfectly discriminating is not in itself heritable, the reason that items differ in heritability both within a group and across groups is that they differ in item difficulty.[2]

Normally, item scores are summed to arrive at the full test score. The information in the five items can be summed to establish the amount of information that test scores provide concerning differences on the latent trait. In figure 2, we display this so-called *test information* of the five items related to the levels of the trait (grey curve). The test information shows that the measurement precision of the test scores is not identical across the entire ability range and depends on the distribution of item difficulties. It is clear that the five-item test is more informative for lower IQ levels. Another way to consider the measurement precision of the test score is to consider the test-characteristic curve (Lord 1980), which is the monotonically increasing relation between the true score on the test and the latent trait. The test characteristic curve for the five-item test is also displayed in figure 2 as the black curve. This function is nonlinear, thereby reflecting the differences across the ability spectrum in the functioning of the test. The steepness of the test characteristic curve is a function of the sum of the item information at that particular level of ability. For instance, at higher IQ levels, this five-item test is much less precise in measuring the underling trait than at lower IQ levels, because the items in the test have difficulties that are relatively low. This is reflected in the test characteristic curve being completely flat at the higher IQ levels. This depresses the test reliability and normally leads to skewed distributions of test scores in higher-scoring samples. If this test is used to estimate the heritability of the trait in the higher-scoring group, the heritability of the trait will appear to be much lower than it actually is, and this downward bias can be substantial (Neale *et al.* 2005; van den Berg *et al.* 2007). For instance, if the trait heritability is 0.80 in all three groups in figure 1, then the estimated heritabilities on the basis of this very limited test will be 0.70, 0.76 and 0.53 for the low, middle, and high-scoring groups, respectively. To conclude, groups that differ in mean ability can differ in heritability of the test scores because tests differ in measurement precision over the ability range, even if the heritability of the trait is equal across groups. From a practical perspective, the problem is potentially most pronounced when the test consists of only one item, i.e. at the item level.

3. HOW THIS APPLIES TO THE RESULTS OF Rushton *et al.* (2007)

Rushton *et al.*'s (2007) conclusions were based on the observation of significant correlations between item heritabilities in one group and between-group differences in item *p*-values or proportion correct. Their interpretation of these correlations relied on two assumptions. First, they had to assume that trait heritabilities, and by extension item heritabilities, were the same across groups, and second, they had to assume that any association between item heritabilities and group differences in proportion correct of items was causal: that genetic influences caused the group differences in items' *p*-values. We have shown that item heritabilities are related to item difficulties, but they also depend on the distributions of ability in the population groups in which they are calculated, thus violating the first assumption. Moreover, though they are related, item heritabilities do not create group differences in item proportion correct. In fact, for completely unrelated reasons, the item heritabilities form an inverted U shape across the distribution of ability measured by the test, with the highest item heritabilities for the items in the middle of the distributions of ability in each group, and the group differences in item proportion correct forms a very similar shape across this same distribution of ability. This creates the correlation Rushton *et al.* observed, but the correlation is artefactual rather than the reflection of a causal effect. We would expect that item heritabilities to be different across groups when these groups differ in terms of the underlying trait.

To show this, we simulated data in three latent ability groups ($n = 10.000$ per group) on an idealized test that comprised 36 items based on the Rasch model. In this test, the equally discriminating items[3] were distributed with difficulties at 0.25 point intervals over the standardized ability range from -5 to 3.25. This reflects the actual item difficulty parameters on the Raven test, as the items tend

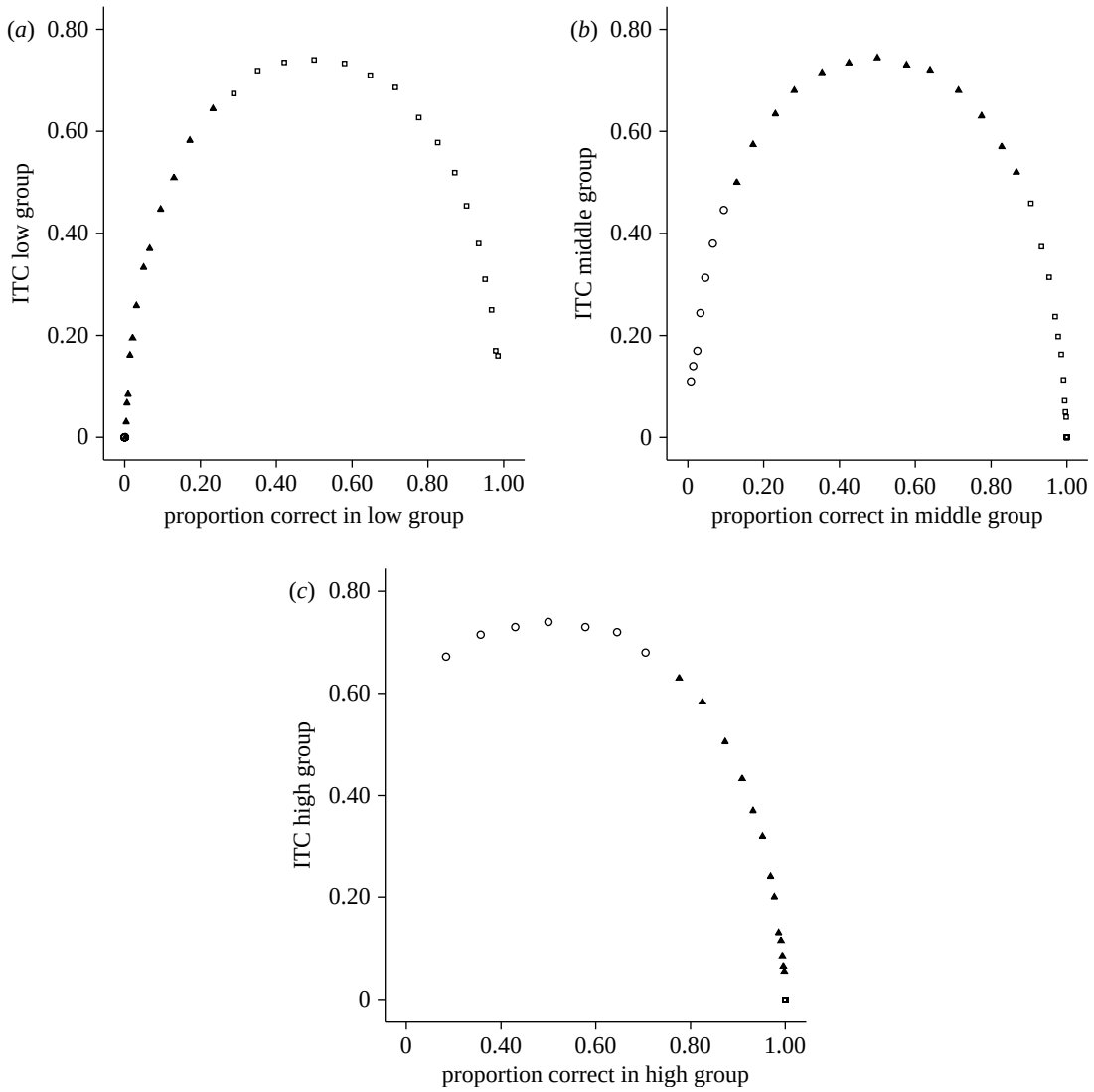

Figure 3. Results from simulation study showing the nonlinear relation between the vector of ITCs and the vector of standardized group differences in *p*-values on the items. (*a*) Low group; (*b*) middle group; (*c*) high group.

to be relatively easy in many samples. The item parameters were chosen to match the item properties in the data analysed by Raven *et al.* Ability was distributed normally within the groups, but the groups differed in mean ability as follows: low group ($M=-2.5$, s.d.$=1$), middle group ($M=0$, s.d.$=1$), and high group ($M=2.5$, s.d.$=1$). Thus, the groups differed greatly in mean ability, as was the case in the samples considered by Rushton *et al.* The average proportions of items correct in the three groups were 0.35, 0.62 and 0.88 for the low, middle and high-ability groups, respectively. The ITCs in each group were the point-biserial correlations between item scores and the trait values.

Figure 3 depicts the ITCs against the proportion correct for each of the three groups in our simulation. The ITCs differ across the three ability groups. For instance, the ITCs of the Low group correlate at $r=-0.88$ with the ITCs in the high group. It is clear that the ITCs are highest for those items that are maximally informative for the group, i.e. for items that show proportions correct around 0.50 *within* the groups. Likewise, *item heritabilities* will be at a maximum for those items with difficulty parameters near the middle of the ability distribution in the group in which the heritabilities are computed.

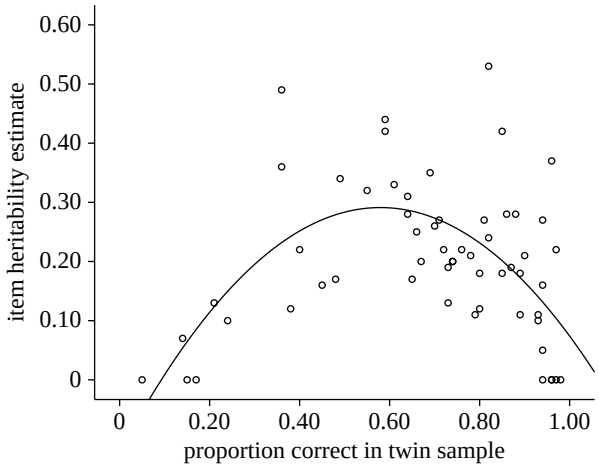

Figure 4. Relation between *p*-value and item heritability in adult twin sample of Rushton *et al.* (2007).

Items that show proportions correct approximately 0.50 in that particular sample are also the items with the highest heritability for that particular sample.

This can also be seen in figure 4, in which we display the item heritability estimates from Rushton *et al.*'s study 2 against the *p*-values of the 58 items in the twin sample.

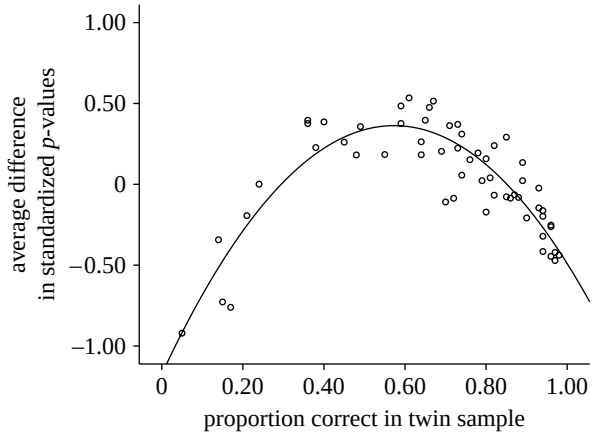

Figure 5. Relation between average of 55 group differences in standardized *p*-values and *p*-value in twin sample of Rushton *et al.* (2007).

The item heritabilities are at their maximum for items that show proportions correct approximately 0.50 in the twin sample (although the inflection point is somewhat higher than 0.50). The quadratic coefficient explains 36 per cent of the variance, which is quite impressive given the relatively small twin sample and hence the unreliability of the data.

The central result of Rushton *et al.* (2007) was based on generally positive correlations between the item heritabilities from the twin sample on the one hand, and the differences in standardized item *p*-values between the various population groups on the other. For instance, in their second study, they compared the (standardized) item *p*-values between Black and White South-African students (average item performance 76% and 91%, respectively) and found that these 58 differences in item *p*-values correlated at 0.54 with the 58 item heritabilities that were based on the North-American twin sample. However, similar to the item heritabilities, these group differences in item proportion correct are lawfully related to the *p*-values of the items of the groups from which the latter are drawn. In figure 5, we display the item averages of the group differences between *p*-values across the 55 group comparisons from Rushton *et al.* in relation to the *p*-values within the twin sample. Again, the relation is clearly quadratic and items with *p*-values approximately 0.50 showing the largest between-group differences. The quadratic trend explains 78 per cent of the variance. This trend is hardly surprising because the relatively easy items have high proportions correct even in the lower-scoring samples, and so cannot show large group differences in proportions correct (either standardized or not). Therefore, the correlation between the item heritabilities and the group differences occurred because both of these vectors are lawfully quadratically related to the item *p*-values in the twin group. Items with *p*-values around 0.50 in the twin sample show the largest heritability and also show the largest between group differences in *p*-values across the groups because these items lie near the middle of the ability spectrum in the adult twin sample which was used as the base.

In figure 6, we display our simulated data, in which the same thing occurs. The item *p*-values from one of the three groups and differences in standardized *p*-values between groups are related, but certainly not in a linear way. This can be seen clearly in the nine scenarios displayed in figure 6. Figure 6*a*(i)–*c*(i) shows group differences in standardized item *p*-values between the low group and the middle group. Figure 6*a*(ii)–*c*(ii) displays group differences in low and high groups, and figure 6*a*(iii)–*c*(iii) displays group differences in the middle and high groups. The three rows show the results based on the item *p*-values in the low group (figure 6*a*(i–iii)), middle group (figure 6*b*(i–iii)) and high group (figure 6*c*(i–iii)). The shapes of the relations vary, even within rows and columns of the figure, but they are generally in the shape of an inverted U, especially for the middle column. Note that the twin sample in Rushton *et al.*'s (2007) second study is comparable with the middle group in the simulation.

Rushton *et al.* modelled their data by computing *linear* correlations between the item heritabilities and group differences in item *p*-values. To replicate their results, we also plotted the ITCs from the three groups against the standardized group differences in item *p*-values and found strongly nonlinear relations in various shapes. The linear correlations on the basis of the simulated data varied widely from −0.41 to 0.91. On average, the correlation equalled $r=0.30$, which is analogous to the correlations reported by Rushton *et al.* It is important to note that we did not impose any structure on the data in our simulations apart from the basic psychometric relation between item scores and trait levels. The correlations between these variables have therefore no substantive meaning besides the psychometric meaning that the item scores in the different groups related because they are based on the same items. Rushton *et al.* (2007) found generally positive correlations between item heritabilities and group differences in item proportions correct on the Raven's because both of these variables are related in a similar nonlinear way to the *p*-values of items in the twin sample.

Whatever the sample and source of item heritabilities, they will correlate with group differences in item proportions correct even when they come from completely different tests and samples, as long as the items contributing to both are ordered by difficulty and the test is not of extreme difficulty or lack of difficulty for the sample used as the base. The correlation occurs because it is the effect of item *p*-value on both item heritabilities and group differences that create the correlation between them. Two examples should make this clear.

First, when the correlations reported by Rushton *et al.* (2007) are appropriately adjusted for item variance[4] and ITC, the mean correlation is reduced from 0.21 ($p=0.055$, one-tailed)[5] to −0.02 (ns). No single correlation in any comparison group was significant. Six of the ten were negative and four were positive. Second, we took the first 58 items from Tellegen's multidimensional personality questionnaire (MPQ; Tellegen 2005) as measured in the Minnesota twin registry (Lykken *et al.* 1990; Krueger & Johnson 2002). In these data, the MPQ consisted of 300 dichotomous items that fall into 11 personality scales, but the items from each scale are distributed randomly throughout the questionnaire. Therefore, the first 58 items do not measure any one of these 11 personality factors, but they vary in item 'difficulties' keyed across the full range from 0 to 1. The mean correlation between the heritabilities of these items

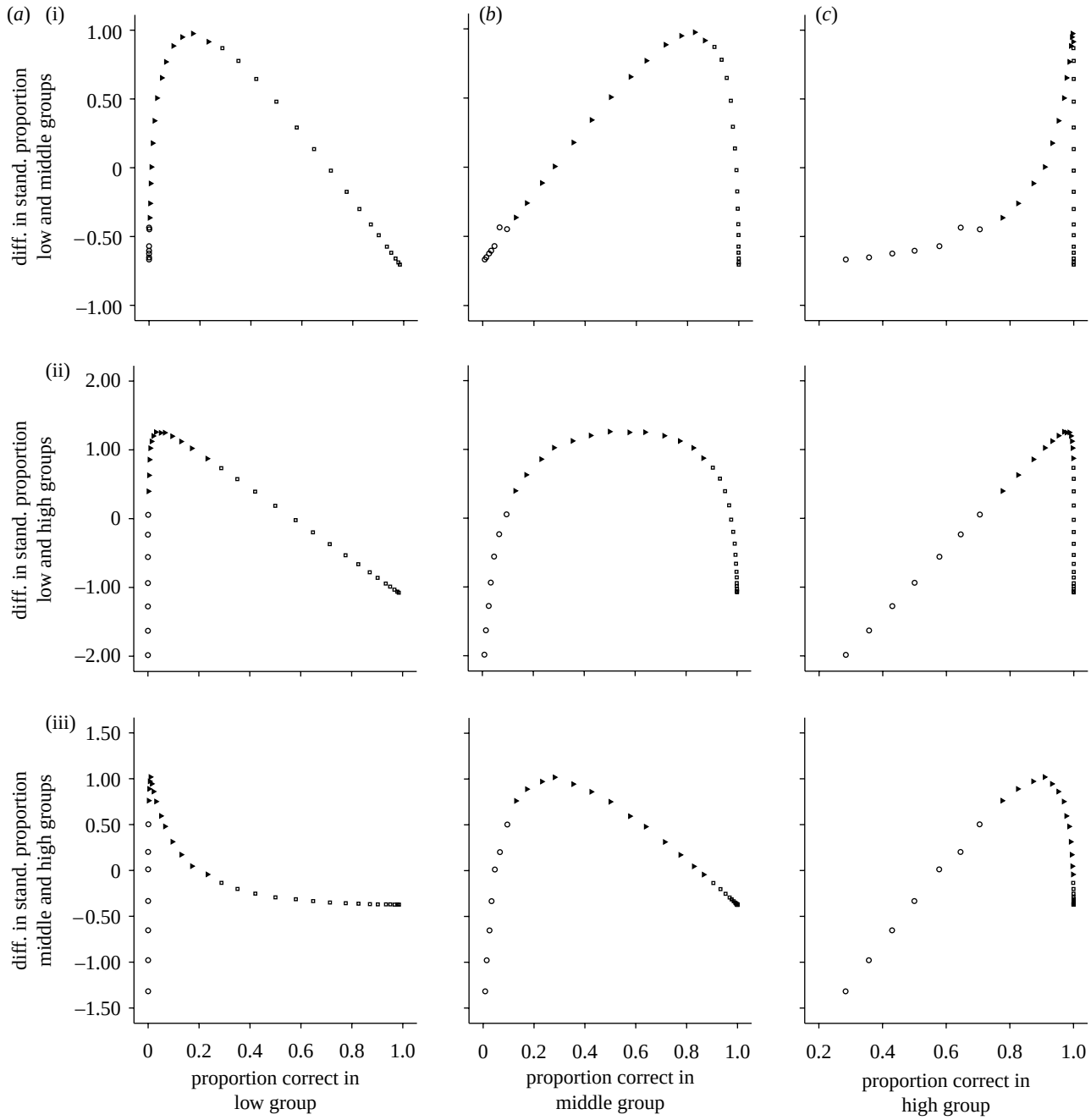

Figure 6. Relation between *p*-values and group differences in standardized *p*-values in the simulation. (Difficulty of item: squares, easy; triangles medium; circles, hard.) (*a*) Proportion correct in low group; (*b*) proportion correct in middle group; (*c*) proportion correct in high group. (i) Comparison of low and middle group; (ii) comparison of low and high group; (iii) comparison of middle and high group.

and the group differences in Raven's proportions correct reported by Rushton *et al.* was 0.20, which is substantial considering that the highest ITC in the group of MPQ items was 0.24 and the average item proportion correct was 0.53, making the MPQ items roughly representative of the low group in our simulation. Corroborating this, the correlations between MPQ item heritabilities and group differences in item difficulties in the two groups with average item proportions correct similar to that from the MPQ items were 0.36 and 0.47, both p's < 0.01. Thus, the correlations between the item heritabilities and the group differences in Rushton *et al.*'s study are manifestations of the psychometric functioning of items in the twin samples, but have no implications for the nature of the group differences in the underlying trait.

4. CONCLUSIONS

Group differences in test score heritability may be due to group differences in measurement precision that arise because of mean group difference in ability. Even when the items of a test function equivalently across groups, these items are not equally informative for different levels of the trait. Hence, if groups differ in mean ability, differences in difficulty of test items will result in differences in the degree to which the test items will be informative for that particular group. The test characteristic curves of most published tests are not linear (unless the test consists of great many items that are uniformly distributed across the entire range of abilities) and so tests are not equally precise in measuring the trait at all its levels. For that reason, differences in measurement

precision may result in spurious heritability estimates on the basis of poor measurement. As we have shown here, this problem is particularly vexing at the item level. If groups differ in mean ability, then the heritabilities computed at the item level in one group cannot be expected to generalize to the other group.[6]

To show the importance of understanding the properties of heritability estimates, we considered a paper by Rushton *et al.* (2007) that concluded that group differences in intelligence test performance were genetic in origin based on analysis of item heritabilities in a single group. We showed that the correlations Rushton *et al.* obtained and used as the basis for their conclusion were inevitable based on the inherent properties of the heritability estimates in the particular group used and the inherent properties of the group differences in item proportions correct. That is, the correlations they obtained were owing to statistical properties of the data rather than to any actual causal link between the genetic influences reflected in the heritability estimates and the group differences. This is an important demonstration of the potential implications of failing to recognize that heritability estimates depend on the population group, trait distribution, measurement validity and environmental circumstances in the sample in question. Although similar problems have been discussed previously (Neale *et al.* 2005; van den Berg *et al.* 2007), the conclusions drawn by Rushton *et al.* (2007) on the generalizability of item heritabilities illustrate that this problem is often not recognized.

We have demonstrated that Rushton *et al.*'s conclusion that the group differences in intelligence test performance they observed were genetic in origin was unwarranted, and that the method they used to draw that conclusion was inappropriate.[7] This does not mean, however, that the group differences were not genetic in origin. Rather, it means that Rushton *et al.*'s analyses shed no light on the question of the origins of the group differences. The correlation between item heritabilities and group differences in item performance does not speak to the nature of group differences in intelligence. Likewise, Rushton *et al.* established a positive correlation between item-test correlations and group differences in item performance and claimed that the Raven's tests measure the same construct across groups. However, this correlation does not establish measurement invariance across the groups. The extent to which the test is measurement invariant across the groups to be considered needs to be rigorously tested (Millsap & Everson 1993). In addition, the heritability of the trait needs to be rigorously established in each population group to be considered, preferably with an approach based on IRT (see van den Berg *et al.* 2007 for an excellent example).

However, at present, the methods available to address the question of the extent of involvement of genetic determinism in group differences in intelligence are not sufficient to resolve it. This is because we do not have sufficient understanding of how genes are involved in intelligence to interpret the heritability statistics we obtain, regardless of the groups within which we obtain them. The techniques used in estimating heritability accurately depend on unidimensionality of the trait and homogeneity of the population and this very dependence precludes their application across population groups that by definition differ on some dimension of interest. As Rushton & Jensen (2005) put it: 'A high heritability within one group does not mean that the average difference between it and another group is due to genetic differences, even if the heritability is high in both groups' (p. 239).

The research by J.M.W. was made possible by VENI grant no. 451-07-016 from the Netherlands Organization for Research (NWO). W.J. holds a Research Council of the UK Fellowship.

ENDNOTES

[1]The first two items were not scored in the data used by Rushton *et al.* leaving 58 items.

[2]Throughout our simulations and discussion, we assumed that the Raven items tap general intelligence in a unidimensional manner, which is an assumption made implicitly in most studies that use the Raven as a measure of intelligence. This meant that we considered all item-specific variance to be random error.

[3]The discrimination parameter was set to two.

[4]Rushton *et al.* claim to adjust for item variance and item reliability, but the formula they use for item variance is incorrect. The item test correlations in Rushton *et al.*'s are based on biserial rather than point-biserial correlations. The biserial correlation is less strongly related to item *p*-values than the point-biserial correlations.

[5]Rushton *et al.* incorrectly reported the *p*-value for this correlation as less than 0.05 without noting whether it was intended to be measured as one- or two-tailed.

[6]The present results are based on item heritabilities that are derived from classic item statistics, such as the intra-class correlation coefficient (as used by Rushton *et al.*) and the phi-coefficient. The tetrachoric correlation coefficient is developed in order to be insensitive to the 'difficulty' of items. However, under the logistic model that we used in our simulation the assumptions underlying the tetrachoric correlation no longer hold, and group differences in the mean ability may also result in group differences in item heritability.

[7]Rushton *et al.* also estimated environmental influences on Raven's performance in the adult twin group and calculated the correlation between item environmentalities and group differences in item proportions correct. Because the average correlation was low and not significant, they rejected the possibility that environmental influences contributed to the group differences (although they were more cautious in their discussion). The method they used is as flawed in this application as in the heritability application, but the environmentalities they calculated were also inaccurate as they reflected item variances alone. As an indication of this, they were positively correlated (0.32), but heritabilities and environmentalities should be negatively correlated as they by definition sum to one.

REFERENCES

Anderson, C. A., Duffy, D. L., Martin, N. G. & Visscher, P. M. 2007 Estimation of variance components for age at menarche in twin families. *Behav. Genet.* **37**, 668–677. (doi:10.1007/s10519-007-9163-2)

Bechger, T. M., Maris, G., Verstralen, H. H. F. M. & Beguin, A. A. 2003 Using classical test theory in combination with item response theory. *Appl. Psychol. Meas.* **27**, 319–334. (doi:10.1177/0146621603257518)

Embretson, S. E. & Reise, S. P. 2000 *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Flynn, J. R. 2007 *What is intelligence?* Cambridge, UK: Cambridge University Press.

Heath, A., Berg, K., Eaves, L. J., Solaas, M. H., Corey, L. A., Sundet, J., Magnus, P. & Nance, W. E. 1985 Education policy and the heritability of educational attainment. *Nature* **314**, 734–736. (doi:10.1038/314734a0)

Krueger, R. F. & Johnson, W. 2002 The Minnesota twin registry: current status and future directions. *Twin Res.* **5**, 488–492. (doi:10.1375/136905202320906336)

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

Lubke, G. H., Dolan, C. V. & Neale, M. C. 2004 Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction. *Twin Res. Hum. Genet.* **7**, 292–298. (doi:10.1375/twin.7.3.292)

Lykken, D. T., Bouchard Jr, T. J., McGue, M. & Tellegen, A. 1990 The Minnesota twin family registry: some initial findings. *Acta Genet. Med. Gemellol.* **39**, 35–70.

Mellenbergh, G. J. 1989 Item bias and item response theory. *Int. J. Educ. Res.* **13**, 127–143. (doi:10.1016/0883-0355(89)90002-5)

Mellenbergh, G. J. 1996 Measurement precision in test score and item response models. *Psychol. Methods* **1**, 293–299. (doi:10.1037/1082-989X.1.3.293)

Millsap, R. E. & Everson, H. T. 1993 Methodology review: statistical approaches for assessing measurement bias. *Appl. Psychol. Meas.* **17**, 297–334. (doi:10.1177/014662169301700401)

Neale, M. C., Lubke, G., Aggen, S. H. & Dolan, C. V. 2005 Problems with using sum scores for estimating variance components: contamination and measurement noninvariance. *Twin Res. Hum. Genet.* **8**, 553–568. (doi:10.1375/twin.8.6.553)

Raven, J. 2008 The Raven progressive matrices tests: their theoretical basis and measurement model. In *Uses and abuses of Intelligence. Studies advancing Spearman and Raven's quest for non-arbitrary metrics* (eds J. Raven & J. Raven), pp. 17–68. Unionville, NY: Royal Fireworks Press.

Raven, J. C. 1941 Standardization of progressive matrices, 1938. *Br. J. Med. Psychol.* **19**, 137–150.

Raven, J. C., Court, J. H. & Raven, J. 1996 *Manual for Raven's progressive matrices and vocabulary scales*. Oxford, UK: Oxford Psychologists Press.

Rushton, J. P. & Jensen, A. R. 2005 Thirty years of research on race differences in cognitive ability. *Psychol. Public Policy Law* **11**, 235–294. (doi:10.1037/1076-8971.11.2.235)

Rushton, J. P., Bons, T. A., Vernon, P. A. & Cvorovic, J. 2007 Genetic and environmental contributions to population group differences on the Raven's progressive matrices estimated from twins reared together and apart. *Proc. R. Soc. B* **274**, 1773–1777. (doi:10.1098/rspb.2007.0461)

Tellegen, A. 2005 *Manual for the multidimensional personality questionnaire*. Minneapolis, MN: University of Minnesota Press.

van den Berg, S., Glas, C. & Boomsma, D. I. 2007 Variance decomposition using an IRT measurement model. *Behav. Genet.* **37**, 604–613. (doi:10.1007/s10519-007-9156-1)

van der Ven, A. H. G. S. & Ellis, J. L. 2000 A Rasch analysis of Raven's standard progressive matrices. *Pers. Individ. Dif.* **29**, 45–64. (doi:10.1016/S0191-8869(99)00177-4)