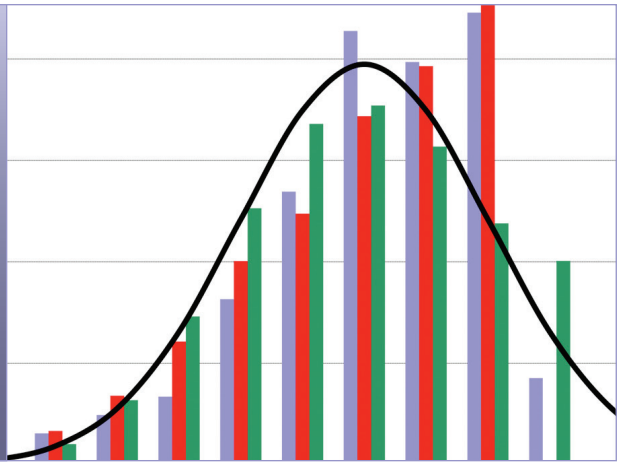
ACT Research
Report Series
2012 (8)



Relationships of Examinee Pair Characteristics and Item Response Similarity

Jeff Allen

October 2012

ACT®

For additional copies, write:
ACT Research Report Series
P.O. Box 168
Iowa City, IA 52243-0168

© 2012 by ACT, Inc. All rights reserved.

Relationships of Examinee Pair Characteristics and Item Response Similarity

Jeff Allen

Table of Contents

Abstract

Detecting unusual similarity in the item responses of a pair of examinees usually conditions on the pair's overall test performance (e.g., raw scores). Doing this, however, often requires assumptions about the invariance of other examinee pair characteristics. In this study, we examined the appropriateness of such assumptions about selected examinee pair characteristics using item response data from the ACT Assessment®. We found that the number of identical incorrect responses on multiple choice tests were slightly higher for same high school, female, and twin pairs. Pairs who took the same mathematics courses and received the same grades had slightly higher levels of identical incorrect responses on the ACT Mathematics test. Across the four ACT tests, twin pairs were about twice as likely to have a similarity level described as a one-in-a-thousand event, all other shared characteristics held constant. The results for twin pairs can be used to set an upper bound for the level of similarity in identical incorrect responses that is due to examinee's shared environment and academic experiences.

Acknowledgements

I thank Steve Fleming, Chi-Yu Huang, and Jim Sconing for their review and helpful suggestions on this paper. Special thanks to Karen Zimmerman for her thorough review and great work formatting the paper. An earlier version of this paper was presented at the Conference on Statistical Detection of Potential Test Fraud held May 23-24, 2012 at the University of Kansas.

Relationships of Examinee Pair Characteristics and Item Response Similarity

Introduction

Cheating on standardized tests undermines the accurate measurement of educational achievement. Students who cheat do not receive accurate information about their areas of academic need, and would be less likely to receive the extra help they need. Students who do not cheat are harmed by those that do, especially when the test is used for high-stakes purposes, such as college admissions or scholarship opportunities.

Given the importance of standardized testing for diagnosing students' academic needs, identifying academic programs appropriate to their achievement, and measuring aggregate school performance for accountability purposes, methods for detecting cheating are increasingly important. Two forms of cheating by students have been the subject of most of the research on cheating detection methods: copying and impersonation (Angoff, 1974). Copying on multiple choice tests will result in highly similar or identical item responses. In particular, the frequency of identical incorrect responses for a pair of examinees is often used as the basis for measures of response similarity (Angoff, 1974; Bellezza & Bellezza, 1989; van der Linden & Sotaridona, 2004).

For a group of examinee pairs, the range of possible identical incorrect responses depends on each examinee's number of correct responses on the test (e.g., raw scores). Moreover, other properties of the frequency distribution of identical incorrect responses (e.g., the mean, median, and standard deviation) are driven by raw scores. Thus, measures of response similarity based on identical incorrect responses must account for raw scores, either directly or indirectly.

Beyond raw scores, it is also possible that other factors affect response similarity. For example, one might expect that two students who had taken the same courses from the same

teachers throughout their K-12 education would be more likely to have answered the same items incorrectly and also to have chosen the same incorrect response options. More generally, it is prudent to investigate whether examinees with shared life experiences, academic backgrounds, cultures, or demographic characteristics have higher average response similarities.

In this study, we examine the extent that selected characteristics of examinee pairs predict level of item response similarity. We address the following research questions:

1) Do examinee pairs from the same high school have greater response similarity than those who attend different high schools?
2) Do pairs who took the same courses (in English, mathematics, and science) and earned the same grades in those courses have greater response similarity on the test (in the subject area) than those who did not?
3) Do pairs from the same racial/ethnic group have greater response similarity than those who are from different racial/ethnic groups?
4) Do same-gender pairs have greater response similarity than pairs of different gender?
5) Do twins have greater response similarity than others?

Addressing these research questions has the potential to lead to enhancements in cheating detection methods that properly account for examinee pair characteristics that are naturally associated with higher response similarity. The research will help developers of cheating detection methods understand if or how examinee pair characteristics need to be considered when assigning probability values to similarity measures.

Methodology

Data from the ACT Assessment program (ACT, 2006) were used for this study. Since 1959, the ACT Assessment has been used as a college admissions test. The ACT Assessment

includes multiple choice tests in English (75 items), mathematics (60 items), reading (40 items), and science (40 items), as well as an optional writing test. The multiple choice items in English, reading, and science have four response options (three incorrect options); the items on the mathematics test have five response options (four incorrect options).

Because recent educational reforms have set college and career readiness as the end goal of K-12 education, many school, districts, and states across the U.S. view adequate performance on the ACT as the appropriate goal for all students. Among U.S. high school graduates of 2011, 49% took the ACT and the ACT was taken by virtually all students in eight states (ACT, 2011). The ACT data were used for two general purposes: 1) constructing large “norm groups" for obtaining distributions of identical incorrect responses (IIR) where copying could not have occurred, for each combination of examinee pair raw scores and for each ACT subject test (English, mathematics, reading, and science); and 2) constructing a large sample of examinee pairs to determine which pair characteristics predict response similarity. Next, the data sets used for these two purposes are described in detail.

Development of Norms for Identical Incorrect Responses.

One way to quantify an examinee pair’s level of response similarity is to first establish a distribution of the similarity measure for a sample of examinee pairs for which copying could not have occurred. This can be done by pairing examinees (who used the same test form) from different testing centers (c.f., Angoff, 1974). Data from the ACT Assessment Program (AAP) from the 2005-2006 through the 2010-2011 testing years was used to develop norm distributions of IIR.

The norms were intended to be used for measuring level of response similarity when situations of suspected copying during ACT administrations occurred. Examinees who attended

the same high school and tested with the same form on the same date, but who tested at different test centers, were paired. This design ensures that the norms are not contaminated by cases of copying, and also controls for any effects of shared high school.

Item responses that were blank or double-gridded were not counted in the computation of IIR for each examinee pair. For certain raw score ranges, we calculated separate frequency distributions of IIR *for each possible combination of raw scores*. For example, for the ACT mathematics test, we considered raw scores between 20 and 59 (a raw score of 60 is perfect and would not lend itself to evidence of copying based on IIR). There are 820 possible raw score combinations for the 40 raw scores between 20 and 59. Table 1 summarizes the structure of the norms for IIR for the four ACT multiple choice tests. Despite having separate norms for every possible raw score combination, the IIR norms are based on very large numbers of examinee pairs, owing to the large volume of ACT testing and the accumulation of data across six years.

Table 1

Summary of Norm Distributions for Identical Incorrect Responses

Test	Number of items	Raw score range	No. raw score combinations	Sample size per raw score combination		
				Min	Median	Max
English	75	30-74	1,035	8,734	85,221	214,175
Math	60	20-59	820	8,171	113,524	227,598
Reading	40	15-39	325	48,749	299,153	602,079
Science	40	15-39	325	12,997	264,567	748,990

The norms for IIR provide a way to quantify an examinee pair's level of item response similarity. For example, a nonparametric estimate of the probability of having an IIR value as or more extreme as an observed value K, assuming a certain combination of raw scores, is obtained from the cumulative norm frequency distribution for that set of raw scores. We refer to this probability as the IIR p-value. This p-value does not condition on test form, and so assumes that

the p-value is invariant to test form or that further analysis of the p-values will account for possible test form effects.

Figure 1 below is the norm frequency distribution of IIR for the ACT mathematics test for pairs of examinees who each had a raw score of 20. Because the ACT mathematics test has 60 items, the maximum possible IIR value is 40 (the minimum possible is 0). The distribution of IIR in this case is bell-shaped with a mode of 8, and at first glance the normal distribution appears to fit the distribution very well.

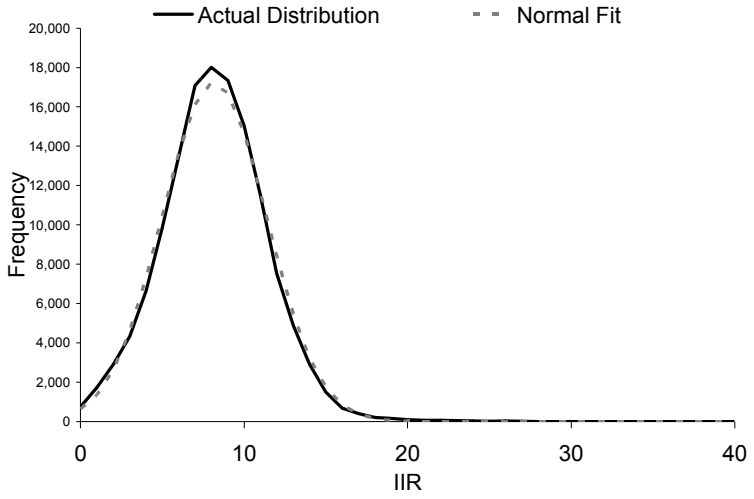


Figure 1. Mathematics norm distribution of IIR for examinee pairs with raw scores of 20.

However, close inspection of the right-hand tail of the distribution (Figure 2) reveals that the normal distribution does not fit well in the extremes of the distribution, which is most important when the goal is to accurately quantify the probability of a possibly extreme level of item response similarity. This phenomenon is not unique to the mathematics test or this particular raw score combination, but rather occurs frequently across the various norm

distributions.[1] For this reason, to identify cases of unusually high IIR we used p-values based on the cumulative probability density functions of IIR from the norm distributions. Again, the IIR norms used examinee pairs from the same high school who tested at different test centers.

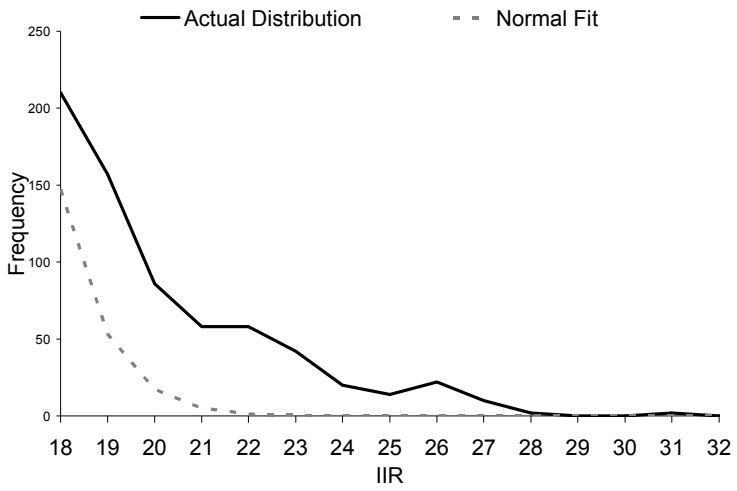


Figure 2. Tail of the mathematics norm distribution of IIR for examinee pairs with raw scores of 20.

Next, we describe the steps taken to construct a data set of examinee pairs with selected shared characteristics.

Construction of Examinee Pair Data Set

We used AAP data from the 2010-2011 testing year to construct a data set of examinee pairs with selected shared characteristics. All examinee pairs tested on the same day with the same test form. In order to apply the IIR norm distributions, we required all examinees to have raw scores on each test within the ranges covered by the norms (30-74 for English, 20-59 for mathematics, 15-39 for reading, and 15-39 for science).

[1] For example, in mathematics, there are 820 norm distributions (one for each raw score combination). Across these 820 distributions, the median percentage of IIR values that are greater than 3 standard deviations from the mean is 0.54%; under the normality assumption, we'd expect this percentage to be 0.13%. The same median percentages for English (0.25%), reading (0.42%), and science (0.40%) are also greater than what would be expected under the assumption that IIR is normally distributed.

To properly identify shared characteristics of examinee pairs, we deleted records for examinees with unreported race/ethnicity, those belonging to two or more racial/ethnic groups, those with unreported gender, and those who did not report course grades for a sufficient number of courses (we required 2 or more reported course grades for English, 3 or more for mathematics, and 2 or more for science). Because we sought to identify twins, we also deleted records for examinees who did not report their last name, first name, date of birth, address, phone number, or social security number. Twins were identified as having the same last name, same date of birth, same phone number, same address, same high school, different first name, and different SSN.

Several data sets were created using various rules for joining examinee pairs together that resulted in large samples of examinee pairs with various shared characteristics. For nearly all characteristics studied, the resulting examinee pairs could be from the same high school (or not) and from the same testing center (or not). An exception to this was the twin pairs, who we required to have the same high school code to maximize our confidence in accurately identifying twins. For most data sets, random sampling of students (among those available for joining in an examinee pair) had to be used to keep the data set manageably small. Again, an exception was the data set for twin pairs – there were too few twin pairs to warrant the need for random sampling before creating the pair data set.

After creating each data set with the purpose of ensuring a large number of pairs with each shared characteristic, we merged the data sets and removed duplicate pairs (pairs may have been joined by virtue of multiple shared characteristics). Each examinee pair could have multiple shared characteristics. To ensure that the effects of shared characteristics could be identified statistically, we also kept over 647,000 pairs with no shared characteristics, except for

possibly test center or high school. (Note that many of the pairs with no shared characteristics were obtained from the data sets created for same high school or same test center pairs). Because we sought to study item response similarity levels that arise naturally (not due to copying), we then removed examinee pairs from the same test center who exhibited such high levels of item response similarity so as to raise serious concerns of copying.[2] The rate of having such high similarity so as to raise concerns of copying was about 3 cases per 10,000 pairs of examinees. It is possible that some cases of no copying were also removed, which would bias the rates of unusually high similarity downwards. We sought to minimize this problem by only removing examinee pairs who exhibited high levels of similarity on multiple tests.

The process described above resulted in a data set with nearly 4.5 million pairs of examinees. Table 2 provides the number of examinee pairs for each shared characteristic, as well as the percentages of those pairs who shared high school and test center.

[2] Examinee pairs were removed if a) their IIR value exceeded the norm group maximum for two or more tests or b) the sum of their four *d* statistics (defined as the number of standard deviations from the norm group mean) exceeded 10 or c) the sum of any two of their four *d* statistics (note that there are six such sums) exceeded the maximum observed among examinee pairs from different test centers.

Table 2

Summary of Examinee Pair Data Set

Shared Characteristic	N pairs	% same high school	% same test center	Definition
High school	2,150,664	100%	75%	Same high school attended
Test center	2,722,990	59%	100%	Tested at the same center
English courses and grades	924,165	69%	69%	Same English courses taken with same grades earned
Math courses and grades	462,059	84%	72%	Same mathematics courses taken with same grades earned
Science courses and grades	564,605	78%	71%	Same science courses taken with same grades earned
Race/ethnicity				Same reported race/ethnicity
African American	95,417	53%	55%	
Caucasian	2,329,829	60%	73%	
Hispanic	174,800	56%	59%	
Gender				Same reported gender
Female	1,427,638	49%	60%	
Male	885,888	49%	61%	
Twins	5,791	100%	98%	Same last name, same date of birth, same phone number, same high school, same address, different first name, different SSN
None	647,245	20%	41%	None of the characteristics listed above in common
Total	4,486,908	48%	61%	

For each examinee pair, we calculated IIR for each of the four ACT multiple choice tests. Using the IIR norms, we assigned to each IIR value the IIR p-value (the proportion of examinee pairs in the norm group that exceeded the IIR value). We also assigned a statistic measuring deviation from the norm group mean, defined as $d = \frac{IIR - \mu}{\sigma}$ where μ is the norm group mean and σ is the norm group standard deviation.

The IIR p-value was used to classify IIR values according to the likelihood of observing high IIR values by chance. P-value classifications of $p<0.01$, $p<0.001$, $p<0.0001$, and $p=0$ were

considered. (Note that p=0 does not mean that the IIR value is impossible, rather it means that the IIR value was larger than any observed in the norm group for examinee pairs with the same raw score combination). We examined whether the occurrence of extreme p-values varied according to the examinee pairs' shared characteristics. Similarly, the *d* statistic was used to examine whether the center of the IIR distribution varied by shared characteristic.

Statistical Modeling

Statistical analyses were conducted to assess the extent that identical incorrect item response similarity varied by examinee pair characteristics. Predictor variables were coded for each possible shared characteristic. For each predictor, we present descriptive statistics for each similarity outcome (mean and standard error of *d* and proportion with small p-value classifications of $p<0.01$, $p<0.001$, $p<0.0001$, and $p=0$). All analyses were conducted for each ACT multiple choice test.

Linear regression was used to examine how the predictors (shared characteristics) explained variance in *d* and logistic regression was used to examine how the predictors explained the probability of having small p-value classifications ($p<0.01$, $p<0.001$, and $p<0.0001$). In the regression models, all predictor variables were included simultaneously so as to estimate the unique effect on similarity of each shared characteristic. Thus, the coefficients estimated from the linear regression model represent estimates of the conditional mean *d* value for each shared characteristic. Likewise, the coefficients from the logistic regression models represent estimates of the conditional log-odds of having an unusually high IIR value. In addition to the indicators for each shared characteristic, the regression models used test form as a nominal predictor variable to control for any test form effects on similarity.

Distinguishing Naturally Occurring Similarity from Copying

In creating the analysis data set, we removed examinee pairs from the same test center who exhibited such high levels of item response similarity to raise serious concerns of copying. However, because the flagging methods employed are unlikely to catch all cases of copying, this does not ensure that the resulting data set (and rates of unusually high IIR across shared characteristics) only measures naturally occurring similarity. In other words, we cannot determine with certainty whether higher rates of unusually high IIR are due to shared characteristics themselves (and hence are naturally occurring) or are due to a higher preponderance of copying within certain shared characteristic subgroups. To help address this problem, we repeated the regression analyses described above, using only the examinee pairs who tested at different test centers, and hence would have been unable to participate in traditional modes of copying. As mentioned earlier, removing examinee pairs with unusually high similarity levels also raises the possibility of mistakenly removing non-copiers, which would bias the similarity levels downward.

Results

Mean IIR d Statistics

We first present the simple mean *d* statistics for each shared characteristic. Because examinee pairs could have multiple shared characteristics, and because some shared characteristics are more likely to co-occur with others, the mean *d* statistics are descriptive only. Because of confounding, we cannot conclude with certainty that larger values for mean *d* are only attributed to the shared characteristic in question. Later, we will present results for analyses that attempt to attribute variation in similarity levels specifically to different shared characteristics.

With such large samples of examinee pairs for each shared characteristic, most of the mean *d* statistics are significantly different from zero (Table A1 in the appendix presents means and standard errors). For descriptive purposes, it is most meaningful to examine the size of mean *d*, rather than if it is different from 0. Examinee pairs with no shared characteristics (other than possibly high school and test center) had average *d* values of 0.02, -0.01, -0.06, and -0.08 in English, mathematics, reading, and science, respectively.

The level of similarity for examinee pairs with shared characteristics is generally larger for the ACT English and Mathematics tests, relative to the Reading and Science tests. Among the shared characteristics studied, twin pairs consistently had the largest mean *d*, ranging from 0.18 in science to 0.50 in English. Same high school pairs had higher similarity levels, with mean *d* ranging from 0.02-0.03 in reading and science to 0.12-0.15 in English and mathematics. Pairs from the same test center showed a similar pattern.

Pairs who took the same courses and received the same grades had higher similarity levels, particularly on the English and mathematics tests. However, having the same courses and grades in a particular subject area was not strongly aligned with higher similarity on the same subject area test. For example, students with the same mathematics courses and grades had an average *d* of 0.16 on the mathematics test, but also an average *d* of 0.13 on the English test. Students with the same science courses and grades had an average *d* of 0.03 on the science test, but an average *d* of 0.14 on the mathematics test.

Caucasian pairs generally had higher similarity, again, particularly on the English and mathematics tests. Caucasian pairs tended to have greater similarity than African American or Hispanic pairs, who surprisingly had negative mean *d* values in reading and science, ranging from -0.07 to -0.08. Same-gender pairs also had higher similarity on the English and

mathematics tests, with female pairs having higher similarity on the English test (mean d=0.14 for female, mean d=0.05 for male).

Rates of Unusually High IIR

We now present the rates of unusually high IIR values, by each shared characteristic. We express the rates as number of pairs per one hundred, one thousand, ten thousand, or one hundred thousand for IIR p-value classifications of <.01, <.001, <.0001, and 0, respectively. Again, because examinee pairs could have multiple shared characteristics, and because some shared characteristics are more likely to co-occur with others, these rates are descriptive only. Later, we will present results for analyses that attempt to attribute odds of unusually high IIR specifically to different shared characteristics.

By definition, the expected rates of unusually high IIR should be near 1 for p<0.01, p<0.001, and p<0.0001. In Tables A2-A5 located in the Appendix, we only present rates for shared characteristics when the approximate expected number of unusually high IIR values is at least five. We did this to avoid the situation where only a few extreme cases can have a large influence on the estimated rates. For example, there are 5,791 twin pairs (Table 2). Therefore, we'd expect approximately 58 to have p<0.01, 5.8 to have p<0.001, and 0.58 to have p<0.0001, under the hypothesis that twin pairs are no more likely than norm group pairs to have unusually high IIR. Thus, for twin pairs, rates are presented for p<0.01 and p<0.001, but not for p<0.0001 or p=0.

English. The rates of unusually high IIR for English are presented in Table A2. Nearly all of the rates are less than 1; this is expected because IIR is a discrete random variable and the p-values are defined as the proportion of pairs in the norm data sets that *exceeded* the value in

question.[3] Across shared characteristics, twin pairs had the highest rates of unusually high IIR for the $p<0.01$ and $p<0.001$ criteria (2.33 per 100, 2.24 per 1,000, respectively). Shared characteristics with the highest rates of extreme IIR values ($p=0$) were male gender (0.56 per 100,000), no shared characteristics (0.46 per 100,000), same test center (0.26 per 100,000), and same high school (0.28 per 100,000). It should be noted, however, that differences in the rates of extreme IIR values ($p=0$) should be interpreted with caution. For example, of the 647,245 examinee pairs with no shared characteristics, only three (0.46 per 100,000) had an extreme IIR value. The rates of extreme IIR are extremely sensitive to single occurrences of $p=0$. The rates of unusually high IIR did not appear to be different for examinee pairs that shared courses and grades – the rates for the $p<0.01$, $p<0.001$, and $p<0.0001$ criteria were all near expected.

It should be noted at this point that the rates of extremely high IIR in English are somewhat difficult to interpret because we cannot discern whether the higher rates are due to naturally occurring extreme similarity or a greater preponderance of copying. For example, we see that male pairs are more likely than female pairs to have an IIR result with a p-value of 0 (0.56 versus 0.07 per 100,000). When creating the data set, we removed cases of probable copying, but it is possible that some copiers remained in the study data set. Thus, we do not know if the male-female difference described above exists because males are more likely to copy, because males are more likely to have extremely high similarity that occurs naturally, or even because of possible confounding with the other shared characteristics. Later, we will discuss analyses that attempt to address this problem.

[3] For example, from the norm sample data, we observed that 0.49 examinee pairs per 100 had a p-value (based on the norm sample distribution) below 0.01 in English. The corresponding rates in mathematics, reading, and science are 0.60, 0.45, and 0.52 per 100, respectively.

Mathematics. The rates of unusually high IIR for mathematics are presented in Table A3. Again, across shared characteristics, twin pairs had the highest rates of unusually high IIR for the $p<0.01$ and $p<0.001$ criteria (1.54 per 100, 2.24 per 1,000, respectively). The rates of unusually high IIR in mathematics did not appear to be much different for examinee pairs that shared mathematics courses and grades. Among the shared characteristics with sufficient sample size, same science course pairs had the highest rate below the $p<0.0001$ criterion, with 0.57 pairs per 10,000. The shared characteristics with the highest rates of extreme IIR values ($p=0$) was no shared characteristics (0.93 per 100,000) and male gender (0.68 per 100,000). Again, it cannot yet be determined whether this higher rate reflects undetected cases of copying or naturally occurring similarity.

Reading. The rates of unusually high IIR for reading are presented in Table A4. Twin pairs had the highest rates of unusually high IIR for the $p<0.01$ and $p<0.001$ criteria (1.38 per 100, 1.73 per 1,000, respectively). Among the shared characteristics with sufficient sample size, Hispanic race had the highest rate of pairs below the $p<0.0001$ criterion, with 0.74 pairs per 10,000. There were very few occurrences of the most extreme IIR values ($p=0$). This is likely due to fact that the raw score range for reading (15-39) was shorter than that for English (30-74) and mathematics (20-59). The shorter range of raw scores causes larger norm group sample sizes for each raw score combination (see Table 1), which in turn makes it harder to obtain a p-value of 0 (the chances of getting a p-value of 0 decrease as the norm group sample size increases).

Science. The rates of unusually high IIR for science are presented in Table A5. Twin pairs had the highest rates of unusually high IIR for the $p<0.01$ and $p<0.001$ criteria (1.47 per 100, 1.73 per 1,000, respectively). The rates of unusually high IIR in science did not appear to

be much different for examinee pairs that shared science courses and grades – the rates were slightly higher for pairs that shared mathematics courses and grades. Among the shared characteristics with sufficient sample size, Hispanic race had the highest rate of pairs below the $p<0.0001$ criterion, with 1.32 pairs per 10,000. Shared characteristics with the highest rates of extreme IIR values ($p=0$) were male gender (0.34 per 100,000) and same high school (0.23 per 100,000).

Conditional Mean IIR d Statistics

Earlier, we presented mean IIR *d* statistics for each shared characteristic (Table A1). We observed *d* statistics above what would have been expected (greater than 0) under the hypothesis that the sample of examinee pairs in this study is no different than the sample of examinee pairs in the norm sample with respect to similarity. Because examinee pairs could have multiple shared characteristics, and because some shared characteristics are more likely to co-occur with others, we could not conclude with certainty that the larger *d* values of Table A1 were only attributed to the shared characteristic in question.

We now present conditional mean IIR *d* statistics derived through a multiple linear regression model. For each test, *d* was regressed on the full set of variables indicating different shared characteristics. The resulting parameter estimates (one for each shared characteristic) measure the contribution of each shared characteristic to *d*, conditional on all other shared characteristics. The results of this analysis are presented in Table A6 (in the Appendix).

As one would expect, the conditional mean *d* values presented in Table A6 are often smaller than the unconditional values presented in Table A1. This occurs because including all shared characteristics in the model leaves less variation in *d* explained by each individual characteristic. Across all four tests, the conditional *d* values were greatest for the twins, with

coefficients ranging from 0.13 (science) to 0.31 (English). High school was another shared characteristic consistently associated with greater similarity levels, with coefficients ranging from 0.04 (reading) to 0.10 (mathematics).

Same test center was associated with greater similarity levels, with coefficients ranging from 0.02 to 0.06. Examinee pairs with the same courses and grades in mathematics had greater similarity on the mathematics test (mean conditional d = 0.03). Examinee pairs with the same courses and grades in English and science had smaller increases in similarity levels.

Female pairs had greater similarity across the four tests (mean conditional d ranging from 0.01 to 0.06), though again the differences were small. Results were inconsistent across racial/ethnic groups. Caucasian had greater similarity levels; while African American and Hispanic pairs had smaller similarity levels. The differences with the largest magnitudes were observed for Hispanic in mathematics (mean conditional d=-0.09), African American in mathematics and reading (mean conditional d=-0.06), and Caucasian in English and mathematics (mean conditional d=0.07).

The multiple regression models used to obtain the conditional d statistics were repeated, but using only the examinee pairs from different test centers (Table A7 in the Appendix). The results are very similar for the two analyses. One notable difference occurred for twins. When only the twins from different test centers were examined, the conditional d statistics dropped noticeably. In English, d dropped from 0.31 to 0.12, in mathematics from 0.15 to 0.09, and in reading from 0.16 to 0.01. In science, d increased from 0.13 to 0.19. Unfortunately, only 136 of the 5,791 twins tested at different centers, so the standard errors for twins in Table A7 are large.

Conditional Odds Ratios for Unusually High IIR

Earlier, we examined rates of unusually high IIR for each shared characteristic (e.g., Table A2). Now, we extend that analysis by examining odds ratios for each shared characteristic, conditioning on all other shared characteristics. The odds ratios measure the percent increase in the probability of having an unusually high IIR, attributed to each shared characteristic. For example, an odds ratio of 1.50 for the $p<0.001$ criterion suggests that the odds of having a one-in-a-thousand similarity level are 1.50 times higher (50% higher) for having the shared characteristic in question, regardless of other shared characteristics. For simplicity, we use the terms *odds* and *probability* interchangeably. (When event probabilities are very small, such as is the case here, event odds and probability are approximately equal).

One-in-a-hundred events. We first present the results for the $p<0.01$ criterion (Table A8 in the Appendix). For twins, the probability increased by 187%, 89%, 170%, and 134% for English, mathematics, reading, and science, respectively. The probability also increased significantly for pairs from the same high school, with probability increases ranging from 16% to 41% across the four tests. Pairs from the same test center also had increased probability of $p<0.01$ similarity, with odds ratios ranging from 1.07 to 1.17.

Pairs with the same mathematics courses and grades had greater odds of $p<0.01$ on the mathematics test (11% increase). Greater odds of $p<0.01$ on the science test were observed for examinee pairs with the same mathematics (9% increase) and science (4%) courses and grades.

African American pairs were more likely to have IIR values with $p<0.01$, with percent increases of 19%, 11%, 29%, and 24% in English, mathematics, reading, and science, respectively. Hispanic pairs were more likely to have IIR values with $p<0.01$ across all four subjects, with percent increases ranging from 27% to 52%. The odds ratios for Caucasian pairs

were statistically significant and greater than one in English, mathematics, and science; but smaller than those for African American and Hispanic students. Female pairs also had higher odds, with probability increases ranging from 12% in reading to 20% in mathematics.

The multiple logistic regression models used to obtain the conditional odds ratios for $p<0.01$ were repeated, but using only the examinee pairs from different test centers (Table A9 in the Appendix). The predictor variable indicating twin pairs was excluded from this analysis because the small sample of twins testing at different test centers (n=136) did not permit stable estimation of the multiple logistic regression parameters. The results for the different-center pairs were mostly very similar to those for all pairs. For example, the odds ratios for $p<0.01$ for high school were 1.16, 1.41, 1.28, and 1.28 for all pairs and 1.13, 1.54, 1.35, and 1.40 for different-center pairs. Notable exceptions occurred for African American and Hispanic pairs. For different-center pairs, the odds ratios for $p<0.01$ were consistently smaller than they were for all pairs for these two racial/ethnic groups. The largest differences occurred for Hispanic pairs in mathematics (odds ratio of 1.48 for all pairs, 1.05 for different-center pairs) and science (odds ratio of 1.52 for all pairs, 1.04 for different-center pairs).

One-in-a-thousand events. For the $p<0.001$ criterion ("one-in-a-thousand" event), twin pairs again had the largest conditional odds ratios (Table A10 in the Appendix). The chances of unusually high IIR similarity were 170%, 300%, 243%, and 163% percent higher for twin pairs (in English, mathematics, reading, and science, respectively), regardless of other shared characteristics. For examinees from the same high school, the probability increased by factors of 1.35, 1.36, 1.47, and 1.65 across English, mathematics, reading, and science, respectively. Examinee pairs from the same test center only had significantly increased odds in English (10% increase).

Statistically significant probability increases were observed for examinee pairs with the same mathematics courses and grades on the mathematics (26% increase) and science (23% increase) tests; other odds ratios for the coursework shared characteristics were smaller or not significant.

Similar to the results for the $p<0.01$ criterion, Hispanic pairs were more likely to have unusually high IIR values using the $p<0.001$ criterion, with percent increases ranging from 72% to 121%. African American pairs had significantly increased odds in English (52%) and reading (68%). The odds ratios for Caucasian pairs were statistically significant in reading and science, but suggested decreased chances of $p<0.0001$.

Relative to the $p<0.01$ criterion, male and female odds ratios were more similar for the $p<0.001$ criterion. The exception was in mathematics, were the chances of unusually high IIR were 29% higher for female-female pairs relative to other pairs.

The multiple logistic regression models used to obtain the conditional odds ratios for $p<0.001$ were repeated for examinee pairs from different test centers (Table A11 in the Appendix). The predictor variable for twin pairs was again excluded from the models because the small sample of twins testing at different test centers did not permit stable estimation of the multiple logistic regression parameters. The results for the different-center pairs appeared mostly similar to those for all pairs. In many cases, odds ratios were not statistically significant, owing perhaps to the smaller sample sizes and the outcome occurring less frequently for the $p<0.001$ criterion. Differences between the results for all pairs and those for different-center pairs were again observed for African American and Hispanic pairs. For different-center pairs, the odds ratios for $p<0.001$ were consistently smaller than they were for all pairs for these two

racial/ethnic groups. The largest differences occurred for Hispanic pairs in science (odds ratio of 1.94 for all pairs, 1.23 for different-center pairs).

One-in-ten thousand events. For the $p<0.0001$ criterion ("one-in-ten thousand" event), the twin pair indicator variable was not included as a predictor in the analysis because of the small sample size of twins and low frequency of outcomes (Table A12 in the Appendix). For examinees from the same high school, the probability increase was not statistically significant in any of the four subject areas, though the confidence intervals for the odds ratios were wide. Similarly, examinee pairs from the same test center did not have significantly increased odds for $p<0.0001$.

Statistically significant probability increases were observed for examinee pairs with the same science courses and grades on the mathematics (67% increase) and reading (93% increase) tests.

Similar to the results for the $p<0.01$ and $p<0.001$ criteria, Hispanic pairs were more likely to have unusually high IIR values using the $p<0.0001$ criterion on the English, reading, and science tests, with percent increases ranging from 142% to 268%. African American pairs had significantly increased odds in reading (187%). Neither male nor female pairs had statistically significant odds ratios for the $p<0.0001$ criterion.

The analyses for the $p<0.0001$ criterion were repeated for examinee pairs from different test centers (Table A13 in the Appendix). Similar to the analysis using all pairs (Table A12), the analysis using pairs from different centers showed increased odds of $p<0.0001$ for Hispanic pairs in English and reading. With a few exceptions, the other odds ratios were not statistically significant. The analysis of the $p<0.0001$ criterion for pairs from different test centers had less

power than other analyses, owing to the smaller sample size and the modeling of such rare events.

Discussion

Revisiting the Research Questions

The study results help us better understand the extent that unusually high similarity levels, as measured by examinee pairs' identical incorrect responses on the multiple choice tests of the ACT Assessment, occur naturally due to shared environments and academic experiences. We now discuss how the study findings address each research question.

1. *Do examinee pairs from the same high school have greater response similarity than those who attend different high schools?*

We found that examinee pairs from the same high school do indeed have greater response similarity levels than those who attend different high schools. All other shared characteristics being held equal, examinee pairs from the same high school have mean IIR values that are between 0.04 and 0.10 standard deviations greater than others. The high school effect on level of similarity appears to be largest for the mathematics and science tests. Examinee pairs from the same high school also showed higher rates of unusually high IIR values. Chances of meeting the $p<0.001$ criterion (one-in-a-thousand event) were 35-65% higher for examinee pairs from the same high school. Surprisingly, the odds ratios for high school were only statistically significant for the $p<0.0001$ criterion when examinee pairs from different test centers were considered. Generally, the high school effect did not diminish when only pairs from different test centers were considered, so we can conclude that the high school effect is not confounded by instances of copying.

We theorize that examinee pairs from the same high school have greater response similarity than those who attend different high schools because of shared academic experiences; many students from the same high school are also likely to have shared elementary and middle schools and live in the same geographical area. Attending the same school (and possibly having the same teachers) may slightly increase the likelihood of choosing the same incorrect response options on the ACT.

2. *Do pairs who took the same courses (in English, math, and science) and earned the same grades in those courses have greater response similarity on the test (in the subject area) than those who did not?*

We found some evidence that examinee pairs who reported taking the same courses and earning the same grades in those courses have greater similarity. However, the evidence was not consistent and often the shared subject did not align with greater similarity on the test in that same subject. Shared mathematics courses and grades seemed to be most related to greater response similarity. For example, the IIR values on the mathematics test were 0.03 standard deviations higher, on average, for pairs with the same mathematics courses and grades. Examinee pairs with the same mathematics courses and grades had greater probability of an unusually high IIR value at the $p<0.0001$ level ("one-in-ten thousand") in reading (41% increase) and science (49%). This is an example of the shared subject not aligning with greater similarity on the test in that same subject.

Again, we theorize that exposure to the same types of courses in high school can lead to small increases in response similarity and small increases in the chances of unusually high IIR values.

3. Do pairs from the same racial/ethnic group have greater response similarity than those who are from different racial/ethnic groups?

All other shared characteristics held equal, African American and Hispanic pairs had lower response similarity levels, on average. This finding is counterintuitive as we expected that all shared characteristics would be associated with at least slightly larger similarity levels, rather than smaller similarity levels. Further work is needed to understand this result. Caucasian pairs, on the other hand, had higher response similarity levels, all other shared characteristics held equal. On average, IIR was between 0.03 and 0.07 standard deviations above the mean for Caucasian pairs.

Contrary to the results from the analysis of the *d* statistics (similarity levels), we found that African American and Hispanic pairs had increased odds of unusually high IIR values (small p-values), all other shared characteristics held equal. In some cases, the odds ratios for African American and Hispanic race/ethnicity were smaller or became non-significant when same-center pairs were removed from the analysis.

4. Do same-gender pairs have greater response similarity than pairs of different gender?

All other shared characteristics held equal, female-female pairs had slightly higher average IIR, ranging from 0.01 to 0.06 standard deviations higher across the four subject tests. Female pairs also had consistently increased odds of unusually high IIR values at the $p<0.01$ criterion, but the female effect was not present or was inconsistent at the other levels of unusual similarity ($p<0.001$ and $p<0.0001$). Male pairs did not show many cases of significantly increased odds, and those that were significant were not consistent across levels of unusual similarity.

5. Do twins have greater response similarity than others?

Of all the shared characteristics studied, twin pairs had the greatest response similarity. This is not unexpected, given that most twin pairs have shared the same home environment and have likely had very similar educational experiences. All other shared characteristics held equal, twin pairs had higher similarity levels in English (mean d=0.31), mathematics (d=0.15), reading (d=0.16), and science (d=0.13). Interestingly, most of these values dropped significantly when twins from the same test center were removed from the analysis. However, with only 136 twin pairs who tested at different centers, it is difficult to speculate on what might have caused the decreases. Twin pairs had significantly increased odds of having unusually high IIR values, with odds ratios for the $p<0.001$ criterion ranging from 2.63 to 4.00. However, the overall rate of unusually high IIR values was still quite small, even for twin pairs. For example, the rate of twin pairs exceeding the $p<0.001$ criterion was 2.24 (per 1,000) in English, 2.24 in mathematics, 1.73 in reading, and 1.73 in science.

Limitations

Despite having an overall sample size of nearly 4.5 million examinee pairs, the study still suffered from some cases of model instability due to data sparseness. This problem resulted because we sought to study extremely unusual events, such as having an IIR value representing a less than one-in-ten thousand event. Also, one of the primary shared characteristics of interest – twin pairs – had just 5,791 cases.

The study relied on students' self-reported data on courses taken and grades earned. Inaccuracies in this data would likely bias the effect estimates towards 0 (e.g., odds ratios of 1.0). This problem is most likely not serious, however, as prior studies have demonstrated a reasonably high degree of accuracy in students' self-reported grades and courses (Sawyer, Laing,

and Houston, 1988). The study would be strengthened by studying other shared educational experiences, such as shared teachers, elementary and middle schools, and instructional programs. Our data set does not enable more detailed measures of shared educational experiences. The study also relied on students' self-reported race/ethnicity and gender.

The inability to completely eliminate cases of copying from the data set used for analyses is another limitation. When cases of copying exist, we cannot discern with certainty whether higher similarity levels occurred naturally due to shared characteristics or whether they are due to a greater preponderance of copying. A possible remedy to this problem is to only include examinee pairs who tested at different centers. We employed this technique, but did not want to rely solely on different-center pairs because we would lose nearly all of the twin pairs. Also, we cannot rule out the possibility that same-center pairs are different than different-center pairs in ways (other than the ability to copy) that affect similarity. For example, 136 of the 5,791 twin pairs tested at different centers. What were the circumstances or factors leading those twins to test at different centers? Could those factors also be related in some way to response similarity?

Another limitation of the analysis is that guessing response patterns were not addressed. When students do not know the answers to test questions or are running out of time, they may resort to guessing. A well-known guessing technique is to respond with all "B" or "C" multiple choice response options. Thus, two examinee pairs that used the same guessing technique would be more likely to have unusually high IIR values. This phenomena could affect the study findings if, for example, students with certain shared characteristics are more likely to use the same guessing technique.

A final limitation of the study is that only identical incorrect responses (IIR) were considered and the normative method for calculating p-values may not have optimal power

related to other procedures, such as those based on the generalized binomial test (Zopluoglu & Davenport, in press) and the ω index (Wollack, 1996). Methods based on identical incorrect responses and identical correct responses have also been used (Saupe, 1960; Sotaridona & Meijer, 2003), and methods are also available that account for all test items (van der Linden & Sotaridona, 2006).

Recommendations

Examinee pair shared characteristics that are related to greater similarity in responses to multiple choice tests should be considered when designing systems for detecting unusually high similarity. Even though we found several cases where shared characteristics were related to greater similarity, the effects were not strong enough to invalidate systems that do not account for shared characteristics. For example, twin pairs had the highest similarity levels, but still their rates of unusually high IIR (e.g., one-in-a-thousand events) were small (approximately two cases per 1,000). This suggests that detection systems that only condition on raw scores are sufficient and need little modification, even when dealing with examinee pairs that share home environment and educational experiences.

The study findings on twin pairs could be used to set upper bounds for the level of similarity that arises naturally, and this information can help stakeholders understand the likelihood that an unusually high similarity level is due to copying (rather than chance or naturally occurring high similarity). For example, by comparing twin pairs to pairs with no shared characteristics (Table 3), we can obtain estimates for the upper bounds of the increase in odds of an unusually high similarity level that is due to shared characteristics. For the one-in-a-thousand criterion, we see that the upper bound odds ratios range from 4.1 (in Science) to 5.9 (in Mathematics).

Table 3

Odds Ratios (95% Confidence Intervals) for Unusual IIR Values, Comparing Twin Pairs to Pairs with No Shared Characteristics

Subject	**Odds Ratio (95% CI)**	
	P<.01 (per 100)	**P<.001 (per 1,000)**
English	4.3 (3.6, 5.1)	4.6 (3.8, 5.5)
Mathematics	3.1 (2.5, 3.8)	5,9 (4.9, 7.0)
Reading	3.5 (2.8, 4.4)	4.5 (3.7, 5.5)
Science	3.4 (2.7, 4.2)	4.1 (3.3, 5.5)

Because same-high school examinee pairs are more likely to have greater similarity in responses to the multiple choice tests of the ACT Assessment, a good design for detecting unusually high similarity levels will use same-school examinee pairs in the comparison group or benchmark sample. Other shared educational characteristics (e.g., same courses and grades) do not appear to be as important and would likely be more difficult to implement in a comparison group design.

References

ACT (2006). *The ACT Technical Manual.* Iowa City, IA: Author.

ACT (2011). *The Condition of College and Career Readiness.* Iowa City, IA: Author.

Angoff, W.H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, *69(345)*: 44-49.

Bellezza, F.S., & Bellezza, S.F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16(3*): 151-155.

Saupe, J. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement, 20*, 475-489.

Sawyer, R., Laing, J., & Houston, M. (1988). Accuracy of self-reported high school courses and grades of college-bound students (ACT Research Report No. 88-1). Iowa City, IA: American College Testing Program.

Sotaridona, L.S., & Meijer, R.R. (2003). *Two new statistics to detect answer copying. Journal of Educational Measurement, 40*, 53-69.

van der Linden, W.J., & Sotaridona, L. (2004). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement*, *41(4)*: 361-377.

van der Linden, W.J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31:* 283-304.

Wollack, J.A. (1996). Detection of answer copying using item response theory. *Dissertation Abstracts International*, 57/05, 2015.

Zopluoglu, C., & Davenport, E.C. (in press). The empirical power and type 1 error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*.

Appendix

Tables A1-A13

Table A1

Mean IIR d Statistics (Standard Error)

Shared Characteristic	English	Mathematics	Reading	Science
High school	0.12 (<0.01)	0.15 (<0.01)	0.02 (<0.01)	0.03 (<0.01)
Test center	0.12 (<0.01)	0.11 (<0.01)	0.01 (<0.01)	0.02 (<0.01)
English courses and grades	0.12 (<0.01)	0.10 (<0.01)	0.01 (<0.01)	0.01 (<0.01)
Math courses and grades	0.13 (<0.01)	0.16 (<0.01)	0.03 (<0.01)	0.04 (<0.01)
Science courses and grades	0.13 (<0.01)	0.14 (<0.01)	0.03 (<0.01)	0.03 (<0.01)
Race/ethnicity				
African American	0.04 (<0.01)	-0.02 (<0.01)	-0.08 (<0.01)	-0.08 (<0.01)
Caucasian	0.13 (<0.01)	0.13 (<0.01)	0.01 (<0.01)	0.01 (<0.01)
Hispanic	0.02 (<0.01)	-0.05 (<0.01)	-0.07 (<0.01)	-0.08 (<0.01)
Gender				
Female	0.14 (<0.01)	0.07 (<0.01)	0.00 (<0.01)	0.00 (<0.01)
Male	0.05 (<0.01)	0.07 (<0.01)	-0.01 (<0.01)	-0.03 (<0.01)
Twins	0.50 (0.01)	0.33 (0.01)	0.19 (0.01)	0.18 (0.01)
None	0.02 (<0.01)	-0.01 (<0.01)	-0.06 (<0.01)	-0.08 (<0.01)

Table A2

Rate of Examinee Pairs below Selected Norm Group P-values: English

Shared Characteristic	Number per 100 below .01 p-value	Number per 1,000 below .001 p-value	Number per 10,000 below .0001 p-value	Number per 100,000 with p-value=0
High school	0.66	0.61	0.44	0.28
Test center	0.65	0.58	0.43	0.26
English courses and grades	0.63	0.57	0.56	0.11
Math courses and grades	0.65	0.53	0.50	.
Science courses and grades	0.65	0.59	0.58	0.18
Race/ethnicity				
African American	0.68	0.78	0.52	.
Caucasian	0.68	0.58	0.39	0.13
Hispanic	0.74	0.91	1.09	.
Gender				
Female	0.71	0.62	0.46	0.07
Male	0.59	0.56	0.60	0.56
Twins	2.33	2.24	.	.
None	0.55	0.50	0.48	0.46

Table A3

Rate of Examinee Pairs below Selected Norm Group P-values: Mathematics

Shared Characteristic	Number per 100 below .01 p-value	Number per 1,000 below .001 p-value	Number per 10,000 below .0001 p-value	Number per 100,000 with p-value=0
High school	0.79	0.55	0.37	0.19
Test center	0.70	0.46	0.32	0.18
English courses and grades	0.70	0.53	0.42	0.43
Math courses and grades	0.82	0.63	0.37	.
Science courses and grades	0.76	0.57	0.57	0.35
Race/ethnicity				
African American	0.65	0.48	0.31	.
Caucasian	0.70	0.46	0.34	0.17
Hispanic	0.86	0.94	0.40	.
Gender				
Female	0.73	0.55	0.39	0.28
Male	0.58	0.43	0.33	0.68
Twins	1.54	2.24	.	.
None	0.50	0.39	0.51	0.93

Table A4

Rate of Examinee Pairs below Selected Norm Group P-values: Reading

Shared Characteristic	Number per 100 below .01 p-value	Number per 1,000 below .001 p-value	Number per 10,000 below .0001 p-value	Number per 100,000 with p-value=0
High school	0.49	0.45	0.30	0.05
Test center	0.45	0.40	0.24	0.04
English courses and grades	0.44	0.37	0.25	0.00
Math courses and grades	0.49	0.49	0.35	.
Science courses and grades	0.48	0.46	0.43	0.00
Race/ethnicity				
African American	0.58	0.75	0.63	.
Caucasian	0.44	0.36	0.22	0.04
Hispanic	0.62	0.73	0.74	.
Gender				
Female	0.47	0.42	0.27	0.07
Male	0.42	0.42	0.26	0.00
Twins	1.38	1.73	.	.
None	0.40	0.39	0.23	0.00

Table A5

Rate of Examinee Pairs below Selected Norm Group P-values: Science

Shared Characteristic	Number per 100 below .01 p-value	Number per 1,000 below .001 p-value	Number per 10,000 below .0001 p-value	Number per 100,000 with p-value=0
High school	0.60	0.60	0.51	0.23
Test center	0.56	0.51	0.42	0.22
English courses and grades	0.55	0.51	0.45	0.11
Math courses and grades	0.61	0.66	0.63	.
Science courses and grades	0.59	0.59	0.50	0.00
Race/ethnicity				
African American	0.62	0.64	0.73	.
Caucasian	0.53	0.47	0.41	0.13
Hispanic	0.80	1.08	1.32	.
Gender				
Female	0.58	0.54	0.43	0.07
Male	0.49	0.53	0.49	0.34
Twins	1.47	1.73	.	.
None	0.44	0.43	0.32	0.00

Table A6

Conditional Mean IIR d Statistics (Standard Error)

Shared Characteristic	English	Mathematics	Reading	Science
High school	0.05 (<0.01)	0.10 (<0.01)	0.04 (<0.01)	0.06 (<0.01)
Test center	0.04 (<0.01)	0.06 (<0.01)	0.02 (<0.01)	0.05 (<0.01)
English courses and grades	0.00 (<0.01)	0.00 (<0.01)	0.01 (<0.01)	0.01 (<0.01)
Math courses and grades	0.02 (<0.01)	0.03 (<0.01)	0.01 (<0.01)	0.02 (<0.01)
Science courses and grades	0.01 (<0.01)	0.02 (<0.01)	0.01 (<0.01)	0.02 (<0.01)
Race/ethnicity				
African American	-0.02 (<0.01)	-0.06 (<0.01)	-0.06 (<0.01)	-0.03 (<0.01)
Caucasian	0.07 (<0.01)	0.07 (<0.01)	0.03 (<0.01)	0.03 (<0.01)
Hispanic	-0.05 (<0.01)	-0.09 (<0.01)	-0.04 (<0.01)	-0.05 (<0.01)
Gender				
Female	0.06 (<0.01)	0.01 (<0.01)	0.02 (<0.01)	0.03 (<0.01)
Male	-0.02 (<0.01)	0.01 (<0.01)	0.01 (<0.01)	0.00 (<0.01)
Twins	0.31 (0.01)	0.15 (0.01)	0.16 (0.01)	0.13 (0.01)

Table A7

Conditional Mean IIR d Statistics (Standard Error), Different Test Centers

Shared Characteristic	English	Mathematics	Reading	Science
High school	0.07 (<0.01)	0.12 (<0.01)	0.05 (<0.01)	0.10 (<0.01)
English courses and grades	0.01 (<0.01)	0.01 (<0.01)	0.02 (<0.01)	0.02 (<0.01)
Math courses and grades	0.03 (<0.01)	0.04 (<0.01)	0.03 (<0.01)	0.03 (<0.01)
Science courses and grades	0.02 (<0.01)	0.03 (<0.01)	0.02 (<0.01)	0.03 (<0.01)
Race/ethnicity				
African American	-0.03 (<0.01)	-0.07 (<0.01)	-0.07 (<0.01)	-0.05 (<0.01)
Caucasian	0.06 (<0.01)	0.06 (<0.01)	0.02 (<0.01)	0.02 (<0.01)
Hispanic	-0.06 (<0.01)	-0.11 (<0.01)	-0.05 (<0.01)	-0.08 (<0.01)
Gender				
Female	0.06 (<0.01)	0.00 (<0.01)	0.01 (<0.01)	0.02 (<0.01)
Male	-0.02 (<0.01)	0.02 (<0.01)	0.01 (<0.01)	0.00 (<0.01)
Twins	0.12 (0.09)	0.09 (0.09)	0.01 (0.09)	0.19 (0.09)

Table A8

Conditional Odds Ratios (95% Confidence Intervals) for IIR Norm Group P-value less than 0.01

Shared Characteristic	English	Mathematics	Reading	Science
High school	1.16 (1.13, 1.19)	1.41 (1.37, 1.45)	1.28 (1.24, 1.32)	1.28 (1.24, 1.32)
Test center	1.07 (1.04, 1.10)	1.12 (1.09, 1.15)	1.09 (1.05, 1.12)	1.17 (1.14, 1.21)
English courses and grades	0.95 (0.92, 0.98)	0.97 (0.94, 1.00)	0.98 (0.95, 1.02)	0.99 (0.96, 1.02)
Math courses and grades	1.01 (0.97, 1.05)	1.11 (1.07, 1.15)	1.04 (0.99, 1.08)	1.09 (1.04, 1.13)
Science courses and grades	0.98 (0.95, 1.02)	1.05 (1.01, 1.08)	1.03 (0.99, 1.08)	1.04 (1.00, 1.08)
Race/ethnicity				
African American	1.19 (1.10, 1.29)	1.11 (1.02, 1.20)	1.29 (1.18, 1.40)	1.24 (1.14, 1.34)
Caucasian	1.16 (1.13, 1.19)	1.10 (1.07, 1.13)	0.99 (0.96, 1.03)	1.03 (1.00, 1.06)
Hispanic	1.27 (1.20, 1.35)	1.48 (1.40, 1.57)	1.43 (1.34, 1.52)	1.52 (1.43, 1.61)
Gender				
Female	1.16 (1.13, 1.19)	1.20 (1.17, 1.23)	1.12 (1.09, 1.16)	1.16 (1.13, 1.19)
Male	1.00 (0.97, 1.03)	0.94 (0.91, 0.97)	1.02 (0.98, 1.06)	1.00 (0.96, 1.03)
Twins	2.87 (2.41, 3.41)	1.89 (1.53, 2.33)	2.70 (2.16, 3.37)	2.34 (1.89, 2.91)

Table A9

Conditional Odds Ratios (95% Confidence Intervals) for IIR Norm Group P-value less than 0.01, Different Test Centers

Shared Characteristic	English	Mathematics	Reading	Science
High school	1.13 (1.07, 1.18)	1.54 (1.47, 1.62)	1.35 (1.28, 1.43)	1.40 (1.33, 1.47)
English courses and grades	0.92 (0.87, 0.97)	0.96 (0.91, 1.01)	0.97 (0.91, 1.03)	1.04 (0.98, 1.11)
Math courses and grades	1.01 (0.94, 1.08)	1.12 (1.04, 1.20)	1.06 (0.97, 1.16)	1.12 (1.04, 1.21)
Science courses and grades	1.00 (0.93, 1.06)	1.04 (0.97, 1.11)	0.99 (0.92, 1.08)	1.02 (0.94, 1.09)
Race/ethnicity				
African American	1.15 (1.02, 1.31)	0.92 (0.81, 1.06)	1.12 (0.97, 1.28)	1.06 (0.93, 1.22)
Caucasian	1.19 (1.14, 1.25)	1.06 (1.01, 1.11)	0.93 (0.88, 0.98)	0.98 (0.93, 1.03)
Hispanic	1.09 (0.99, 1.20)	1.05 (0.95, 1.17)	1.16 (1.04, 1.30)	1.04 (0.93, 1.16)
Gender				
Female	1.14 (1.10, 1.19)	1.18 (1.13, 1.23)	1.10 (1.04, 1.16)	1.18 (1.12, 1.24)
Male	0.98 (0.93, 1.03)	0.96 (0.90, 1.01)	0.99 (0.93, 1.06)	0.93 (0.88, 0.99)

Table A10

Conditional Odds Ratios (95% Confidence Intervals) for IIR Norm Group P-value less than 0.001

Shared Characteristic	English	Mathematics	Reading	Science
High school	1.35 (1.23, 1.47)	1.36 (1.23, 1.51)	1.47 (1.31, 1.64)	1.65 (1.49, 1.82)
Test center	1.10 (1.01, 1.20)	0.90 (0.82, 1.00)	1.04 (0.94, 1.16)	1.04 (0.95, 1.14)
English courses and grades	0.97 (0.87, 1.07)	1.04 (0.94, 1.16)	0.90 (0.80, 1.02)	0.96 (0.87, 1.07)
Math courses and grades	0.89 (0.78, 1.02)	1.26 (1.11, 1.44)	1.20 (1.04, 1.39)	1.23 (1.08, 1.39)
Science courses and grades	1.00 (0.89, 1.13)	1.14 (1.01, 1.29)	1.12 (0.97, 1.28)	1.06 (0.94, 1.20)
Race/ethnicity				
African American	1.52 (1.20, 1.93)	1.08 (0.80, 1.46)	1.68 (1.31, 2.14)	1.16 (0.89, 1.51)
Caucasian	1.08 (0.99, 1.18)	0.97 (0.88, 1.07)	0.81 (0.73, 0.90)	0.89 (0.81, 0.98)
Hispanic	1.74 (1.47, 2.07)	2.21 (1.87, 2.62)	1.72 (1.42, 2.08)	1.94 (1.65, 2.27)
Gender				
Female	1.13 (1.04, 1.24)	1.29 (1.17, 1.42)	1.14 (1.03, 1.27)	1.12 (1.02, 1.23)
Male	1.07 (0.96, 1.19)	1.02 (0.90, 1.15)	1.17 (1.03, 1.32)	1.13 (1.02, 1.27)
Twins	2.70 (1.56, 4.68)	4.00 (2.31, 6.93)	3.43 (1.84, 6.41)	2.63 (1.41, 4.91)

Table A11

Conditional Odds Ratios (95% Confidence Intervals) for IIR Norm Group P-value less than 0.001, Different Test Centers

Shared Characteristic	English	Mathematics	Reading	Science
High school	1.37 (1.18, 1.59)	1.48 (1.26, 1.74)	1.72 (1.44, 2.05)	1.58 (1.35, 1.85)
English courses and grades	0.94 (0.79, 1.12)	0.85 (0.70, 1.04)	0.64 (0.50, 0.82)	0.84 (0.69, 1.02)
Math courses and grades	0.94 (0.74, 1.20)	1.30 (1.03, 1.64)	1.41 (1.09, 1.83)	1.43 (1.14, 1.78)
Science courses and grades	1.09 (0.89, 1.35)	0.88 (0.69, 1.12)	1.04 (0.80, 1.36)	1.06 (0.85, 1.33)
Race/ethnicity				
African American	1.18 (0.79, 1.76)	0.97 (0.63, 1.49)	1.04 (0.68, 1.60)	0.60 (0.36, 1.01)
Caucasian	0.98 (0.85, 1.14)	0.89 (0.76, 1.05)	0.64 (0.53, 0.77)	0.85 (0.73, 0.99)
Hispanic	1.29 (0.96, 1.75)	1.71 (1.30, 2.25)	1.27 (0.92, 1.75)	1.23 (0.91, 1.66)
Gender				
Female	1.10 (0.95, 1.27)	1.21 (1.04, 1.41)	0.91 (0.76, 1.09)	1.08 (0.93, 1.25)
Male	1.11 (0.94, 1.32)	0.90 (0.74, 1.09)	1.25 (1.04, 1.51)	1.01 (0.85, 1.21)

Table A12

Conditional Odds Ratios (95% Confidence Intervals) for IIR Norm Group P-value less than 0.0001

Shared Characteristic	English	Mathematics	Reading	Science
High school	0.92 (0.67, 1.25)	1.16 (0.82, 1.64)	1.38 (0.89, 2.13)	1.17 (0.85, 1.62)
Test center	0.92 (0.69, 1.22)	0.75 (0.55, 1.04)	0.81 (0.54, 1.21)	0.73 (0.53, 0.99)
English courses and grades	1.33 (0.95, 1.84)	1.15 (0.79, 1.68)	0.93 (0.57, 1.50)	1.02 (0.71, 1.45)
Math courses and grades	1.14 (0.73, 1.80)	0.94 (0.56, 1.59)	1.41 (0.81, 2.45)	1.49 (0.99, 2.25)
Science courses and grades	1.36 (0.92, 2.02)	1.67 (1.11, 2.50)	1.93 (1.20, 3.12)	1.11 (0.73, 1.68)
Race/ethnicity				
African American	1.13 (0.46, 2.80)	0.71 (0.22, 2.25)	2.87 (1.20, 6.86)	1.80 (0.82, 3.92)
Caucasian	0.72 (0.53, 0.97)	0.83 (0.59, 1.15)	0.98 (0.63, 1.52)	1.04 (0.75, 1.44)
Hispanic	2.42 (1.47, 3.98)	0.98 (0.45, 2.12)	3.68 (1.94, 6.98)	3.23 (2.01, 5.19)
Gender				
Female	1.00 (0.73, 1.37)	0.92 (0.66, 1.29)	1.18 (0.77, 1.81)	0.96 (0.70, 1.33)
Male	1.38 (0.99, 1.93)	0.82 (0.54, 1.25)	1.19 (0.72, 1.95)	1.10 (0.77, 1.58)

Table A13

Conditional Odds Ratios (95% Confidence Intervals) for IIR Norm Group P-value less than 0.0001, Different Test Centers

Shared Characteristic	English	Mathematics	Reading	Science
High school	1.29 (0.81, 2.05)	0.99 (0.56, 1.77)	1.28 (0.63, 2.62)	1.03 (0.58, 1.82)
English courses and grades	1.28 (0.78, 2.11)	1.31 (0.73, 2.36)	1.69 (0.81, 3.50)	1.83 (1.05, 3.16)
Math courses and grades	0.64 (0.25, 1.60)	0.68 (0.24, 1.92)	0.97 (0.29, 3.26)	1.29 (0.57, 2.93)
Science courses and grades	0.90 (0.44, 1.83)	1.37 (0.65, 2.85)	0.21 (0.03, 1.55)	0.60 (0.24, 1.53)
Race/ethnicity				
African American	1.08 (0.34, 3.46)	2.30 (0.81, 6.49)	1.56 (0.37, 6.64)	1.65 (0.51, 5.37)
Caucasian	0.61 (0.37, 1.01)	0.78 (0.44, 1.41)	0.46 (0.20, 1.05)	0.90 (0.52, 1.56)
Hispanic	2.98 (1.65, 5.37)	2.56 (1.14, 5.75)	3.49 (1.50, 8.09)	1.32 (0.47, 3.71)
Gender				
Female	1.16 (0.75, 1.81)	1.63 (1.01, 2.65)	1.20 (0.62, 2.32)	1.15 (0.68, 1.95)
Male	1.44 (0.89, 2.34)	0.37 (0.15, 0.96)	1.22 (0.57, 2.61)	1.31 (0.73, 2.37)

05020812 0
Rev 1