ANALYZING THE GENDER GAP ON AN ENTRANCE EXAM FOR MATHEMATICALLY TALENTED STUDENTS

J. BUTTERFIELD, H. KEYNES, J. ROGNESS, and J. SUKIENNIK
School of Mathematics Center for Educational Programs, University of Minnesota
Minneapolis, MN 55455
rogness@math.umn.edu

Abstract

We investigate the qualifying entrance exam for the University of Minnesota Talented Youth Mathematics Program (UMTYMP), a five-year accelerated program covering high school- and undergraduate-level mathematics. The exam is used to assess the computational, numerical reasoning, and geometric skills of hundreds of fifth-, sixth-, and seventh-grade students annually. It has accurately identified qualified students in past years, but female participants consistently have had lower overall scores. Based on our belief that they are equally well qualified, in 2011 we began an extensive investigation into the structure and content of the exam to determine the possible sources for these differences. After gathering and analyzing data, we made relatively modest changes in 2012 which essentially eliminated the gender bias on one version of the entrance exam, increasing the percentage of females who qualified. The other unmodified versions in 2012 exhibited the typical gender difference from previous years. We continue to analyze the possible reasons for the gender differences while monitoring the overall student performance upon entering the Program.

Introduction

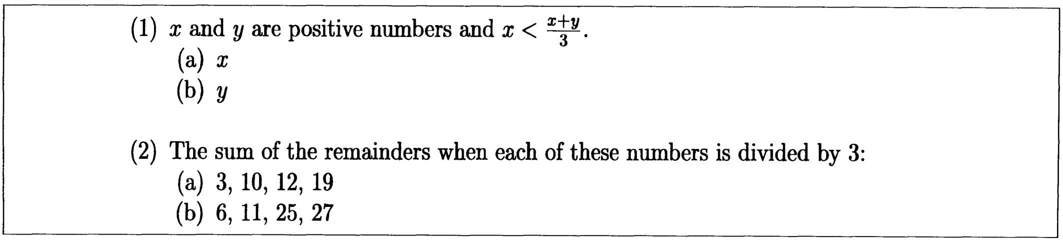
The University of Minnesota Talented Youth Mathematics Program (UMTYMP, pronounced "um-tee-ump") is a highly accelerated program for middle school and high school students who are talented in mathematics. Each year, approximately 500 participants take their mathematics courses through UMTYMP, instead of their regular schools, meeting once per week for two hours. In the first two years of the Program, students cover honors-level algebra, geometry, and precalculus at an accelerated pace; during the following three years, students earn University of Minnesota credits for a sequence of courses covering calculus, linear algebra, multivariable calculus, and vector analysis. Students regularly finish UMTYMP as tenth graders and take upper-division mathematics courses at the University for the remainder of their high school careers [1].

Like many other accelerated mathematics programs, UMTYMP has historically had more male participants than female. In the mid-1990s, a multifaceted intervention funded by the Bush

Foundation resulted in incoming classes with female enrollment of over 30% [2]. Once the funding ended for these costly initiatives, the percentage decreased to between 25% and 30%. Beginning in 2010, this percentage has been rising, and in the 2012-13 academic year 35% of the admitted UMTYMP algebra class was female. These increases are particularly exciting because we do not currently have an intervention program targeting female enrollment. Rather, the results seem to be in large part due to an analysis of our qualifying exam, and subtle changes in the problem content and difficulty levels to make it more gender neutral in identifying the best candidates for the Program. This article discusses our initial efforts and describes which adjustments had an effect on the results.

The Entrance Exam—Testing Process

Students in grades 5-7 who wish to enter UMTYMP must achieve a satisfactory score on an entrance exam developed by our academic staff. The exam covers a variety of concepts in arithmetic, numerical reasoning, mathematical modeling, geometry, and spatial reasoning. In each question, students are given two quantities and must decide if one is always larger than the other, if they are always equal, or if there is not enough information to decide (see Figure 1). The format is based on the Quantitiative Scholastic and College Ability Test (SCAT) used by the Center for Talented Youth at Johns Hopkins. In the early 1980s, the actual SCAT was used to identify potential UMTYMP students. The exam has traditionally been comprised of fifty questions to be answered in twenty minutes, giving students an average of 24 seconds to work on each problem. The purpose of this exam design was for higher scores to indicate the ability to quickly process and understand mathematical concepts that are necessary to be successful in algebra.



(1) x and y are positive numbers and $x < \frac{x+y}{3}$.
(a) x
(b) y

(2) The sum of the remainders when each of these numbers is divided by 3:
(a) 3, 10, 12, 19
(b) 6, 11, 25, 27

Figure 1. Practice questions for the UMTYMP Algebra Entrance Exam—Students must determine the size relationship between the two quantities in each question.

Historically, the passing score has hovered around 40/50, although it has changed at times due to test item analysis or other factors. As part of their registration form, students answer essay questions about their interest in mathematics and UMTYMP; these responses are used as part of the evaluation process, especially for students close to the passing line. In some years, for example, we have admitted all students scoring at least 41, and then a subset of the students with a score of 40 based on their essay answers.

Two entrance exams are given each year: the "Early Exam" is generally held in February, and the "Regular Exam" is given in late March or early April. The Early Exam is part of a larger optional program called UMTYMP Opportunities, which gives students a chance to learn more about the testing process. One week before the Early Exam, students come to campus to work through a series of sample problems with our instructors, culminating in a short practice test. The value of this opportunity is not the mathematical content; rather, we find that exposing students to the testing environment a week before the entrance exam makes them much more comfortable during the actual test. Furthermore, students who do not qualify based on their Early Exam score can register for the Regular Exam later that spring, giving them an extra chance to pass an entrance exam.

The distinction between the Early and Regular Exam pools is very important when evaluating results. The Early testers know more about the exam, and by the very act of enrolling for that exam they have maximized their chances of qualifying. Most Regular Exam testers see the exam format for the first time at the exam sitting. Not surprisingly, both the overall results and the gender breakdown of the scores on the Early Exam are often different than those on the Regular Exam.

Year-to-year comparisons can be tricky even when focusing exclusively on one of these exam pools. We have decades' worth of scores on UMTYMP entrance exams, along with the corresponding transcripts of students who enrolled in the Program, and it is tempting to use this data to make sweeping statements about longitudinal performance. However, experience has taught us that the entrance exam data is highly variable over time due to many factors. In the past, it was common for over 1,000 or even 1,500 students to take the exam. In recent years, our recruiting has become more targeted and we now annually test 600-800 students who score higher, on average, than the students in the past. On a related note, the mathematical climate in Minnesota has changed in the last few years with the introduction of a state mandate that all students take algebra by eighth grade. This has pushed more pre-algebra curriculum into earlier grades, which means our current testing pools are likely better prepared for an algebra course than our pools from even five years ago. Long-term comparisons of entrance exam data are therefore difficult. On a shorter time scale, we have surmised that the testing pool from one year to the next would be relatively stable, but even this assumption may be tenuous.

The Entrance Exam—Effectiveness

It is reasonable to ask whether this process is the best way to identify potential UMTYMP students. For example, although the pace of our course requires students to be able to process mathematics quickly, there is no particular data-driven reason that students should have twenty-four seconds to answer each entrance exam question as opposed to twenty seconds, or thirty-five. The main reason we have continued to use this entrance exam is that for decades it has proven to be highly effective in identifying students who are capable of succeeding in the Program.

In 2011-12, for example, 142 students enrolled in our algebra course after passing the exam, and all but one of them did well enough to continue on to the second year of the Program. Overall, of the approximately 500-600 students registered in the entire Program each year, the number of students whose grades are too low to continue is generally ten or fewer. This group includes students who are capable of succeeding in UMTYMP, but self-report that they are giving a higher priority to other courses or extracurricular activities. In other words, the entrance exam has very few false positives. Furthermore, among the students who have enrolled in the Program, we have observed a positive correlation between higher scores (above 45) and success in UMTYMP, both in terms of number of semesters completed and grades earned. Hence, the exam not only identifies a pool of capable students, but provides a good indication of who the particularly strong and committed students are.

However, these empirical observations do not preclude the possibility that the exam has a number of false negatives—students who could succeed in UMTYMP but do not achieve a predetermined passing score. We are particularly concerned that the false negatives may be concentrated among the female applicants to the Program because of three observations:

1) Historically, on any given entrance exam the average score of the female students has been lower than that of the males.

2) In the (very) few instances when females have been admitted with scores below 40 based on their essay responses, their performance in UMTYMP (in terms of grades and longevity in the Program) has equaled or exceeded that of males who scored 41 or 42 on the same exam.

3) Conversely, the male students who are admitted with scores at or just above the passing line have lower retention rates and grades than other students in the Program.

In other words, the entrance exam appears to adequately identify and rank appropriate male students; males who score higher on the exam are more likely to succeed in UMTYMP, and their overall performance roughly correlates with their entrance exam scores. However, female students with scores near the passing line tend to perform at a higher level than males with similar scores. Their success may be due as much to work ethic, study habits, and overall maturity as mathematical ability, but this suggests that a fixed passing line might fail to identify qualified females with scores that are one or two points below the line.

These observations caused us to wonder whether the exam could be improved so that students with similar scores would have similar success rates within UMTYMP, regardless of gender. This led to a large-scale analysis of our exam results and the admitted students' performance in UMTYMP, with a specific focus on the differences between genders. To give the reader a better context, we begin with a case study of a typical entrance exam.

A Case Study: The Regular 2009 Exam

We evaluate the gender gap on an exam in multiple ways. First, we examine the rough descriptive statistics, comparing median and mean scores. Depending on the context, we might compare the averages for all males and females, or we may analyze a specific subgroup: e.g., fifth-grade males and females; or, sixth-grade males and females who have taken a previous version of our entrance exam. Second, and perhaps more importantly, we scrutinize the students at the top of the pool to determine whether a certain passing line would result in an entering class whose ratio of males to females more closely reflects the proportion in the testing pool. This is similar to the frequently used method of comparing 90th percentile scores among the male and female pools, but allows us to focus on the demographics of a potential entering class. Finally, once students are in the Program, we track their progress to determine whether students whose exam scores were comparable performed at a comparable level in the Program, both in terms of grades and continued enrollment.

Table 2(A)
Gender Comparisons for Regular 2009 Exam

	Males	Females
Number	329	164
Mean Score	33.47	29.94
Median Score	34	30

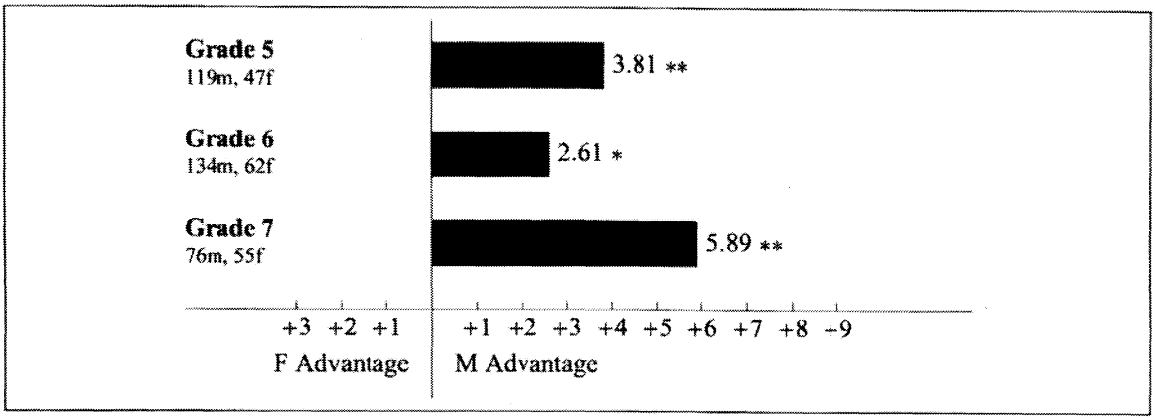
Note: Overall statistics by gender for Regular 2009 Exam. The difference in means is statistically significant (*p<0.0001*).

Table 2(B)
Gender Comparisons for Regular 2009 Exam

Score	# Males	# Females	F% of Potential Entering Class
50	0	0	
≥49	0	1	100%
≥48	2	1	33%
≥47	7	3	30%
≥46	13	5	28%
≥45	23	7	23%
≥44	37	8	18%
≥43	49	11	18%
≥42	58	16	22%
≥41	69	19	22%
≥40	86	28	25%

Note: Scores achieved by male and female students. For each potential passing line, the last column shows the percentage of the admitted students who would be female. The overall testing pool was 33.27% female.

The results of the Regular 2009 Exam are typical and illustrate the types of disparities we have observed between males and females. Table 2(A) shows the overall statistics according to gender. The statistically significant difference in mean scores is persistent across all grade levels (see Figure 3).



* indicates a gap which is statistically significant at the $p<0.05$ level.
** indicates $p<0.01$.

Figure 3. Difference between average male and female scores on Regular 2009 Exam by grade level.

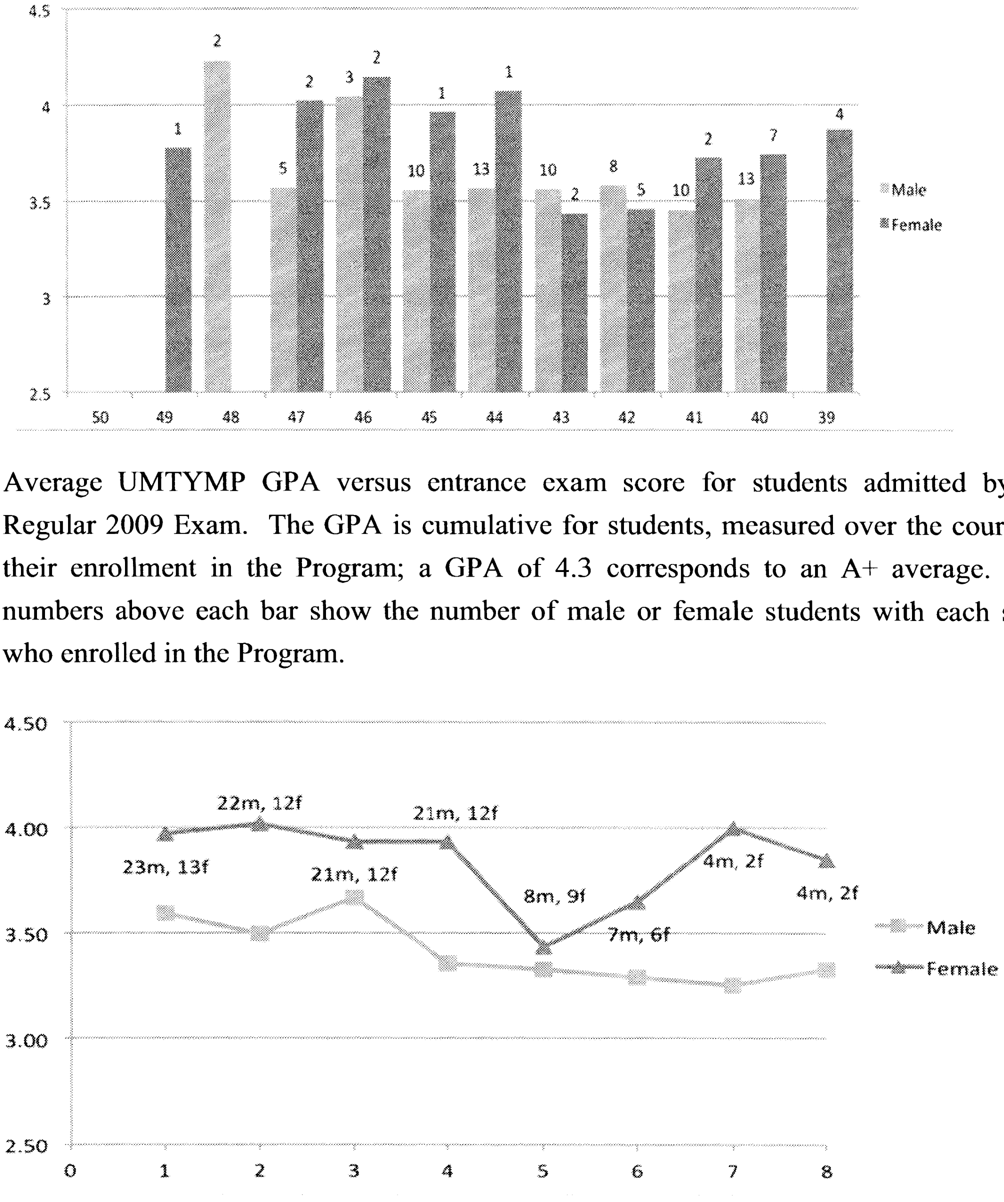
The gap was not simply due to a number of low-scoring outliers, but existed among the top performers as well. Over 33% of the testing pool was female, but Table 2(B) shows the qualifying students were disproportionately male. On this particular exam, the passing line was 40/50, although a handful of female students who scored 39/50 were admitted after evaluating the essay responses on their applications.

A consistent observation in our analysis is that female students generally omit problems at a much higher rate than males, especially toward the end of the exam. Table 4 shows the omission rates by gender for the last twenty problems on the Regular 2009 Exam. There is no penalty for wrong answers on the exam and students are encouraged to guess, so omitted problems generally indicate a student ran out of time to finish the exam. Given the omission rates in Table 4, the gap in average scores is less surprising because female students are completing less of the exam. However, this cannot entirely explain the gap. Even if we compute the percentage of correct responses among questions answered, the male students still outperformed the females by 4.5% (or 2.25 on a 50-point scale, $p < 0.001$).

Table 4
Omission Rates by Gender on the Last Twenty Questions of the Regular 2009 Exam

Question	M	F	Question	M	F
31	5.2%	12.0%	41	17.4%	24.0%
32	4.3%	10.8%	42	20.4%	26.3%
33	6.1%	14.4%	43	22.0%	29.9%
34	5.5%	12.0%	44	22.6%	29.9%
35	6.1%	13.2%	45	24.7%	33.5%
36	7.3%	16.2%	46	27.1%	34.1%
37	11.6%	18.6%	47	32.0%	43.7%
38	13.1%	21.0%	48	35.7%	46.7%
39	19.2%	25.1%	49	35.1%	46.7%
40	19.5%	29.9%	50	37.5%	47.9%

Students who were admitted in 2009 have now completed up to four years of UMTYMP, which allows us to analyze their performance and longevity in the Program. As mentioned earlier, we have observed that female testers admitted with scores at the lower end of the historical passing range tend to be more successful in the Program than male testers with similar scores. Figure 5(A) illustrates this phenomenon for the Regular 2009 Exam testers; recall that the passing line was 40/50, with a few females admitted with scores of 39. Females admitted with a score below 42 had a higher cumulative GPA in the Program than males admitted with a score below 42 (in fact, they outperform males with scores up to 45). Figure 5(B) shows more detail for students admitted by the Regular 2009 Exam with a score below 42, tracking their average grades for each semester of the Program. Although both groups suffered significant attrition, the female students consistently outperformed their male counterparts. This suggests that, at least within this range of scores, the exam results should be interpreted differently for male and female students.



A) Average UMTYMP GPA versus entrance exam score for students admitted by the Regular 2009 Exam. The GPA is cumulative for students, measured over the course of their enrollment in the Program; a GPA of 4.3 corresponds to an A+ average. The numbers above each bar show the number of male or female students with each score who enrolled in the Program.

B) Average UMTYMP grade versus semester in the Program for students admitted with a score below 42 on the Regular 2009 Exam. The numbers at each data point show the number of male and female students enrolled that semester.

Figure 5. Comparison of Regular 2009 Exam scores versus performance in UMTYMP.

We can also measure success in the Program by retention rates. Students leave UMTYMP for many reasons: some are unable to continue due to grades (they must earn a B− or better in the first four semesters and at least a B in the remainder of the Program to proceed); some decide that the format is not appropriate for their learning style; and, some have difficulty with the commute or schedule. As one might expect, students admitted with comparatively low entrance exam scores have a much lower retention rate. However, among this high-risk population, a significantly greater proportion of female students remain enrolled which again indicates that our exam might have incorrectly identified them as marginal (see Table 6). In the remainder of this article, we will refer to the combination of course grades and continued enrollment in the Program under the blanket term "Program Performance."

Table 6
Enrollment Rates of Students Admitted by the Regular 2009 Exam by Score Range and Gender

	Algebra	**Geo/MA**	**Calc 1**	**Calc 2**	**Still Enrolled**	**Retentio n Rate**
45-50						
Male	26	24	20	15	**17**	**65%**
Female	6	5	5	5	**5**	**83%**
42-44						
Male	31	23	16	11	**12**	**39%**
Female	8	7	7	3	**4**	**50%**
39-41						
Male	23	21	8	4	**4**	**17%**
Female	13	12	9	2	**4**	**31%**

Note: Due to deferral, leaves of absence, and other special cases, not all currently enrolled students have completed all four years. Hence, the retention rate may include students who were admitted with this exam, but have only completed *Calculus I.*

Potential Issues

Based on our statistical analysis, literature review, and anecdotal evidence, we initially identified the following possible reasons why female students persistently have lower scores than the males—and why, among those students who are admitted and enroll in the Program, female students have lower scores than would be suggested by their eventual Program Performance.

The Content Balance Hypothesis — We cycle through a number of different versions of the entrance exam; in particular, we never use the same exam for the Early and Regular testing pools in a given year. All of the versions have arithmetic, algebraic or spatial reasoning problems, but some versions might have disproportionately many problems of one type. If males and females were to perform differently on certain types of problems, this content imbalance could generate a gap in performance.

For example, studies such as those in Casey, et al. have indicated that the gender gap among middle school students on an assessment based on TIMSS problems could be traced to a difference in spatial-mechanical reasoning skills [3]. Our entrance exams typically include ten to fifteen problems that incorporate geometric or spatial reasoning, and are taken by students in grades 5-7, creating a potential for a performance gap among our testers. It should be noted that the more recent study described in "New Trends in Gender and Mathematics Performance: A Meta-Analysis" found no significant difference between male and female students' mathematical performance regardless of the problem content [4].

The Bubble Hypothesis — Students record their answers on a bubble sheet, and the exam proctors report that the female students often seem to take much more time carefully filling in the bubbles. Although this distinction may seem trivial, on a fast-paced exam like ours it can be crucial. For example, a student who spends an extra six seconds per problem filling in each bubble would run out of time after forty questions, never getting a chance to answer the remaining problems. If females tend to spend more time than males filling in their answer sheets, it could account for some of the difference in omission rates illustrated in Table 4.

The Guessing Hypothesis — It is difficult to look at an answer sheet and identify which responses were guesses, but anecdotally our proctors have reported conversations with students after the exam in which females have been more reluctant than males to guess on the exam. This is another potential cause for the data in Table 4.

The Arrangement Hypothesis — Related to both the Content Balance Hypothesis and the fact that females are less likely to finish (or perhaps even read) the last ten questions on the exam, the placement of certain questions could affect male versus female performance. If easier problems are heavily concentrated toward the end of the exam, females may never read or answer them, whereas males are more likely to finish the entire exam.

Although some of these potential causes have some basis in the literature, most are anecdotal and may not be valid for our exam and testing pool. However, each of them allows at least a limited opportunity for testing via modest changes to the exam or by adjusting the amount of time per problem. These hypotheses all assume that females and males who take the exam are equally well qualified for UMTYMP, but the following possibility must also be mentioned.

The Testing Pool Hypothesis — On average, the female students in our testing pool may be less mathematically qualified for UMTYMP than the male students.

Recent literature indicates that at the relevant grade levels, there is no longer a significant gender gap in mathematical ability among Minnesota students [5]. However, even if the female and male students in the Twin Cities metropolitan area were equally qualified for UMTYMP, it is possible that parents, teachers, and other educators who recommend UMTYMP to students are not doing so in a gender neutral way. This issue requires investigation, but for the remainder of this article we will focus on the first four hypotheses which deal with modifications to the entrance exam.

Methods

We use multiple versions of the entrance exam, which are rotated between the Early and Regular Exam pools from year to year; for the purposes of this article, we will refer to two specific versions as Form A and Form B. In 2011, when we began our large-scale analysis, we were scheduled to use Form B for the Early Exam, and Form A as the Regular Exam; this happened to match what was used in 2009. This section describes the changes made to these exams in 2011 and 2012. The modifications are important to describe, but they are fairly detailed so the reader may wish to skim the comprehensive description and refer to the following summary and Table 7 as needed. In all, we implemented three different modifications to the exam and testing process:

1) In 2011, we created Forms A2 and B2 by *rearranging* the problems on Forms A and B, respectively. This allowed us to evenly distribute the difficult problems.

2) In 2012, we gave the Early testers a shorter, 40-question version of Form A2. These forty questions represented a better balance of topics than the full 50-question exam. This modification will be referred to as "rebalancing."

3) In 2012, as a consequence of the shorter exam, the Early testers had more time per problem.

Table 7
Description of Each Exam from 2009 and 2011-2012

Exam	Form Name	Questions	Time per Question	Rearranged	Test Pool
Early 2009	B	50	24s	No	137m, 66f
Regular 2009	A	50	24s	No	329m, 165f
Early 2011	B	50	24s	No	69m, 34f
Early 2011	B2	50	24s	Yes	85m, 45f
Regular 2011	A	50	24s	No	126m, 83f
Regular 2011	A2	50	24s	Yes	163m, 88f
Early 2012	A2(40)	40	30s	Yes	176m, 82f
Regular 2012	A2	50	24s	Yes	304m, 161f
Regular 2012	B2	50	24s	Yes	89m, 41f

Modifications in 2011

We began in 2011 by attempting to address the Content Balance and Arrangement Hypotheses previously described. This required the classification of each problem according to difficulty and content. Based on student performance on Forms A and B in previous years, each question was given a difficulty rating: "Easy," answered correctly by over 80% of all students; "Medium," answered correctly by 55-80%; and "Hard," answered correctly by fewer than 55% of the students. In addition, each question was categorized according to its content type, chiefly arithmetic, numerical reasoning, and geometric reasoning, with a small number of questions in modeling and statistical reasoning.

It was immediately clear that Form A had an issue with the distribution of its difficult problems, with a concentration of Hard problems in the last ten questions. Form A also had far more geometric reasoning problems than other versions of the entrance exam: eighteen questions compared to just seven comparable problems on Form B. Moreover, the geometric reasoning problems on Form A were exceptionally difficult compared to other versions, with twelve of them in the "Hard" category.

The paucity of geometric reasoning problems on Form B led to a surplus in other areas, especially arithmetic, which comprised twenty-three of the fifty problems on Form B. This version of the exam had fewer Hard problems overall, but they were concentrated at the end of the exam: thirteen Hard problems overall, but twelve of the last eighteen.

With these discrepancies, it was clear that we would eventually want to replace problems on each form to make them more similar, but during our initial analysis we wished to keep the same problems in order to preserve as much comparability as possible to the 2009 test results. We therefore focused on rearranging the problems on each form with two goals in mind: 1) making the first forty questions from each form more balanced in terms of difficulty and content; and, 2) making the first forty questions of Forms A and B more comparable to each other.

Because of its excess of Hard geometric problems, rebalancing the questions on Form A forced us to make the final ten questions very imbalanced. In this article, the modified Form A will be referred to as Form A2. The first forty questions of Form A2 had nine spatial reasoning questions and thirteen each of arithmetic and numerical reasoning, as well as eight Hard problems, with three each from numerical and spatial reasoning. The last ten questions included nine spatial reasoning problems, eight of which were Hard problems. Hence, the 40-question sub-exam on Form A2 had a very different profile than the 50-question Form A2.

The modifications on Form B were quite different. The most pressing issue was the heavy imbalance of Hard problems at the end (twelve of the last eighteen). Hence with Form B, the major change in the modified version was to make sure eleven of the Hard problems appeared in the first forty questions, including eight from arithmetic and numerical reasoning. In this article, the modified Form B will be referred to as Form B2. The first forty questions of Form B2 also smoothed out the content type distribution, with seventeen arithmetic reasoning, eleven numerical reasoning, and all seven spatial reasoning questions. Thus, the 40-question sub-exam of Form B2 was at least equal to the entire 50-question Form B and, in some ways, slightly more difficult. These two modified exams, Forms A2 and B2, were used in the following situations :

- Forms B and B2 were used for the Early Exam in 2011. The 233 students who signed up for the Early Exam were separated into a control group of 103 who took the original Form B and a group of 130 who took Form B2. (Due to the logistics of scheduling and testing rooms, splitting the pool exactly in half was not feasible.)
- Similarly, a control group of 209 students took Form A as the Regular 2011 Exam, while the remaining 251 students in the Regular pool took Form A2.

Modifications in 2012

The results of the 2011 exams were promising enough that we continued our experiment in 2012. In addition to using a rebalanced exam, our primary goal in 2012 was to address the Bubble and Guessing Hypotheses. We therefore used the first forty questions of Form A2 for the Early Exam, but kept the same time limit. This gave students thirty seconds per question instead of twenty-four, hopefully ensuring that all students (particularly the females) would have time to finish the entire exam. This version of the exam will be referred to as Form A2(40), emphasizing that only the first forty questions were used.

The results of the Early 2012 Exam were more gender neutral than other recent exams. As a control group for these results, we used the full 50-question Form A2 as the Regular Exam, with the standard twenty-four seconds per question, and all of the discrepancies from previous years immediately returned. Also note that 130 students took Form A2(40) as the Early Exam, did not qualify, and decided to re-test at the Regular 2012 Exam. Rather than giving them Form A2, a longer version of the exam they had just taken, these re-testers were given the full Form B2.

Overall Results

Tables 8(A) and 8(B) summarize the performance by gender for each of the exams in 2011 and 2012, with 2009 included for comparison. Both mean and median scores are supplied to give a more nearly complete picture. With our large sample sizes, the median is often too coarse a measure, but it can be very useful in those instances where the mean score is affected by a large number of outlying scores. Consider Form A2 in 2011, which was given to 163 males and 88 females: the median male and female scores were equal, but the difference in mean scores was a statistically significant 2.26, due to a few female students who scored 15 and below. Without those students, the gap in average scores would have been less than 1.5 (with $p = 0.156$).

Table 8
Performance by Gender on All Exams in 2009, 2011, and 2012

Exam	Form Name	Mean Male Score	Mean Female Score	Difference	*p*-value
Early 2009	B	34.87	31.97	2.90	0.018

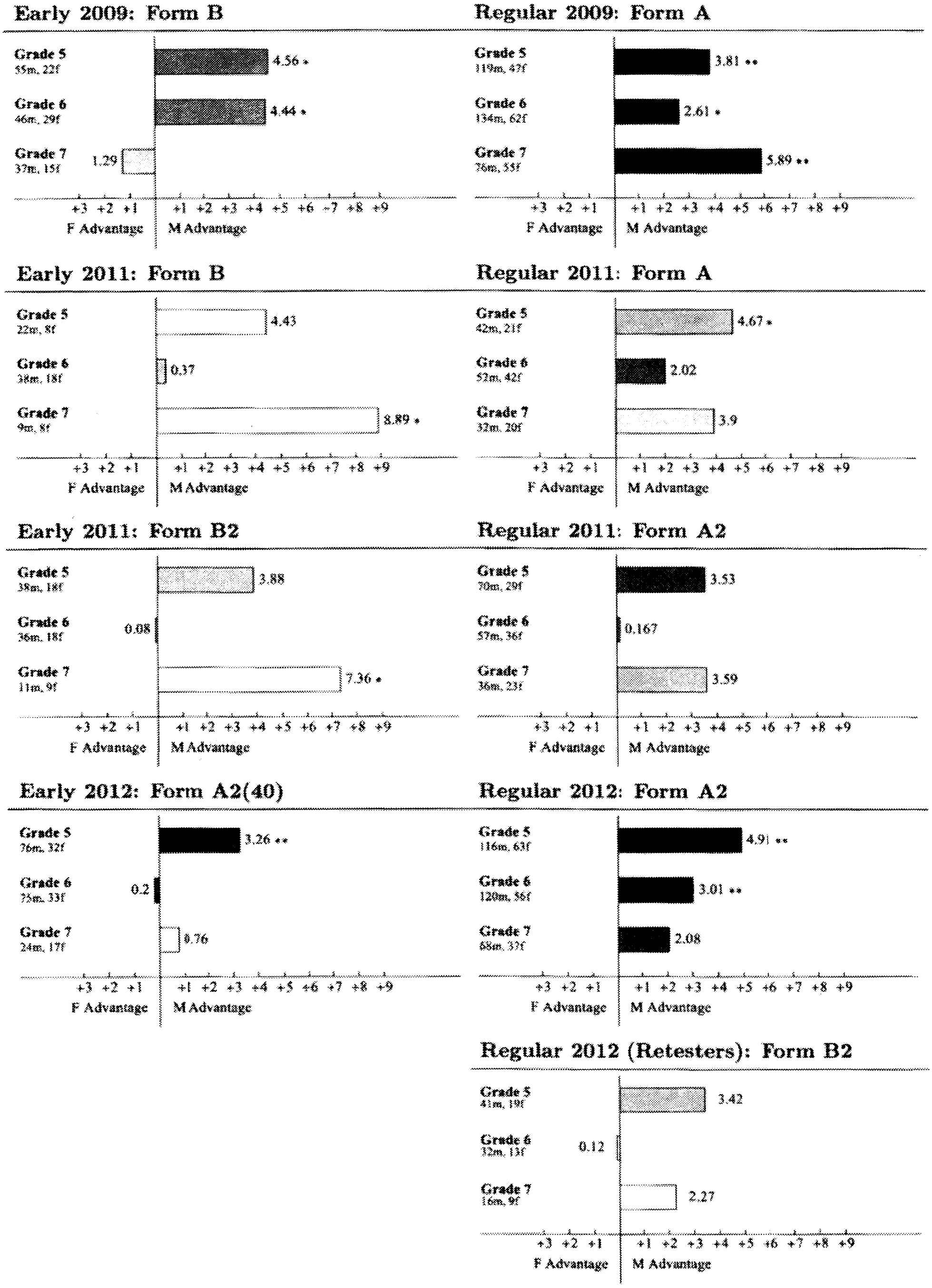
Regular 2009	A	33.47	29.94	3.53	<0.001
Early 2011	B	36.00	33.32	2.85	0.028
Early 2011	B2	36.86	34.07	2.79	0.042
Regular 2011	A	35.56	32.59	2.97	0.007
Regular 2011	A2	34.84	32.58	2.26	0.035
Early 2012	A2(40)	33.74	32.48	1.27	0.058
Regular 2012	A2	35.31	31.65	3.66	<0.001
Regular 2012	B2	38.65	36.59	2.07	0.087

A) Mean scores by gender.

Exam	Form Name	Mean Male Score	Mean Female Score	Difference
Early 2009	B	36	33.5	2.5
Regular 2009	A	34	30	4
Early 2011	B	36	33.5	2.5
Early 2011	B2	38	35	3
Regular 2011	A	37.5	34	3.5
Regular 2011	A2	35	35	0
Early 2012	A2(40)	35	34	1
Regular 2012	A2	35	33	2
Regular 2012	B2	39	38	1

B) Median scores by gender.

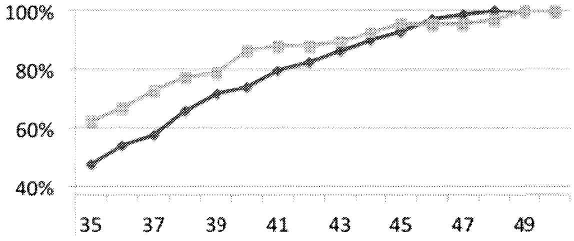
The mean scores are broken down further by grade level in Figure 9, which mirrors Figure 3. Recall that the shading of each bar corresponds to the size of the sub-pool, and statistically significant differences are marked. Hence, the large gaps among seventh graders on the Early 2011 Exams, while significant, represent variation in small numbers of students. Other pools, such as the seventh graders on the Regular 2011 Exams, have far more students, but fail to have a statistically significant gap. (For seventh graders in 2011, $p = 0.08$ for the gap on Form A, and $p = 0.068$ on Form A2.) The cumulative frequency graphs in Figure 10 give a further visual representation of the performance on each exam, broken down by gender.



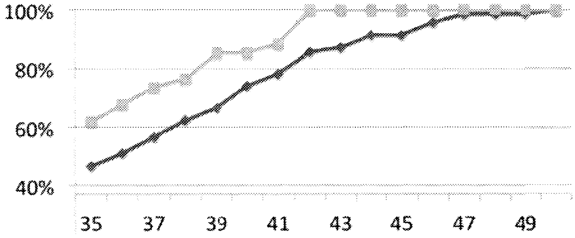
* indicates a gap which is statistically significant at the *p<0.05* level.
** indicates *p<0.01*. NOTE: The bars are shaded according to the size of the pools, with darker bars corresponding to more students.

Figure 9. Difference between average male and female scores on all exams in 2009, 2011, and 2012.

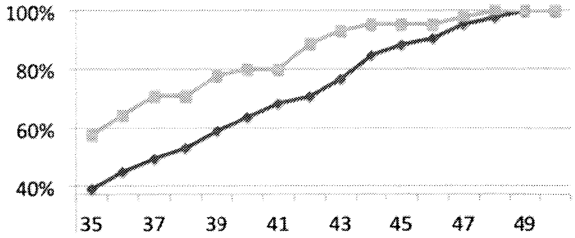
Early 2009: Form B



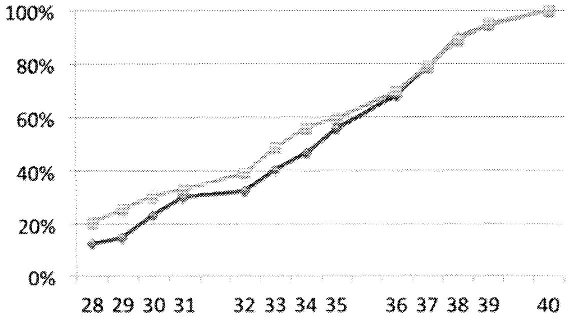
Early 2011: Form B



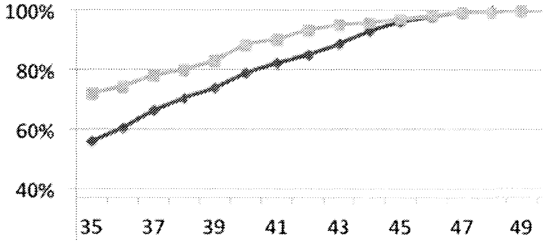
Early 2011: Form B2



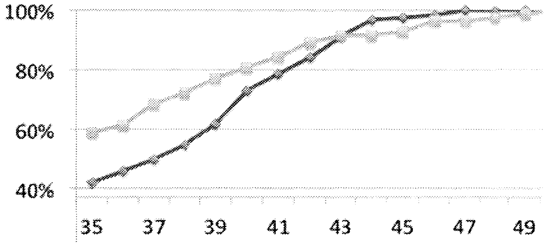
Early 2012: Form A2(40)



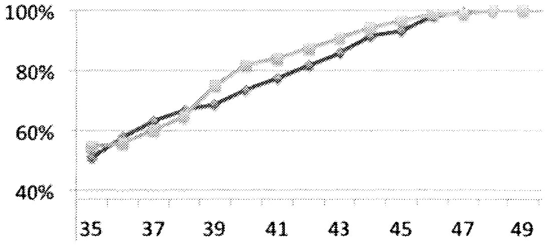
Regular 2009: Form A



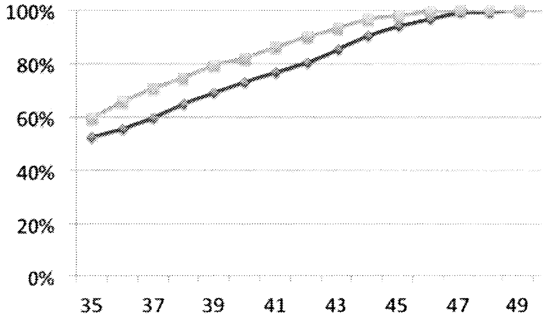
Regular 2011: Form A

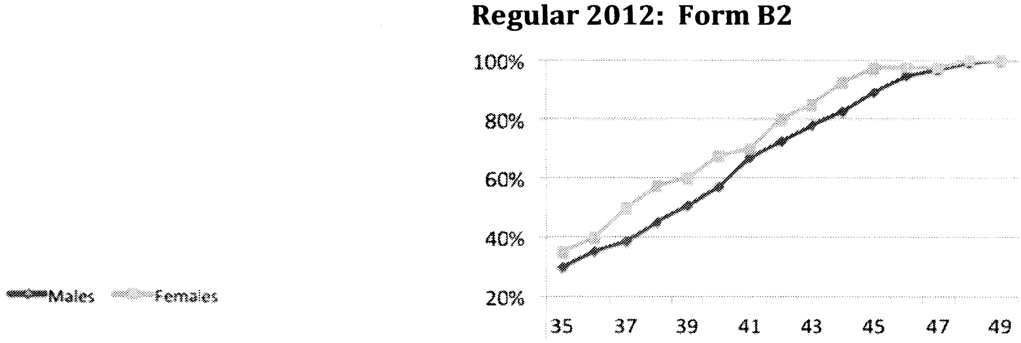


Regular 2011: Form A2



Regular 2012: Form A2





Regular 2012: Form B2

Figure 10. Cumulative frequency of scores by gender on exams.

A few striking patterns are immediately noticeable in these figures:

- With few exceptions (Early 2009 Exam and Regular 2012 Exam), the gender gap is considerably smaller among sixth graders.

- In 2011, the mean values on the rearranged versions of the tests, Forms A2 and B2, had smaller gender gaps than the original Forms A and B. The median values on Forms A2 and B2 had either smaller gaps or were very similar. On Form A2, the median male and female scores were equal; this fact alone was encouraging enough for us to continue the project through 2012.

- In particular, the cumulative frequency graphs in Figure 10 show that, for scores of 35 and up, females and males performed very comparably on the Regular 2011 Exam, Form A2. At first glance, the graph for Form A in 2011 looks very promising, with females outperforming males in the upper range of scores, but there were very few students in this pool overall, and there was a large gap in performance in the 35-40 range. This is the reason we chose to use Form A2 as the basis for the 2012 experiments.

- The Early 2012 Exam had the most gender neutral results of those whose scores are presented here. Although there was a significant gap among the fifth graders, we happen to know there was a group of four female students who scored below 20 (out of 40) and had a large effect on these statistics; if we toss out all scores below 20, the gap between males and females in fifth grade shrinks to 1.64, and is no longer statistically significant. There were no such outliers among the male testers.

Not all of our exams have outlying students with low scores, but we have noticed that when such students exist, they tend to be disproportionately female. We currently have no definitive explanation for this phenomenon, or for why sixth grade females do better relative to their male counterparts than fifth or seventh grade females.

Omission Rates

As previously mentioned, females have historically omitted more questions than their male counterparts. This continued to be the case in 2011, but changed considerably in 2012 (see Table 11). Recall that the Early Exam used Form A2(40), with thirty seconds per question instead of twenty-four. This enabled nearly every student to finish the exam. Omission rates plummeted to near zero, with only small differences between the males and females (compare to Table 4).

Table 11
Omission Rates by Gender on the Last Ten Questions of Each 2012 Exam

Question	M	F	Question	M	F	Question	M	F
31	0.0%	2.4%	41	13.1%	26.4%	41	3.3%	2.5%
32	0.0%	1.2%	42	8.8%	22.0%	42	5.5%	5.0%
33	0.6%	1.2%	43	12.7%	30.2%	43	5.5%	5.0%
34	0.0%	1.2%	44	11.4%	27.0%	44	5.5%	5.0%
35	0.0%	1.2%	45	11.4%	26.4%	45	6.6%	5.0%
36	0.0%	1.2%	46	12.1%	25.8%	46	7.7%	5.0%
37	0.0%	1.2%	47	19.3%	32.1%	47	9.9%	7.5%
38	0.0%	1.2%	48	15.0%	29.6%	48	11.0%	7.5%
39	1.7%	4.9%	49	19.9%	33.3%	49	9.9%	7.5%
40	1.7%	3.7%	50	25.5%	38.4%	50	12.1%	10.0%
(A) Early: Form A2(40)			(B) Regular: Form A2			(C) Regular (Re-Testers): Form B2		

Students who took the Regular Exam were given Form A2 or B2, depending on whether they were re-testers who had already taken Form A2(40). Both Regular Exam pools had the traditional twenty-four seconds per problem. The results for Form A2 reverted to the typical outcome, with much higher omission rates among the female students toward the end of the exam. Interestingly, the re-testers had much lower omission rates overall, and the females actually omitted fewer questions than the males.

Overall, this indicates that familiarity with the exam process may be as important a predictor of omission rates as gender, but there is insufficient data to make any definitive conclusions. Although the gender gap was reversed among the 2012 re-testers, we have examined the data and found that this has not been the case in previous years. Furthermore, although it is tempting to conclude from Table 11 that familiarity with the exam helps reduce omission rates, especially among females, it may simply represent a selection bias. Among the students eligible to re-test in 2012, perhaps more of the strong female students returned than the strong male students. In short, although there was a correlation between re-testing and lower omission rates in 2012, we cannot yet conclude causation.

Effects of Rearranging

In 2011, students took Forms B and B2 for the Early Exam, and Forms A and A2 for the Regular Exam. In both cases, the rearranged versions (A2 and B2) had smaller gender gaps in average scores, as noted earlier. Tables 12 and 13 give further details about the distribution of scores on these exams, and the percentage of female students in each grade range. These are essentially tabular versions of the cumulative frequency graphs in Figure 10. The results of Forms A2 and B2 both seem to suggest that rearranging the problems was helpful for the female testers, but this was evident in different ways.

Table 12
Scores Achieved by Male and Female Students on Early 2011 Exam, Forms B and B2

Score	# Males Form B	# Females Form B	F% of Potential Entering Class	# Males Form B2	# Females Form B2	F% of Potential Entering Class
50	1	0	0%	0	0	-
≥49	1	0	0%	2	0	0%
≥48	1	0	0%	4	1	20%
≥47	3	0	0%	8	2	20%
≥46	6	0	0%	10	2	17%
≥45	6	0	0%	13	2	13%
≥44	9	0	0%	20	3	13%
≥43	10	0	0%	25	5	17%
≥42	15	4	21%	27	9	25%
≥41	18	5	22%	31	9	23%
≥40	23	5	18%	35	10	22%
≥39	26	8	24%	40	13	25%
≥38	30	9	23%	43	13	23%
≥37	34	11	24%	47	16	25%
≥36	37	13	26%	52	19	27%
≥35	44	15	25%	55	25	31%

Note: Testing pool was 33% female for Form B and 35% for Form B2.

Table 13
Scores Achieved by Male and Female Students on Regular 2011 Exam, Forms A and A2

Score	# Males A	# Females A	F% of Potential Entering Class	# Males A2	# Females A2	F% of Potential Entering Class
50	0	1	100%	0	0	-
≥49	0	2	100%	0	0	-
≥48	0	3	100%	1	1	50%
≥47	2	3	60%	3	1	25%
≥46	3	6	67%	11	3	21%
≥45	4	7	64%	14	5	26%
≥44	11	7	39%	23	8	26%
≥43	20	9	31%	30	11	28%
≥42	27	13	33%	37	14	27%
≥41	34	16	32%	43	16	27%
≥40	48	19	28%	51	22	30%
≥39	57	23	29%	54	31	36%
≥38	63	26	29%	60	35	37%
≥37	68	26	28%	69	39	36%
≥36	70	34	33%	80	40	33%
≥35	75	39	34%	88	45	34%

Note: Testing pool was 40% female for Form A and 35% for Form A2.

On the Early Exam, there was a significant gender gap on Form B, which was given to 103 students. In particular, no female scored above 42, compared to ten males who scored 43 or higher. On Form B2, which was taken by 130 students, the gap was still evident (see Figure 10), but there were a small number of females with high scores. In other words, both exams were difficult for students, but the score distribution shifted higher from B to B2 for the female students, more so than for the males.

The differences on the Regular Exam were more noticeable. Females performed significantly better on Form A2 relative to the overall pool than on Form A. In Table 13, for example, we see that 28% of the students who earned scores of 40 or higher *on Form A* were female; recall that the overall pool for Form A was 40% female. For comparison, a full 30% of the students who scored 40 or higher on Form A2 were female, although they only comprised 35% of the overall pool. The decreased gender gap is visibly apparent in Figure 10, where the data points for males and females are closely aligned.

The improved results on Form A2 are particularly interesting, given that the last ten questions on that version are nearly all difficult geometry problems, a type of problem with which female students have sometimes struggled. One possible explanation is that putting all of these problems at the end of the exam helped ensure that females were more likely to answer questions 1–40, which they found easier, and then answer some portion of the remaining questions.

Effects of Rebalancing

Recall that Form A2 was highly imbalanced with respect to content, but the first 40-question sub-exam, Form A2(40), was extremely well balanced. To illustrate the result of a balanced exam, Table 14 compares the results of these two forms on the Early and Regular Exams in 2012. The blank lines in the left half of the table are an attempt to arrange comparable scores next to each other; for example, a score of 36/40 on the Early Exam corresponds to a score of 45/50 on the Regular Exam.

Table 14
Scores Achieved by Male and Female Students on Early 2012 Exam, Form A2(40), and Regular 2012 Exam, Form A2

Score	# Males A2(40)	# Females A2(40)	F% of Potential Entering Class	Score	# Males A2	# Females A2	F% of Potential Entering Class
40	9	4	31%	50	0	0	-
				≥49	2	0	0%
≥39	18	9	33%	≥48	2	0	0%
≥38	37	17	31%	≥47	10	1	9%
≥37	56	25	31%	≥46	18	3	14%
≥36	78	33	30%	≥45	29	5	15%
				≥44	45	11	20%
≥35	94	36	28%	≥43	60	16	21%
≥34	105	42	29%	≥42	71	22	24%

≥33	119	50	31%	≥41	82	29	26%
≥32	123	55	31%	≥40	94	33	26%
				≥39	107	41	28%
≥31	135	57	30%	≥38	123	47	28%
≥30	150	61	30%	≥37	135	47	26%
≥29	154	65	30%	≥36	145	65	31%
≥28	158	68	30%	≥35	166	70	30%

Note: Testing pool was 32% female for Form A2(40) and 35% for Form A2.

It is risky to draw definite conclusions by comparing these two exams, because students had more time per problem on Form A2(40), and also because the Early and Regular Exams have different populations of testers. However, a gender gap is clearly evident on Form A2, yet nearly nonexistent at all levels on Form A2(40). It remains to be seen whether Form A2(40) has accurately predicted students' Program Performance, regardless of gender; however, initial data from those students' first year in the Program is very promising.

Discussion

Students who were admitted based on the 2011 and 2012 exams have not yet accumulated enough retention data and course grades for us to assess their Program Performance; hence, any study of the effectiveness of these exams in predicting future success within UMTYMP must wait for a future longitudinal study. For now, we can analyze the exam scores to see whether our modifications resulted in a testing process in which male and female students pass the exam in proportion commensurate with their proportions of the overall testing population. The answer appears to be a cautious "yes," and the appropriate testing process seems to be a blend of having well-balanced exams with respect to both difficulty and type of problem, as well as giving students slightly more time to complete the exam. Again, definitive answers must wait until we verify that the admitted students' Program Performance is consistent with their scores.

The number of parameters involved in a large entrance exam administered to a large and changing pool makes it very difficult to point to a specific exam modification and definitely conclude that it had a specific, permanent effect. Some parameters that are certainly relevant to our study include the following factors: 1) differences between Early and Regular testers; 2) which students have chosen to take the test each year and why; and, 3) which students decide to re-test within the same year. However, it is highly suggestive that the most gender-neutral results on any exam were on Form A2(4)—the only exam which was rearranged, rebalanced, and on which students had thirty seconds per question instead of twenty-four. For the 2013 exams, we attempted to replicate this pattern. Results were encouraging and will be described in a future paper, once we have more data from 2014 and beyond to allow us to draw stronger conclusions.

Although the Content Balance and Rearrangement Hypotheses may turn out to be valid, it is harder to make any definitive conclusions about the Guessing and Bubble Hypotheses, which gave possible explanations for why females tended to have higher omission rates. Any definite conclusions about these possible causes would be beyond the scope of our current work, requiring extensive observations with stopwatches and post-test interviews to determine which answers were or were not guesses. However, an exact determination of how much each of these hypotheses might explain the high omission rates may not be necessary because, whatever the cause, the omission rates for both genders decreased to near zero once we gave students thirty seconds per question on Form A2(40) (see Tables 11 and 4). Regardless of whether females were taking too long to fill in bubbles, or were unlikely to guess, the omission rate problem has largely been solved.

However, as already mentioned, we now need to explore whether having previously taken an UMTYMP entrance exam is a better predictor of omission rates than gender. If re-testing is more beneficial for female students, we could improve our female passing rates by encouraging more females to take the test a second time. It is worth noting that, although re-testers tend to improve their response rate, they do not necessarily improve their score. Hence, familiarity with the exam process might make students work faster, but not necessarily more accurately.

An interesting byproduct of this project was the in-depth analysis of students' longitudinal success in UMTYMP compared to their entrance exam scores; this was described for students admitted on the Regular 2009 Exam, but results from other years have been very similar. Recall that females admitted by the Regular 2009 Exam with a score below 42 consistently had higher GPAs than males admitted with a score below 45, or even with a score below 42; this was the case despite the fact that some female students were admitted with a score of 39 and no male students were. Anecdotally, it was always suggested that females with lower scores were in fact comparable to males with higher scores, but this has now been verified. As an important consequence, this could justify setting separate, lower passing lines for females in an effort to increase female enrollment in UMTYMP. However, this would not be an entirely satisfactory solution from a public relations point of view. We will therefore continue to explore whether we can create an exam on which students with similar scores have similar Program Performance, regardless of gender.

Finally, while compiling data for this study we analyzed our testing pool more closely than had been done before. This year-to-year analysis showed that our testing pool is much more variable than we had realized, and highlights a recruitment problem. Our first priority is certainly to ensure that our measuring instrument is as fair as possible, as well as capable of correctly identifying qualified students. However, in order to improve the gender balance among enrolled students (which is already high for such a program) we need to encourage more qualified female students to take the entrance exam in the first place. Studies suggest that elementary school girls are aware of the stereotype that men are considered to be better at math than women, but do not personally believe the stereotype [6]. However, studies also indicate that susceptibility to stereotype threat becomes a problem at around twelve years of age, which means that some but not all of our testing pool is likely to be affected [7]. Achieving gender balance in our enrollment could therefore require a combination of both modifications to our exam and targeted intervention programs like those described in “Can Equity Thrive in a Culture of Mathematical Excellence?” which had very positive results on female enrollment in the University of Minnesota Talented Youth Mathematics Program twenty years ago [2].

References

[1] H. Keynes and J. Rogness, "Historical Perspectives on a Program for Mathematically Talented Students," *The Montana Mathematics Enthusiast*, **8**(1) (2011) 189-206 [Originally presented at the 11th International Congress on Mathematics Education; this issue of TMME served as the proceedings of the ICME-11's Study Group 6].

[2] H. Keynes, "Can Equity Thrive in a Culture of Mathematical Excellence?" in W.G. Secada, E. Fennema, and L. Byrd (eds.), *New Directions for Equity in Mathematics Education*, Cambridge University Press, 1995.

[3] M.B. Casey, R. Nutall, and E. Pezaris, "Spatial-Mechanical Reasoning Skills versus Mathematics Self-Confidence as Mediators of Gender Differences on Mathematics Subtests Using Cross-National Gender-Based Items," *Journal for Research in Mathematics Education*, **32**(1) (2001) 28-57.

[4] S.M. Lindberg, J.S. Hyde, J.L. Peterson, and M.C. Linn, "New Trends in Gender and Mathematics Performance: A Meta-Analysis," *Psychological Bulletin*, **136**(6) (2010) 1123-1135.

[5] J.S. Hyde, S.M. Lindberg, M.C. Linn, A.B. Ellis, and C.C. Williams, "Gender Similarities Characterize Math Performance," *Science*, **321** (2008) 494-495.

[6] J. Steele, "Children's Gender Stereotypes about Math: The Role of Stereotype Stratification," *Journal of Applied Social Psychology*, **33**(12) (2003) 2587-2606.

[7] C. Good and J. Aronson, *The Development of Stereotype Threat: Consequences for Educational and Social Equality*, Erlbaum Associates, Mahwah, NJ, 2007.