The Gender Paradox in School Mathematics

Malcolm Cunningham

Carleton University

Current quantitative gender and mathematics ability and postsecondary access research exist in different literatures, each drawing different conclusions. But if they are studied together as a single model it may be possible to demonstrate previously unrecognized associations. Thus using 106,473 standardized Grades 3, 6 and 9 Ontario student test responses from a single cohort, the current study investigates the likelihood of enrolling in Grade 9 academic mathematics against elementary achievement and gender in a single logistic regression model. Results indicate that males achieved higher, (0.028≤d≤0.118), and occupied both achievement extremes in greater numbers, (1.04≤VR≤1.10), while females were 1.5 times more likely to enroll in academic courses. These are paradoxical results which are discussed in relation to the utility of achievement and enrollment as effective metrics in gender and mathematics research.

La recherche quantitative actuelle sur la différence en mathématiques entre les filles et les garçons d'une part, et la recherche sur l'accès aux études secondaires d'autre part, se déroulent dans deux domaines différents et les deux arrivent à des conclusions différentes. Toutefois, en les étudiant ensemble selon un modèle unique, il se peut que l'on puisse démontrer des associations non reconnues auparavant. Ainsi, puisant dans 106 473 réponses aux examens normalisés d'une seule cohorte d'élèves en 3e, 6e et 9e années en Ontario, nous avons étudié, au moyen d'un modèle de régression logistique, la probabilité de s'inscrire à des cours de mathématiques théoriques en 9e année en relation avec le sexe et le rendement à l'école élémentaire. Les résultats indiquent que le rendement des garçons était supérieur (0.028≤d≤0.118), que leurs réponses se situaient aux deux pôles de rendement plus souvent (1.04≤VR≤1.10) et que les filles s'inscrivaient aux cours académiques 1,5 fois plus souvent que les garçons. Nous discutons de ces résultats paradoxaux par rapport à l'utilité du rendement et de l'inscription comme mesures dans la recherche portant sur la différence en mathématiques entre les filles et les garçons.

At least for a century gender and mathematics research has bolstered the broad societal belief that girls and women are not as mathematically talented as boys and men (Li, 1999; Syzmanowicz & Furnham, 2011). Although the pervasiveness of this belief may explain why females are underrepresented in fields requiring mathematical expertise, in today's world it also raises questions about how gender evidence is used to inform theory. Quantitative gender and mathematics research, for example, can be broadly organized as two inquiry lines. The first, gender and mathematics ability research, generally relies on achievement metrics—aggregate scores that summarize students' responses to test questions—to inform theoretical claims about ability (e.g., Hedges & Nowell, 1995; Knapp, Kelly-Reid, & Ginder, 2012; Schoenfeld, 2007). Achievement scores are advantageous proxies for ability because they can be robustly analyzed using established statistical methods. The second inquiry line involves the relation between gender and postsecondary access. Here, studies rely on student enrollment data in degree programs (e.g., Bachelor of Science), which are also well suited to statistical treatment, to inform theoretical claims about relative group success. It is reasonable to initially assume that mathematics ability and postsecondary access are positively related to both sexes.

The trouble is, even when very large samples and common metrics are used, ability and access studies arrive at very different conclusions. And, although this perpetuates a lively societal debate, it does little to resolve matters. For as things stand, robust analyses of school achievement evidence simultaneously support claims of a slight male ability advantage, a slight female advantage, and no gender difference at all (e.g., Hedges & Nowell, 1995; Lindberg, Hyde, Petersen, & Linn, 2010; Voyer & Voyer, 2014). Analyses of enrollment data, meanwhile, consistently indicate a growing female advantage in preparation for, enrollment in and success with completing postsecondary programs (e.g., Riegle-Crumb, 2010). Overall, therefore, gender and mathematics evidence points toward a perplexing paradox. For how is it that ability studies, especially those using similar methods and metrics, arrive at such different conclusions? And why are gender, mathematics ability, and access variables not positively related?

It has been argued that comparing school achievement and postsecondary enrollment data is fraught because they represent completely different contexts. There is considerable conceptual distance between variables leading to the possibility of many, as yet unaccounted for, intervening factors (Hango, 2014). So even if a theoretical case could be made, there is no methodological precedent for a study of these variables as they are currently defined. But if ability and access could be redefined with respect to a single context it may yet be possible to mitigate this distance. Thus I propose to restrict the current study to a school context using data derived from related instruments administered over time to a single cohort of students. Achievement can serve as a proxy for mathematics ability while enrollment in secondary school programs designed for children with postsecondary aspirations, can serve as a proxy for access. If it is known which students are enrolled in such programs in secondary school then elementary achievement and gender can serve as predictors in a single binary response model. This research design is a defensible basis from which to investigate current assumptions informing our understanding of the relations between gender, ability and access.

Literature Review

Research Metrics

Quantitative gender and mathematics claims often rest on outcomes of three statistical metrics: effect size, variance ratio, and enrollment. Effect size—mean achievement difference between sexes expressed as a ratio of the pooled standard deviation—is used to compare mean achievement outcomes between genders. It is often used across different populations and is of particular utility in meta-analyses (Hedges & Nowell, 1995). By convention, when d is positive males have a performance advantage, and when d is negative females have the advantage. Effect sizes are also frequently interpreted relative to Cohen's power scale, such that: $d \le |0.2|$ is negligible, $|0.2| < d \le |0.5|$ small, $|0.5| < d \le |0.8|$ medium, and $d \ge |0.8|$ is interpreted as a large gender effect (Cohen, 1977; Hyde, Fennema, & Lamon, 1990). Variance ratio (*VR*) is defined as the ratio of male to female achievement distributional variances. By convention, if VR > 1 there is greater male variance and VR < 1 indicates greater female variance (Hyde et al., 1990). Postsecondary enrollment data, meanwhile, are reported as counts or percents and are used as a rough indication of relative group success (Riegle-Crumb, 2010).

These metrics are not an exhaustive list but they are important insofar as they are widely used to back theoretical claims (e.g., Hyde, Lindberg, Linn, Ellis, & Williams, 2008). They exist, however, in separate literatures and involve different populations. Hence effect sizes and variance ratios are found mainly in mathematics ability studies while student enrollment data is often encountered in postsecondary access studies.

Gender and Mathematics Ability Research

Gender and mathematics ability research includes a very large number of studies. Hyde and colleagues originally coined the *gender similarities hypothesis* to describe their contention that the sexes do not differ in mathematics ability (Hyde, 1981). This characterization is useful for the purposes of the current argument as many studies have the opposing view, characterized here as the gender differences hypothesis.

The gender similarities hypothesis. The gender similarities hypothesis asserts that evidence of ability difference likely arises from differences in students' opportunities to learn. As a result, males and females do not differ in aptitudes so much as from differences in social and cultural opportunity. Hyde (1981) expounded the view through an analysis of data taken from Maccoby and Jacklin's 1974 book *The Psychology of Sex Differences*. She used what was then new meta-analytic techniques, including effect size and variance ratio metrics, to challenge the authors' original claims that girls have greater verbal ability while boys are better visual-spatially and mathematically. Hyde's findings indicated a moderate Cohen effect (d = 0.43) favoring males. But she argued that, this value notwithstanding, mean score sex differences accounted for only about 1% of total population variance and, thereby, failed to support Maccoby and Jacklin's original position. Hyde et al. (1990) followed up with a larger and more comprehensive meta-analysis. Data published between 1963 and 1988 were taken from 100 selected studies ranging in sample sizes (30 to 90,000) and participant ages (5- to 27-years) yielding 259 effect sizes. Findings indicated an overall weighted mean effect size of 0.15 consistent with a negligible Cohen interpretation. The researchers reported that elementary and middle school females enjoyed slight performance advantages in computation and were comparable to males in problem solving ability. Indeed, achievement differences favoring males did not appear unless secondary school or precocious student data were specifically selected. For these reasons, Hyde and colleagues concluded that results failed to reject the similarities hypothesis.

Faced with criticism about robustness of Hyde's earlier findings, Lindberg et al. (2010) combined nationally representative samples with new analytic methods (see also, Else-Quest, Hyde, & Linn, 2010; Hyde et al., 2008). They hypothesized that differences in sampling methodologies do not markedly influence results and tested this proposition by comparing analyses of mixed sampling methods reminiscent of Hyde's earlier work with analysis of large datasets based on probabilistic sampling (c.f., Hedges & Nowell, 1995). The mixed group included samples of variable size and age composition. Results indicated that weighted effect sizes and mean variance ratios slightly favored males (d = 0.05, VR = 1.07). As noted earlier, no effect size differences were recorded among elementary school samples although a slight difference favoring males did appear in secondary school and college samples. The probabilistic sampling group, meanwhile, included nationally representative achievement data from the National Longitudinal Study of Youth—NLSY-97 (1997–2002), the National Educational Longitudinal Study—NELS-88 (1988–1992), the Longitudinal Study of American Youth—LSAY

(1987–1992), and the National Assessment of Educational Progress—NAEP-92 (1992–2004). Findings indicated an averaged weighted effect size and an averaged weighted variance ratio that were similar to the mixed samples analysis (*d* = 0.07, *VR* = 1.09). Lindberg and coworkers found that there was insufficient evidence in the probabilistic sampling group to reject the NULL hypothesis (i.e., males and females are similarly able) and thereby concluded that choice of sampling methodology does not affect outcome.

In a separate line of inquiry, Else-Quest and colleagues analyzed trans-national data derived from 493,495 14-16 year-old student responses to the 2003 Trends in International Mathematics and Science Study (TIMSS-2003) and the Programme for International Student Assessment (PISA-2003) (Else-Quest et al., 2010). The study estimated differences in mathematics achievement, attitudes, and affect by sex across 69 countries. Findings indicated that although averaged national effect sizes varied widely ($-0.42 \leq d \leq 0.40$), collectively they failed to reject the similarities claim (i.e., $\text{mean}_d = 0.15$). Else-Quest and colleagues conjectured that variable effect sizes likely mirrored cultural differences in social opportunities experienced by these children.

The gender differences hypothesis. Without the aid of Cohen's power scale for interpretation, Hedges and Nowell (1995) established a contrary position to the gender similarities hypothesis. They noted that in the previous 30 years, although varying widely in reported effect sizes and variance ratio magnitudes, an overwhelming number of studies supported a slight male achievement advantage and greater male distributional variance. Hedges and Nowell hypothesized that magnitude differences across studies are likely the result of variable sampling methods so they proposed to restrict analysis to just six large datasets, each constructed using probabilistic sampling. These included Project Talent-1960 (Talent-60), National Longitudinal Study—1972 (NLS-72), NLSY-80, High School & Beyond—1980 (HS&B-80), NELS-88, and NAEP-69. A slight male achievement advantage, $0.03 \leq d \leq 0.26$, and greater male distributional variance, $1.05 \leq VR \leq 1.25$ were the main findings of the study. The researchers reported greater numbers of females in the bottom 10% and greater numbers of males in the top 10% of the achievement spectrum than could be accounted for by chance (see also, Hedges & Friedman, 1993). Based on these findings, Hedges and Nowell contested the gender similarities hypothesis and concluded that analysis using probabilistic sampling methods support a claim of sex ability differences in mathematics.

Gender ability differences claims not based on Cohen's scale also tend to favor males in large-scale international aptitude and achievement test results (c.f., Else-Quest et al., 2010). Brochu, Deussing, Houme, and Chuy (2013) reported Canadian results for PISA, an aptitude test designed to measure the extent to which 15-year-olds acquire knowledge and skills necessary to become full participants in modern society. The PISA-2012 test was focused primarily on mathematics literacy and administered to 20,994 Canadian students across ten provinces (3,652 of whom were Ontario students). Unweighted findings show a small male achievement advantage (Canada *d* = 0.035, Ontario *d* = 0.043) that is consistent with a real ability difference claim. A comparison of multiple years of PISA data, moreover, came to similar conclusions (Stoet & Geary, 2013). Russo, Barbaranelli, and Caponera (2014) analyzed 3,391 Italian student responses to the TIMSS-2011 achievement test and also reported a small male achievement advantage.

Even mathematics achievement studies reliant on multilevel methodologies report a slight male achievement advantage. Rogers et al. (2006) studied, among other things, gender and achievement in a sample of Canadian Grade 6 students. They pointed out the importance of interpreting student-, classroom-, and school-level predictors in a single model; students exist within classrooms which are embedded within schools. If effects at each of these levels are not taken into account, claims pertaining to any given level will likely be confounded. Using Alberta Provincial Language Arts and Mathematics Achievement Test results, they analyzed 3,643 student responses from 198 classes and 129 schools. Altogether, 26 student-, 62 classroom-, and 59 school-level variables were included in a grand-mean centered three-level hierarchical linear model. Rogers and colleagues found that students within classes accounted for 75% of model variance and that mathematics achievement slightly favored males (d = 0.074). These findings, moreover, were replicated in Alberta provincial achievement data analyzed by Pope, Wentzel, Braden, and Anderson (2006).

Some have argued that findings are influenced by the kinds of instruments used to collect data. Large-scale achievement tests are often used to determine whether or not students reach predetermined curriculum-related benchmarks but are otherwise less appropriate when determining changes across either time or human development (Robinson & Lubienski, 2011; Lubienski, Robinson, Crane, & Ganley, 2013; Miller & Halpern, 2014). Lubienski et al. (2013) investigated gender and mathematics achievement among other things using data from a single longitudinal instrument: the Early Childhood Longitudinal Study–Kindergarten Class of 1998-99 (ECLS-K). The research design included a developmental achievement scale and an adaptive staged-design that tracked students' achievement through five waves as they progressed from kindergarten to Grade 8. At each wave, students were first given routing tests to determine their understanding followed by appropriately leveled tests. Hence findings could be arrayed along a single developmental scale. Results indicated that although no gender difference in achievement was detectable at the outset, by the end of kindergarten a noticeable gap favoring males appeared at the top of the achievement spectrum. This gap expanded to include the entire achievement range by Grade 5, widening to a maximum effect size of 0.24. After Grade 5, it narrowed but never entirely disappeared. The researchers noted that this achievement disparate was more apparent among students from economically advantaged families than it was from economically disadvantaged families. They concluded that findings are consistent with a combination of innate ability and socially conditioned differences between the experiences males and females have at school and home over time (see also, Miller & Halpern, 2014).

More recently, evidence of a slight female achievement advantage has emerged when classroom derived data, rather than standardized achievement test data, are used in analysis. Duckworth and Seligman (2006) maintained females are more self-disciplined than males and that this is relevant when it comes to interpreting achievement over a span of time such as a school semester (see also, Russo et al., 2014). Voyer and Voyer (2014) concurred, arguing that the pervasive mythology surrounding male achievement advantage is, for the most part, wholly reliant on interpretations of meta-analyses of standardized achievement scores. Classroom-based assessments of ability are, by contrast, associated with important social and temporal factors not easily captured in test settings. So they conducted a meta-analysis of teacher-generated achievement data to investigate evidence of gender differences. They were also interested in any moderating factors that might help to explain observations. Voyer and Voyer used hierarchical linear modeling as it requires no assumptions about equality of sample sizes among and between model levels. The final two-leveled model was comprised of 502 effect sizes in the first level and 369 samples in the second. Results indicated that mean mathematics achievement accounted for a small effect size favoring females (d = -0.069; originally reported as a positive d value but shown here as a negative value to maintain consistency). The authors discussed various socio-cultural influences that might explain findings (e.g., family beliefs about mathematics, stereotype threat) but left this to future studies.

Gender and Postsecondary Access Studies

The theoretical morass that is gender and mathematics ability research starkly contrasts with the unequivocal clarity of access findings; females enroll in and complete postsecondary programs in greater numbers than males. In a report for the U.S. National Center for Educational Statistics, Hussar and Bailey (2014) summarized college enrollment and degree completion rates over a fourteen year period. They found that females enrolled in bachelor programs in greater numbers than males in the period 1997 to 2011, (male 41% vs. female 48%), and are projected to widen this margin from 2011 to 2022 (male 9% vs. female 18% increases). Not surprisingly, the same trend is seen and predicted for degree completion rates (historical 1997-2011, male 41% vs. female 48%; projected completion increase 2011-2022, male 11% vs. female 22%). These are not unique findings, moreover, as similar U.S. and international enrollment and degree completion rate differences are attested elsewhere (e.g., Jacob, 2002; Knapp et al., 2012; Peter & Horn, 2005; Riegle-Crumb, 2010).

There is evidence that Canadian females enroll in and graduate from universities in higher numbers than males, but as Hango has argued, there are also important sex differences in program choice and how they are related to secondary school mathematics achievement (Hango, 2013, 2014). Combining biannual Youth in Transition Survey (YITS) data—a survey of major transitions in education, training and work—with PISA data, Hango studied the association between mathematics achievement and university program choice. University-bound students' achievement performance in the PISA-2000 survey was combined with six YITS cycles (including participants up to age 25-years). Among students who participated in the PISA-2000 mathematics test, and regardless of original achievement score, 50% of females selected social sciences programs (vs. 32% of males), 20% selected science, technology, engineering and mathematics programs (vs. 44% of males), 14% selected business (vs. 14% of males), and 12.3% chose health (vs. 5.5% of males). This suggests, at best, tenuous connections between university access, program choice and secondary school achievement. And Hango concluded that postsecondary access cannot be fully explained by prior mathematics achievement.

Taking Stock: Gender, Ability and Access

When taken together gender, mathematics ability and postsecondary program access claims present such confusing and contradictory evidence that their relation can only be described in terms of paradox. Complicating matters, posing any substantive questions about their relations is methodologically fraught because research claims are defended in separate literatures. Hango's comparison of gender, students' mathematics ability at 15-years-old, and eventual university program choice is a notable exception. But even here, while datasets, analytic approaches, and interpretations were cleverly integrated post hoc to address research questions, a myriad of intervening factors could have helped to guide students' eventual choices (Hango, 2013, 2014). It is unclear, for example, just how postsecondary access and course enrollment are related. Paradoxical findings across literatures may be explainable, therefore, as an artifact of the conceptual distance that exists between school achievement and postsecondary enrollment. Thus a simpler approach, one where prior achievement and access are coherently defined with respect to a single context, may yet mitigate distance between variables and reveal hitherto unrecognized dependencies. I propose to study the relation between these variables in a school context.

For a study of school mathematics it is essential to establish a consistent definition for achievement and to establish what counts as a significant achievement result. Guskey (2013) broadly defines achievement as the accomplishment of mathematics learning goals. It is associated with specific curricular targets or aptitudes yet ubiquitously represented as aggregate scores; summaries of students' item-level test responses. These characteristics are common in a great many papers that are reliant on achievement results. Type of instrument used to collect data, however, still allows for interpretive latitude. Achievement is variously defined as scores derived from teacher-generated classroom tests, standardized cross-sectional tests, and standardized longitudinal tests (e.g., Hyde et al., 2008; Hedges & Nowell, 1995; Lubienski et al., 2013; Voyer & Voyer, 2014). It is also encountered as scores derived from standardized aptitude tests and scores derived from standardized curriculum-based tests (Brochu et al., 2013; Mullis, Martin, Foy, & Arora, 2012). About the only reliable characteristic of studies in this respect seems to be a lack of discussion about what achievement is or how different instruments influence interpretations about achievement and, ultimately, claims about mathematical ability. In the current paper, therefore, achievement is defined as aggregate scores associated with students' responses to standardized curriculum-based tests. This is a reasonable provisional definition that is consistent with many published gender and mathematics ability research studies. What counts as significant achievement results, meanwhile, is more a matter of convention. The disparate between similarities and differences in ability claims arguably rests on the relative importance afforded to Cohen's scale on what is otherwise broad similarity in data, analytic methods and results. Gender similarities studies use Cohen's scale to back interpretations of results while gender differences studies generally do not. To remain consistent with the majority of published papers, therefore, the current study will not rely on Cohen's scale.

For a study of school mathematics it is also essential to redefine program access. If achievement is defined as aggregate scores associated with students' responses to standardized curriculum-based test questions then access can be redefined as student enrollment/non-enrollment in a university-bound secondary mathematics program stream. This at least notionally links school access to the current postsecondary enrollment literature. Likelihood of school enrollment in a university-bound program can then serve as a dependent variable in a logistic regression with gender and prior elementary achievement as predictors (Long, 1997). This research design establishes a single interpretative context to study gender, ability and access, a coherent methodological approach, and establishes a basis from which evidence, if any, of gender paradox can be further investigated.

Data Source

The province of Ontario gathers information about student mathematics achievement through annual full-census large-scale standardized tests. These tests are designed and administered by an arm's-length agency, the Education Quality and Accountability Office (EQAO), which oversees Grade 3, 6 and 9 administrations (Education Quality and Accountability Office, 2011, 2012). Tests are intended to produce reliable evidence of student performance as it relates to the provincial mathematics curriculum and results are used to inform stakeholders about the effectiveness of the Ontario curricula. The long-term goal is incremental system-wide improvement.

In each of these tests, students' achievement scores—although assessing different mathematics curricula—are interpreted to be psychometrically linked. The agency achieves instrument consistency by careful item design, testing, use of common test formats and consistent administration procedures and scoring. A child's proficiency in number sense and numeration in the elementary grades, for example, is presumed to be indicative of continued success in Grade 9 (e.g., Shulman, Hinton, Zhang, & Kozlow, 2014). All tests are comprised of a mixture of multiple choice and open response type questions. Open response items are scored by trained markers using an item-specific rubric and a 4-point scale to indicate quality of student answers. Multiple choice items are dichotomously scored by machine. Based on their answers to open-response and multiple choice questions, students are assigned a leveled achievement score from 1 to 4. A score of 1 indicates that one is below, a score of 2 indicates one is approaching, a score of 3 indicates one is at, and a score of 4 that one is above pre-established provincial expectations. Consistency in design, administration, and interpretation of Ontario tests, although not strictly longitudinal, makes them an appropriate source of data for the current project.

When Ontario students enter secondary school they are assigned to different streams (applied, academic) in courses (e.g., science, English) that, in combination, lead to different postsecondary opportunities (Taylor, Krahn, & Levine-Rasky, 2009). In mathematics, the applied stream is intended to deal with essential topics. Students learn through practical applications and concrete examples. The academic stream, meanwhile, deals with theoretical topics and requires students to manage more abstract problems (Ministry of Education, 2005a).

Purpose

In Ontario, the Ministry of Education clearly articulates the relation between effort, achievement and motivation in Grade 9 mathematics as (Ministry of Education, 2005b): "Students who make the effort required and who apply themselves will soon discover that there is a direct relationship between this effort and their achievement, and will therefore be more motivated to work" (p. 4). Given this position, it is reasonable to initially assume that student access to Grade 9 academic programs and mathematics ability are related. What is not clear is the role that gender plays.

I hypothesize that gender influence in school mathematics will manifest in one of two ways; either there will be less or there will be at least as much evidence of paradox compared to previous research findings. For the vast majority of students, gender is indelibly designated and codified in the elementary grades. As a consequence, model interpretation primarily depends on the relation between prior achievement and Grade 9 program enrollment, the influence of gender being later inferred. Thus if Grade 9 program enrollment and prior achievement are positively related in a simpler design it is reasonable to assume there will be less evidence of paradox. This assumption is consistent with previous claims about gender and achievement difference/similarity as an inherent human/social attribute with the added benefit of a logical association with gender enrollment numbers in Grade 9 academic mathematics. Evidence of a positive association between variables would arguably point to a problem comparing a single context model and earlier results—there being likely fewer confounding factors in the smaller, more cohesive design. But there is also a possibility that prior achievement and Grade 9 academic access are negatively related or unrelated. Even with the interpretative benefit afforded by a simpler design, this case likely leads to paradox. For if access and achievement are negatively related or unrelated then their association with gender is not straightforward and, as a result, gender will interact differently with each. Making sense of paradoxical results in both a simpler design and earlier findings arguably raises a more difficult theoretical question about how these variables are defined, measured and interpreted.

I will test these hypotheses first with a descriptive picture of the Ontario dataset. This will be followed by two logistic regression models. The first model will estimate the likelihood of students enrolling in Grade 9 academic programs against a full set of available predictors. Modeling the full set tests the explanatory power of the three variables of interest when they are situated in the ecological complexity afforded by the Ontario dataset. A second logistic regression model using only gender and prior achievement as predictors will follow. Evidence of gender paradox, if present, will be apparent when gender, prior achievement and Grade 9 enrollment descriptive and modeled results are compared.

Logistic regression models differ from other regression models in that they are designed to accommodate categorical and limited dependent variables (Long, 1997; Tutz, 2012). This means that whereas regression models such as ordinary least squares are linear, most logistic regression solutions are not. And non-linearity presents unique challenges when it comes to interpreting modeled results that go beyond significance attributions. Students who enroll in Grade 9 academic and applied programs, for example, are converted to a continuous scale between 0 and 1, which is interpretable as the probability or simple odds.

Method

Grade 3 (2005), Grade 6 (2008), and Grade 9 (2011) Ontario large-scale mathematics test results from a single student cohort served as data. This included 106,473 English language student records (French language records were excluded) and 18 variables: gender (0 = male, 1 = female), achievement (Grades 3, 6, 9; levels 0.1 to 4.9), Grade 9 stream (applied, academic), first language (English, non-English), English as a Second Language (ESL) program (Grades 3, 6; not enrolled, enrolled), formal Identification and Placement Review Committee designation (Grades 3, 6; not exceptional, exceptional), Individual Education Plan (IEP) (Grades 3, 6; no IEP, IEP), special education program (Grades 3, 6; no placement, placement), student attitude responses about whether they liked mathematics (Grades 3, 6; yes, I like math; sometimes I like math; no, I do not like math), student attitude responses about whether they were good at math (Grades 3, 6; yes, I am good at math; sometimes I am good at math; no, I am not good at math). Two attitude variables and ESL program enrollment were the only predictors that included small numbers of missing values.

Descriptive Analysis

Numbers of males and females by grade and program, mean achievement, independent samples t-test confidence intervals (Wald statistics, 99% confidence level), effect sizes, effect size standard deviations, and variance ratios were computed. Results were summarized and presented in a table.

Logistic Regression

Two logistic regression models were estimated. A Full model expressed the log-odds of Grade 9 academic enrollment as a linear combination of all available predictors (rms package, R Development Core Team, 2015). Full model results provided a baseline to gauge the explanatory power of gender and prior achievement as predictors in the full dataset. A second logistic regression, the Predictive model, followed. This model included only the three variables of interest. It is reasonable to assume that if the interaction between variables of interest remains stable across models this stability will be reflected as similar intercept values. Comparison of models, therefore, is an important test. Both Full and Predictive model fit were also independently estimated using relative goodness of fit tests such as the likelihood ratio test and Brier, Somer's *Dxy*, and *ROC* discrimination indices.

Results

Descriptive Findings

Table 1 summarizes descriptive statistics for the Ontario data. Sample size (n) and mean mathematics achievement ($\bar{x}$) are computed by Grade and by gender. Mean achievement difference confidence intervals (99%CI), effect sizes, their associated standard deviations (d(SD)), and variance ratios (VR) are computed by Grade. Despite a balanced sex ratio in Grades 3 and 6, mean mathematics achievement and gender differences confidence intervals indicate a slight male advantage, a conclusion also reflected by differences in respective effect sizes. Variance ratio magnitudes, meanwhile, suggest there is more variability in male achievement distributions than females. Males outnumber females at the achievement extremes—(1) Grade 3:

Table 1

Descriptive Statistics for Ontario School Data

Grade	Gender	*n*	$\bar{x}$	99%CI $[\bar{x}_{male} - \bar{x}_{female}]$	*d*(SD)	VR
3	male	53,242	3.275			
	female	53,231	3.246			
	Total	106,473		[0.018,0.041]	0.048(0.006)	1.102
6	male	53,242	3.249			
	female	53,231	3.231			
	Total	106,473		[0.007, 0.029]	0.026(0.006)	1.05
9 Applied	male	15,623	2.891			
	female	13,014	2.789			
	Total	28,637		[0.076, 0.129]	0.118(0.009)	1.082
9 Academic	male	37,619	3.404			
	female	40,217	3.355			
	Total	77,836		[0.037, 0.060]	0.078(0.006)	1.037

achievement level < 1.0, n_{male} = 954 vs. n_{female} = *706*; achievement level > 4.0, n_{male} = 6068 vs. n_{female} = 5288 and (2) Grade 6: achievement level < 1.0, n_{male} = *263* vs. n_{female} = *239*; achievement level > 4.0, n_{male} = 7045 vs. n_{female} = 6308—while females outnumber males in the middle portion of the achievement range (Figure 1).

By comparison, Grade 9 findings do not reflect a balanced sex ratio as 2,609 more males enrolled in the applied stream and 2,598 more females enrolled in the academic stream (Table 1). Despite this difference, there remains a small but robust mean achievement advantage favoring males in both streams, and this is reflected in mean achievement difference confidence intervals and effect size estimates. Once again, variance ratio statistics indicate greater distributional achievement variance among males which is associated with greater numbers of males at the achievement extremes—(1) Grade 9 applied stream: achievement level < 1.0, n_{male} = 610 vs. n_{female} = 485; achievement level > 4.0, n_{male} = 1487 vs. n_{female} = 805 and (2) Grade 9 academic stream: achievement level < 1.0, n_{male} = 123 vs. n_{female} = 86; achievement level > 4.0, n_{male} = 4177 vs. n_{female} = 3768. As with elementary findings, greater numbers of females occupy the middle achievement range (Figure 2). To summarize, there is surprising regularity in descriptive findings across Grades that is consistent with the gender paradox hypothesis.

Logistic Regression Models

Full model. Full model results are summarized in Table 2, columns 2-5. Regression estimates and standard errors of regression appear in columns 2 and 3 (Est and SE respectively). Wald statistical significance test results appear in the fourth column (Pr(> |z|)). Predictors appear in the table rows, organized in descending order of their contribution to explained model deviance reduction (column 5, % dev) as computed in a separate analysis of variance.

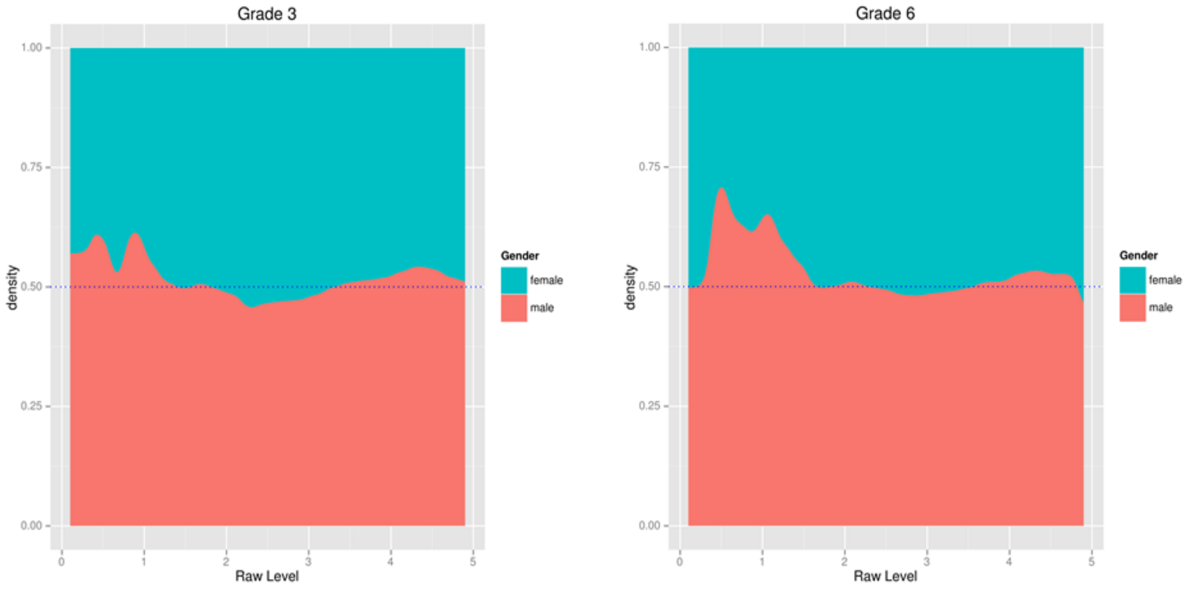

Figure 1. Elementary stackplots illustrating relative numbers of females and males across the achievement spectrum.

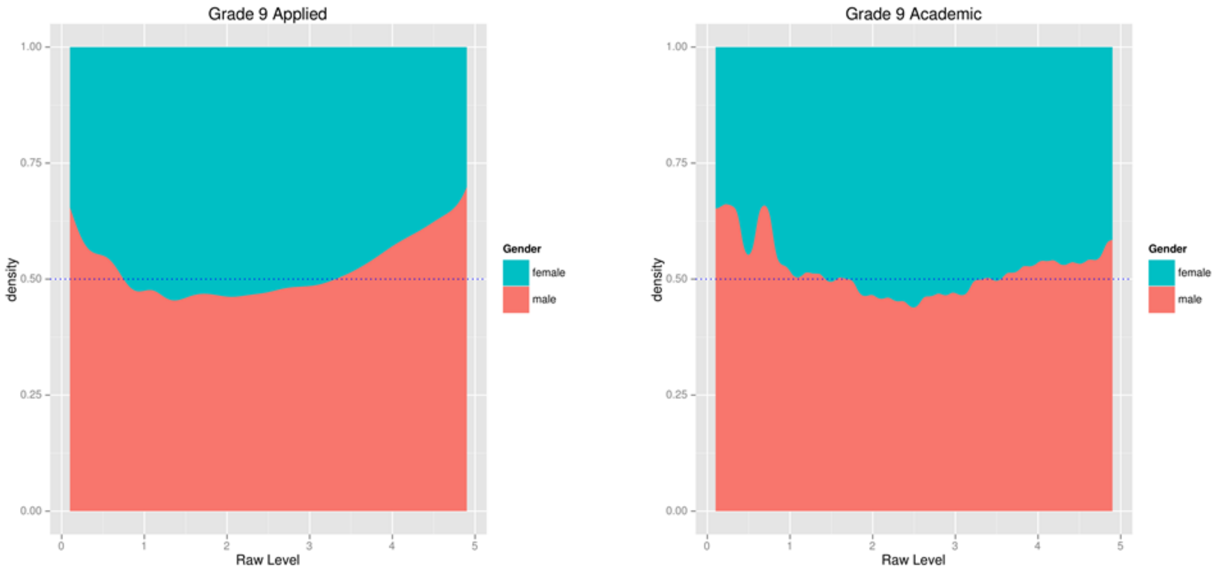

Figure 2. Grade 9 applied and academic program stackplots illustrating relative numbers of females and males across the achievement spectrum.

Ten variables and two interaction terms significantly predict the likelihood that students enroll in Grade 9 academic programs. Grade 3 achievement accounts for the largest deviance reduction (11.7%) while Grade 6 achievement accounts for the next largest (8.2%). Gender and achievement interaction are much farther down the list with smaller model deviance reductions (0.3% and 0.1% respectively). The eight remaining significant predictors include Grades 3 and 6 responses to "I am good at math", first language, Grade 6 IEP identification, Grade 6 responses to "I like math", Grade 3 IPRC identification, and Grade 6 special education involvement. Of these, Grade 6 IEP identification is the strongest predictor. Grade 3/6 interaction to "I am good at math" questions is also a significant model predictor but, as the vanishingly small deviance reduction of 0.02% clearly attests, it had little influence over the model as a whole.

The likelihood ratio test statistic for the full model is 20154.4 with 23 degrees of freedom and associated with a significant p-value. This suggests that modeled results are an improvement over the intercept only model. Moreover, the Brier statistic is 0.123, Somer's *Dxy* is 0.750, and the ROC value is 0.875, all indicating reasonably good fit between observed and predicted values.

Predictive model. Predictive model results are summarized in Table 2, columns 6-9. All estimates are associated with significant t-test results. Grades 3 and 6 achievement account for 11.74% and 8.19% reductions in model deviance while gender and Grade 3/6 achievement interaction account for 0.26% and 0.15%, respectively.

The likelihood ratio test statistic is 35622.4 with 4 degrees of freedom and is associated with a significant p-value. Although noticeably larger than the related likelihood ratio statistic in the Full model, this still indicates that gender and achievement significantly improved model fit over the intercept only model. Meanwhile, the Brier statistic (0.134), Somer's *Dxy* (0.707), and ROC (0.854) all indicate good fit between observed and predicted values although not as convincingly as was observed in the Full model.

Table 2

Full and Predictive logistic regression model results

	Full Model				Predictive Model			
Variable	Est	SE	Pr(> |z|)	% dev	Est	SE	Pr(> |z|)	% dev
NULL	-1.708	0.230	<0.0001	--	-4.958	0.189	<0.0001	--
G3 Level	-0.364	0.072	<0.0001	11.68	-0.267	0.063	<0.0001	11.74
G6 Level	0.672	0.075	<0.0001	8.20	1.075	0.062	<0.0001	8.19
G3 IEP	-0.744	0.913	0.41	2.79				
G3 GoodMath	-0.335	0.042	<0.0001	2.56				
G6GoodMath	-0.693	0.040	<0.0001	1.96				
First Language	0.838	0.030	<0.0001	1.22				
G6 IEP	2.397	0.330	<0.0001	1.20				
G6 LikeMath	-0.099	0.028	<0.0001	1.14				
G3 IPRC	0.370	0.083	<0.0001	0.91				
G6 Spec Ed	-3.670	0.329	<0.0001	0.71				
G3 LikeMath	-0.056	0.033	0.09	0.69				
G3 Spec Ed	-0.068	0.913	0.94	0.49				
Gender	0.505	0.019	<0.0001	0.26	0.408	0.017	<0.0001	0.26
G6 IPRC	-0.055	0.058	0.34	0.21				
INT G3:G6 Level	0.325	0.024	<0.0001	0.15	0.297	0.021	<0.0001	0.15
G3 ESL	0.088	0.055	0.11	0.07				
INT G3:G6 IEP	0.882	1.045	0.40	0.03				
INT G3:G6 GoodMath	0.092	0.022	<0.0001	0.02				
G6 ESL	0.211	0.098	0.03	0.00				
INT G3:G6 ESL	-0.233	0.127	0.07	0.00				
INT G3:G6 LikeMath	0.029	0.015	0.05	0.00				
INT G3:G6 IPRC	-0.275	0.116	0.02	0.00				
INT G3:G6 Spec Ed	-0.322	1.045	0.76	0.00				

Interpreting the influence of Grades 3 and 6 achievement levels (Table 2: G3 Level, G6 Level) is complicated by the achievement interaction term. Grade 3 achievement is negatively associated (i.e., -0.267) while Grade 6 achievement and achievement interaction are positively associated with logit estimates for Grade 9 academic program enrollment at 1.075 and 0.297, respectively. The partial effect of Grade 3 prior achievement on Grade 9 academic enrollment ranges from a minimum of -0.237 when Grade 6 achievement is 0.1 to 1.188 when Grade 6 achievement is 4.9. Points along this line are positive, therefore, when Grade 6 achievement level is greater than 0.9. The partial effect of Grade 6 achievement on Grade 9 academic program enrollment, meanwhile, ranges from a minimum of 1.105 when Grade 3 achievement is

0.1 to a maximum of 2.530 when Grade 3 achievement is 4.9. Overall, therefore, Grades 3 and 6 achievements are positively associated with Grade 9 academic access.

Interpreting the effect of gender on Grade 9 academic enrollment is straightforward as the intercept estimate for gender is positive (0.408). Converting this value to simple odds indicates that females are, on average, 1.5 times more likely to enroll in the Grade 9 academic stream than are males.

Comparison of Logistic Models

Full model results account for greater model deviance reduction than Predictive model results (34.30% and 20.30% respectively, Table 2). Yet estimate magnitudes and deviance reduction estimates are about the same when compared across models (e.g., Estimate values: -1.708 vs. -4.985, Grade 3 level: -0.364 vs. -0.267, Grade 6 level: 0.672 vs. 1.075, gender: 0.505 vs. 0.260, Grades 3/6 interaction: 0.325 vs. 0.297). Relative stability of estimates suggests that the relation between gender, achievement, and access remains stable across the two models. This not only affirms the existence of a relation between principal variables, but comparing these results with Table 1 also confirms that this relation is paradoxical.

Discussion

I have argued that an understanding of the relation between gender, mathematical ability and postsecondary access is methodologically fraught because related research is mostly carried out in separate literatures. But even in cases where they have been studied together, the conceptual maw between school achievement and postsecondary choice has arguably stymied interpretation (e.g., Hango, 2013, 2014). So while it may not be surprising that comparisons of existing literatures suggest gender paradox, previously unrecognized dependencies may yet become apparent if the conceptual distance were reduced. This is the rationale motivating the current design. Hence, all variables were defined relative to the Ontario school context: enrollment in Grade 9 academic mathematics programs as a proxy for access; prior achievement as a proxy for ability. School achievement presumably says something about ability and this is, in turn, likely related to success with enrolling in programs requiring such ability. Thus I initially hypothesized that in a simpler single context model, achievement and program enrollment would be positively related and this would lead to less evidence of gender paradox. But the hypothetical corollary—namely that these variables are negatively related or unrelated—was also possible. In this case, variables interact with gender less predictably with results leading to evidence of paradox.

Gender Paradox and the Single Context Model

Results of the current study clearly and consistently indicate a paradoxical association between gender, mathematics achievement and Grade 9 academic program enrollment. Comparison of Full and Predictive modeled results, moreover, indicate that despite the influence of competing indicators in the Full model, gender and prior achievement remain stable and meaningful predictors of Grade 9 program enrollment in the simpler model. This conclusion is supported by descriptive results where an equal gender ratio notwithstanding, more females are found in Grade 9 academic programs while more males are found in Grade 9 applied programs. In the

Ontario school context, therefore, we can conclude like Hango (2014) that students' eventual Grade 9 program choice is not necessarily associated with prior achievement. Unlike Hango, however, there is no obvious way to explain the gender discrepancy.

Evidence of gender paradox in the Ontario school data reveals an important theoretical dilemma. For if females are, on average, not as cognitively capable in mathematics yet are more successful accessing programs requiring such ability (a negative ability/positive access bias) then males are, on average, more capable in mathematics but are less successful accessing programs requiring ability (a positive ability/negative access bias). On the face of things, it would appear, to embrace paradox is to accept that the relation between ability and access differs by gender. But looking at this conclusion a little more closely it is apparent that evidence in the current study and, indeed, comparisons of earlier gender and mathematics studies, rest on untested assumptions. Unlike sex, which was indelibly codified early on in the Ontario dataset, ability and access are theoretical constructs and, as such, are only interpretable via measurable proxies. This means that while mathematics ability has been explained as a product of innate cognitive and/or social factors favoring certain groups—e.g., Hedges & Nowell, 1995; Hyde et al., 2008; Lubienski et al., 2013; Voyer & Voyer, 2014—these conclusions rest entirely on achievement evidence. But, as argued, achievement definitions are not consistently applied in gender and mathematics ability research, bringing into question the wisdom of achievement as the mainstay for a theory of ability. Similar challenges emerge when program enrollment is used as the sole proxy for a theory of access (hence conclusions about relative group success). Indeed, evidence of paradox arising from achievement and enrollment information in a simpler model underscores the need for a fresh approach.

Gender, Achievement and Claims about Ability

Although an achievement difference favoring males is statistically significant throughout the Ontario data, effect size magnitudes are small enough that they could also be interpreted in the negligible region of Cohen's scale (Table 1; Cohen, 1977). Interpretation is a critical consideration because it ultimately determines the theoretical tenor of subsequent claims. For on the one hand, Ontario findings concur with studies using large-scale achievement test results to substantiate ability difference claims—studies increasingly bolstered by probabilistic sampling and robust statistical tests (e.g., Hedges & Nowell, 1995; Hussar & Bailey, 2014; Lubienski et al., 2013). Test takers' scores are ranked relative to one another so we can say that ability difference claims rest on a norm-referenced warrant. On the other hand, gender similarities claims rest on results interpreted relative to Cohen's scale—a criterion-referenced warrant (e.g., Hyde et al., 1990, 2008). Gender similarities studies otherwise present just as compelling a case by bolstering claims with probabilistic sampling and robust statistical tests (e.g., Lindberg et al., 2010). So as things stand in the large-scale gender and mathematics test literature, dueling ability theories rest on indisputable normative evidence of achievement difference and equally indisputable criterion-referenced evidence of achievement similarity.

Yet a better understanding about the relation between gender and mathematics ability is at the heart of many of these papers. This is contentious, however, as the connection between achievement and ability is not well understood. Hedges and Nowell (1995), for example, conceded that while their findings were consistent with an innate male ability advantage, a normative interpretation of achievement scores does not shed much light on the nature of ability difference. Addressing this question, they continued, requires further research. Pope et al.

(2006) similarly cautioned against interpreting achievement results as indicative of ability in the absence of additional information about mathematical tasks. Devine, Fawcett, Szucs, and Dowker (2012) speculated that gender ability differences may be rooted in mathematics anxiety (discomfort associated with performing mathematics tasks) and test anxiety (discomfort associated with completing tests) differences; girls being more prone to elevated mathematics, and mathematics test, anxiety. Even Hyde and colleagues warned against interpreting statistical analysis of gender achievement scores in the absence of information about differential performance in such things as problem solving, spatial reasoning, and computation (Hyde et al., 1990, 2008; Lindberg et al., 2010).

Despite these shortcomings achievement remains a basic research metric to support theoretical claims about mathematics ability. But by their nature achievement scores are aggregate values summarizing one's total test experience. They have otherwise lost any association with individual mathematical tasks comprising the test. In other words, achievement scores are non-unique with respect to ability. An Ontario leveled achievement score of 3, say, although likely reflecting something of ability, situates all students with respect to their provincial peers along a common four-point scale. The score otherwise retains no item-level information about such things as content, format, required mathematics knowledge, or skill. Consequently, it is possible to receive a score of 3 in multiple ways because there are many item-level response permutations that can possibly aggregate to 3; each with different content and cognitive implications (Harnisch, 1983). Regardless of scale, representativeness of sampling, or robustness of tests, there is no way to unequivocally map achievement scores back to item-level knowledge and skills. Hence, gender and mathematics ability research claims rest on a non-unique measure of group standing—an external indicator—when a unique measure for cognitive ability—an internal indicator—is actually required. As long as achievement remains the basic metric for ability, therefore, we will never achieve sufficient analytic granularity to settle any claims about gender and its relation to mathematics ability.

The same criticism, moreover, extends to gender and mathematics findings that are not reliant on large-scale test interpretations of achievement. Studies using teacher-generated achievement scores, for example, report effect size magnitudes and establish theoretical claims that clearly challenge the large-scale test interpretation (e.g., Duckworth & Seligman, 2006; Voyer & Voyer, 2014). This challenge, however, ignores the fact that teacher-generated achievement scores, when used in analysis, are also summary scores and every bit as inscrutable with respect to item-level mathematical content and cognitive demand as test-generated scores. Instead of presenting a viable counterclaim, teacher-generated findings merely introduce an alternative way of defining achievement, claims about female ability advantage being otherwise based on the same kind of normative statistical analysis. So we are left to differentiate between a large-scale test interpretation and a teacher-generated interpretation. But the existence of multiple interpretations, rather than settling matters, actually reinforce the need to challenge the appropriateness of using achievement, however defined, in the first place.

Gender, Achievement Distribution and Claims about Ability

Ontario school results reveal distributional variance magnitudes favoring males that differ slightly from earlier studies (Table 1, Hedges & Nowell, 1995; Lindberg et al., 2010). There are greater numbers of males at both achievement extremes—Hedges and Nowell reported greater numbers of males only at the upper extreme—and there is a persistent tendency for females to dominate in the middle achievement range, a finding not reported elsewhere (Figures 1, 2). But interpretation is complicated because distributional variance findings are themselves derived from achievement scores. And this means that distributional variance claims are vulnerable to the same criticisms. It is possible, for example, that males and females respond differently to standardized testing conditions—(e.g., Devine et al., 2012; Voyer & Voyer, 2014)—or that these differences arise from, as yet, an only partially realized set of social, environmental, and cognitive factors (e.g., Halpern, 2012). For like achievement, distributional variance values merely establish that difference exists but are otherwise not sufficiently granular to address more nuanced questions about the nature of this difference.

Gender, Enrollment and Claims about Access

Ontario girls are 1.5 times more likely to enroll in Grade 9 academic mathematics programs than boys, a finding that is consistent with gender and postsecondary enrollment results reported elsewhere (e.g., Hango, 2014; Hussar & Bailey, 2014). I initially argued that student enrollment in Grade 9 academic mathematics is a reasonable proxy for access because entry to the academic stream is necessary for the bulk of those wishing to pursue later mathematics courses and, ultimately, postsecondary study. As it turned out, 73% of the sample enrolled in academic programs, so it is accurate to say that the majority of Ontario students had opportunity.

Enrollment is a reasonable proxy for access only if we have faith that academic enrollment is meritoriously adjudicated. Streaming, however, remains controversial. Taylor et al. (2009) argued that streaming in Canada grew out of the practice of grouping by ability as a way of assisting students to find their proper place in society (see also, Ireson & Hallam, 2001). But there are undoubtedly many reasons why students opt for academic Grade 9 courses and why parents, teachers, and administrators might encourage such choices. Proponents argue that students do better in school if they are grouped with other children most like themselves while critics contend that streaming actually increases inequity. Minority students and those who come from economically disadvantaged households disproportionately populate lower ability groups while Caucasians and those from wealthier households populate upper ability groups. But as gender was the focus, ethnic and socioeconomic data were not collected in the current study so it is difficult to determine whether enrollment was influenced by such factors.

Future Directions

Paradoxical findings indicate the care with which achievement and enrollment metrics should be used. Achievement, for example, is perfectly suitable as an external metric when used to sort students relative to their peers but not as well suited when used as an internal metric to sort students by cognitive attributes (such as ability). Likewise, enrollment in a secondary school stream is useful for sorting students relative to their peers but less useful when considering attributional differences.

Future gender and mathematics studies could use item-level instead of summary analysis methods. An item-level analysis can provide a probabilistic solution to the estimation of ability while, at the same time, linking results back to specific mathematics content and skills. This offers not only a basis from which to develop a theory of mathematics ability but also its relation to gender. It would be possible, for example, to test whether females and males perform differently when asked to address text-heavy problems versus spatial reasoning tasks (c.f., Lindberg et al., 2010). Number and type of mathematics courses students opt to take in later grades, meanwhile, might be better measures for access than Grade 9 academic enrollment. Then an item-level analysis of student ability could be combined with mathematics course enrollment in a single ordinal regression model. Indicators for Grades 3, 6, and 9 (divided by Grade 9 streams) ability could serve as predictors of the number of courses students voluntarily participate, and successfully complete, in Grades 10 through 12. This represents a nuanced model, one that could be used to construct a theory of ability from which other grouping characteristics such as cognitive strand (e.g., spatial reasoning vs. problem solving), question format (e.g., multiple choice vs. open response), and achievement definition (e.g., large-scale test vs. teacher-generated results) could be tested.

References

Brochu, P., Deussing, M.-A., Houme, K., & Chuy, M. (2013). *Measuring up: Canadian results of the OECD PISA Study: The performance of Canada's youth in reading, mathematics, and science*. Retrieved from http://www.cmec.ca/Publications/Lists/Publications/Attachments/318/PISA2012_CanadianReport_EN_Web.pdf

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (revised). New York: Academic Press.

Devine, A., Fawcett, K., Szucs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions, 8*(33), 1-9.

Duckworth, A. L. & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98*(1), 198-208.

Education Quality and Accountability Office. (2011). *EQAO's technical re- port for the 2010-2011 assessments: assessments of reading, writing and mathematics, primary division (grades 1-3) and junior division (grades 4-6); grade assessment of mathematics and Ontario secondary school literacy test*. Toronto: Queen's Printer for Ontario.

Education Quality and Accountability Office. (2012). *About the EQAO*. Retrieved from http://www.eqao.com/AboutEQAO

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103.

Guskey, T. R. (2013). Defining student achievement. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 3-7). New York: Routledge.

Halpern, D. F. (2012). *Sex differences in cognitive abilities*. East Sussex, UK: Psychology Press.

Hango, D. W. (2013). *Gender differences in science, technology, engineering, mathematics and computer science (stem) programs at university*. Statistics Canada.

Hango, D. W. (2014). *Ability in mathematics and science at age 15 and program choice in university: Differences by gender*. Beaconsfield, QC: Statistics Canada/Canadian Electronic Library.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20*(2), 191-206.

Hedges, L. V. & Friedman, L. (1993). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research, 63*(1), 94-105.

Hedges, L. V. & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41-45.

Hussar, W. J. & Bailey, T. M. (2014). Projections of education statistics to 2022. NCES 2014-051. *National Center for Education Statistics*.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using *w2*and *d*. *American Psychologist, 36*(8), 892.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321*(5888), 494–495.

Ireson, J. & Hallam, S. (2001). *Ability grouping in education*. London: Sage.

Jacob, B. A. (2002). Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review, 21*(6), 589-598.

Knapp, L. G., Kelly-Reid, J. E., & Ginder, S. A. (2012). Enrollment in post- secondary institutions, fall 2010; financial statistics, fiscal year 2010; and graduation rates, selected cohorts, 2002-07. First look. NCES 2012-2080. *National Center for Education Statistics*.

Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: A review. *Educational Research, 41*(1), 63-76.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123-1135.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Lubienski, S. T., Robinson, J. P., Crane, C. C., & Ganley, C. M. (2013). Girls' and boys' mathematics achievement, affect, and experiences: Findings from ECLS-K. *Journal for Research in Mathematics Education, 44*(4), 634-645.

Miller, D. I. & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences, 18*(1), 37-45.

Ministry of Education. (2005a). *The Ontario curriculum: Grades 1–8 mathematics*. Retrieved from http://www.edu.gov.on.ca/eng/curriculum/elementary/math18curr.pdf

Ministry of Education. (2005b). *The Ontario curriculum: grades 9 and 10 mathematics*. Retrieved from http://www.edu.gov.on.ca/eng/curriculum/secondary/math910curr.pdf

Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Boston College, TIMSS & PIRLS International Study Center; International Association for the Evaluation of Educational Achievement. (ED544554).

Peter, K. & Horn, L. (2005). Gender differences in participation and completion of undergraduate education and how they have changed over time. Postsecondary education descriptive analysis reports. NCES 2005-169. *US Department of Education*.

Pope, G. A., Wentzel, C., Braden, B., & Anderson, J. (2006). Relationships between gender and Alberta achievement test scores during a four-year period. *Alberta Journal of Educational Research, 52*(2006): 4-15.

R Development Core Team. (2015). *R: A language and environment for statistical computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Riegle-Crumb, C. (2010). More girls go to college: exploring the social and academic factors behind the female postsecondary advantage among Hispanic and white students. *Research in Higher Education, 51*(6), 573-593.

Robinson, J. P. & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268-302.

Rogers, W. T., Ma, X., Klinger, D. A., Dawber, T., Hellsten, L., Nowicki, D., & Tomkowicz, J. (2006). Examination of the influence of selected factors on performance on Alberta learning achievement tests. *Canadian Journal of Education, 29*(3), 731-756.

Russo, P., Barbaranelli, C., & Caponera, E. (2014). The influence of broad and specific personality traits on mathematics achievement. *Personality and Individual Differences, 60*, S73.

Shulman, R., Hinton, A., Zhang, S., & Kozlow, M. (2014). *EQAO research: Longitudinal results of province-wide assessments in English-language schools: Trends in student achievement and implications for improvement planning in mathematics and literacy*. Toronto: Queen's Printer for Ontario.

Stoet, G. & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of PISA data. *PLoS One, 8*(3), e57988.

Syzmanowicz, A. & Furnham, A. (2011). Gender differences in self-estimates of general, mathematical, spatial and verbal intelligence: Four meta analyses. *Learning and Individual Differences, 21*(5), 493-504.

Taylor, A., Krahn, H., & Levine-Rasky, C. (2009). Streaming in/for the new economy. In C. Levine-Rasky (Ed.) *Canadian perspectives on the sociology of education* (pp. 103-123). Don Mills, ON: Oxford Press.

Tutz, G. (2012). *Regression for categorical data*. Cambridge: Cambridge University Press.

Voyer, D. & Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychological Bulletin*. doi: http://dx.doi.org/10.1037/a0036620

Malcolm Cunningham is an Adjunct Research Professor in the Institute of Cognitive Science at Carleton University. His interests include a better understanding about how to define and measure mathematics ability, the association between ability and learning and what can be done to improve current practice.