REPLY

Are Schools Shortchanging Boys or Girls? The Answer Rests on Methods and Assumptions: Reply to Card (2014) and Penner (2014)

Joseph P. Robinson-Cimpian and Sarah Theule Lubienski
University of Illinois at Urbana-Champaign

Yasemin Copur-Gencturk
University of Houston

Our target article (Robinson-Cimpian, Lubienski, Ganley, & Copur-Gencturk, 2014) used nationally representative data to examine the development of gender gaps in math achievement. We found that when boys and girls demonstrate equivalent math test performance and are perceived by their teachers to be equally well behaved and engaged with the material, teachers tend to rate girls as less proficient in math than boys (Study 1). Moreover, this underrating of girls’ proficiency appeared to contribute substantially to a widening gender gap in early elementary school (Study 2). In this response, we use the thoughtful comments of Card (2014) and Penner (2014) as a springboard for discussing the methodologies and assumptions of some of the most recent research using nationally representative data to explore gender inequities. In the process, we shed light on how recent works using the same data reach different conclusions. We also make recommendations regarding the use of such data for understanding the development of the gender gap and for designing effective interventions.

Keywords: achievement gap, gender, mathematics, elementary school, quasi-experimental designs

Do schools comparatively advantage boys or girls? At the root of the debate around this question is the more basic question of whether there are innate cognitive differences between boys and girls, and the extent to which schools perpetuate or reduce such differences. Although most researchers acknowledge interplay between biology and environment, the current debate is characterized by the degree to which the environment is hypothesized to play a role (Halpern et al., 2007). Thus, there are two predominant views in this debate: The first view is that boys and girls, on average, have similar cognitive abilities and that gender differences that emerge in achievement stem primarily from environmental factors (see, e.g., Bussey & Bandura 1999; Hyde, 2005, 2014; see also Beilock et al., 2010; Else-Quest et al., 2010). The second view is that there are innate cognitive differences between boys and girls (see, e.g., Baron-Cohen, 2002; Geary, 1996, 1998), with some scholars arguing that “noncognitive” differences between boys and girls may help to mitigate the manifestation of cognitive differences (e.g., Cornwell, Mustard, & Van Parys, 2013; DiPrete & Jennings, 2012). The first viewpoint has support from a range of studies, including international comparisons showing wide variability in gender achievement gaps (Else-Quest et al., 2010) and smaller math gaps in countries with greater cultural gender equality (Guiso et al., 2008), and studies demonstrating that larger gender achievement gaps are found in populations that hold stronger gender-stereotypic views (across states, Pope & Sydnor, 2010, and nations, Nosek et al., 2009). The second viewpoint has support from studies showing males outperforming females in advanced

1 Note, however, that even proponents of a model placing greater emphasis on biological differences often acknowledge some differences arising from a social influence. For example, Geary (1996, p. 245) noted that a teacher’s style of interaction with students can influence the math gains of boys and girls differentially.

2 Noncognitive is a term used by some economists and sociologists to describe skills such as perseverance, attentiveness, and self-control.
math courses (Benbow & Stanley, 1983), greater variability in males’ test scores (Hedges & Nowell, 1995), and links between markers of prenatal testosterone exposure (e.g., finger-length ratios) and young children’s math and reading performance (Brosnan, 2008; Fink, Brookes, Neave, Manning, & Geary, 2006). Both sides of the debate often rely on correlational evidence and typically smaller scale experimental studies. But, is there other evidence that can shed light on this debate?

In the next section, we propose the expanded use of nationally representative data sets and quasi-experimental techniques to inform our understanding of gender differences in math. In the process, we examine conflicting interpretations on how to remedy gender gaps from researchers using such data sets. In subsequent sections, we discuss how the use of quasi-experimental and measurement-focused techniques can be simultaneously considered when performing analyses of developmental processes, and we reflect on the implications of the target article for interventions.

**Nationally Representative Data Sets and Quasi-Experimental Methods: Potential for New Insights**

As Card (2014) notes, nationally representative data sets are underused resources for exploring the potentially causal effects of social and biological factors. In our target article (Robinson-Cimpian, Lubinski, Ganley, & Copur-Gencturk, 2014), we used one such data set (Early Childhood Longitudinal Study—Kindergarten Class of 1998-99 [ECLS-K]) to study the effects of teachers’ perceptions on the development of the gender gap in math test performance. Although many studies using the ECLS-K to study gender differences have focused on descriptive relations (see, e.g., Fryer & Levitt, 2010; Robinson & Lubinski, 2011), quasi-experimental modeling techniques can be applied to gain insights into plausibly causal pathways.

In this response to commentaries, we begin by illustrating the importance of causal analysis with large data sets, discussing two recent articles on gender differences that also used the ECLS-K data set but reached very different conclusions from our own. Both studies—Cornwell et al. (2013) and DiPrete and Jennings (2012)—explored teacher ratings using analyses similar to our Study 1. In all of these studies (including our own), the researchers estimated the gender gap in math test scores before accounting for measures of behavior and engagement, and then again after accounting for them. The models accounting for behavior and engagement showed greater losses over time for girls in math test performance. DiPrete and Jennings (2012) interpreted this change as indicating that “the math gap would be . . . larger . . . but for the female advantage in social and behavioral skills” (p. 8). This is a reasonable interpretation, but without continuing to examine causal mediators of the growing gap, DiPrete and Jennings’s analysis leaves us wondering why the gap would be larger. Would it be larger because girls have lower cognitive abilities (but perhaps greater noncognitive abilities), or would the gap be larger because of some external/social source such as parent or teacher expectations?

Although Cornwell et al. (2013) conducted a similarly descriptive (i.e., noncausal) analysis, they went on to provide a stronger interpretation:

> [If] young girls display a more developed ‘attitude toward learning’ and teachers (consciously or subconsciously) reward these attitudes by giving girls higher marks than warranted by their test scores, the seeds of a gender gap in educational attainment may be sown at an early age, because teachers’ grades strongly influence grade-level placement, high school graduation, and college admission prospects. Consequently, our results may spur further educational innovation at the early grade levels, such as developing ways to improve boys’ noncognitive skills, creating alternative methods of instruction to communicate more effectively to boys who have different noncognitive skill sets, and experimenting with single-gender instruction. (p. 263)

Although we agree with Cornwell et al. (2013) that finding ways to improve boys’ noncognitive skills is a beneficial avenue for future research and intervention attempts, we believe some aspects of this statement deserve greater scrutiny. First, although engagement predicts teachers’ ratings, the notion that teachers “reward” girls runs counter to our findings. We find that girls are, indeed, rated as being more engaged with the material and persevering at tasks—attributes associated with greater learning (see, e.g., Blackwell, Trzesniewski, & Dweck, 2007; Duckworth, Kirby, Tsukayama, Berstein, & Ericsson, 2011; Dweck, 2006). However, once we account for differences in behavioral and engagement ratings (ratings that are unlikely to reflect bias in terms of teachers artificially rating girls higher; see DiPrete & Jennings, 2012), we find that teachers rate boys’ math proficiency higher. Moreover, this higher rating of boys (compared with similarly behaving and performing girls) occurs throughout the range of math achievement (see Figure 1). Therefore, if there is any “rewarding” suggested by the ECLS-K data, it favors boys when comparing students with similar performance, behavior, and engagement. In other words, girls have to be perceived as more engaged with the material to be rated equal to boys in math proficiency. Hence, we

---

3 It may appear as if we reached a similar conclusion to that of Cornwell et al. (2013) in a previous article (Robinson & Lubinski, 2011). In that article, we stated that teachers might conflate behavior with learning, resulting in an overrating of girls’ math proficiency. We concluded that this overrating might actually hurt girls if teachers mistake girls’ good behavior for learning new material, which might explain why girls lose ground to boys in math during elementary school. (Note that Robinson & Lubinski, 2011, did not account for any behavioral factors.) In the target article, we explored these relationships in a series of rigorous analyses. Among them, we matched girls and boys on dimensions such as test performance and perceived behavior and engagement and found that teachers actually underrated girls’ math proficiency. Thus, in the target article, we found no support for our initial hypothesis in Robinson and Lubinski (2011) that an overrating of their proficiency might hurt girls; in fact, we found that an underrating of their proficiency hurts girls.

4 Cornwell et al. (2013) and DiPrete and Jennings (2012) performed analyses similar to ours and observed similar changes in the estimated gender gap. The difference is that they do not interpret the resulting pattern of underrating girls as suggestive of bias against girls. By contrast, we interpret this pattern as bias for two reasons. First, intuitively, there is little reason to suspect that teachers’ ratings of math proficiency reflect actual gender differences in math achievement after matching boys and girls on measures of all prior and current math achievement, prior teacher ratings of math proficiency, prior and current behavior and engagement, and other demographics. Second, in Study 2, we are able to test empirically whether prior teacher ratings (i.e., our instrumental variables) actually predict future growth (independent of their effect on current teacher ratings), and we find no evidence to suggest that they do (see Table 3 in the target article). This is important because it provides statistical support for the idea that the ratings favoring boys are not rooted in any real knowledge that these boys will blossom but rather are simply reflecting bias teachers hold in favor of boys vis-à-vis math.
interpret this as a bias against girls, likely due to perceptions that girls have lower innate abilities. Second, Cornwell and colleagues did not perform any analyses that would warrant their recommendation for single-gender instruction. In fact, there is little to no empirical support for the effectiveness of single-gender instruction in the literature (Halpern et al., 2011). Quite the contrary—boys in co-ed classrooms with high female-to-male ratios have been shown to have higher achievement in both experimental (Whitmore, 2005) and quasi-experimental studies (Hoxby, 2000; Lavy & Schlosser, 2011) relative to boys in classrooms where there were fewer girls.

As Card (2014) noted, one strength of research using nationally representative data is generalizability, and the studies by Cornwell et al. (2013) and DiPrete and Jennings (2012) are strong in that respect. Perhaps because of this generalizability, research using such data sets is likely to attract the attention of the media (see, e.g., Sommers, 2013, in the New York Times covering Cornwell et al.). Given the attention that generalizable data attract, researchers using these data sets must strive for robust causal analyses when making claims about bias and implications about the effects of such bias on educational and life outcomes.

### A Hierarchy of Analytic Concerns

This issue of causality relates to other comments made by Card (2014) regarding measurement and, more generally, to establishing a project-specific hierarchy of modeling choices to maximize both validity and parsimony. In the target article, we used several different techniques associated with quasi-experimental designs to address limitations of working with nonexperimental data. First, we used teacher fixed effects to address concerns of nonrandom sorting of students across classes. Second, we used propensity-score matching to match boys and girls with similar characteristics and reduce the reliance on functional-form assumptions in the mediation analysis. Third, and most importantly, we used an instrumental-variables approach to address simultaneity bias concerns about the current teacher’s ratings reflecting what she saw—and what she influenced—in a student’s test score. Using teacher fixed effects and propensity-score matching did little to alter the estimated paths from gender to teachers’ ratings in the mediation analyses; however, the use of instrumental variables greatly changed the estimates—and we argued, the validity—of the paths from teachers’ ratings to math performance, and consequently of the direct paths between gender and math performance. In this instance, it would seem that the most important analytic concern to be addressed was the simultaneity bias, and lower order concerns were differential sorting and the functional-form assumption.

Card (2014) raised measurement concerns that are valid in a general sense, but which are lower priorities in the case of the target article. For example, he noted that one should account for item-specific differences in aggregate scales; however, because the ECLS-K scores for the direct cognitive assessment and the teachers’ ratings of math proficiency were calculated using item response theory, they are tapping into latent constructs and removing between-item variability. Card also noted the importance of accounting for between-student variation in the amount of time between tests (e.g., two students may take a test on the same date in kindergarten but on different dates in first grade, thereby creating a larger “lag time” for one student). We are fortunate that in the ECLS-K data set, we do not see the amount of variability in lag times as is seen in the example cited by Card (Selig, Preacher, & Little, 2012). Moreover, there is no gender difference in lag times, and our models all include the child’s age (to the day) at the date of each assessment, thereby accounting for between-child differences in lag times.

---

**Figure 1.** Teacher ratings of math proficiency by student gender and external math test score. Both panels show smoothed curves, predicted after regressions of third-grade teacher ratings on prior and current external math and reading achievement scores, all prior teachers’ ratings of the student’s math proficiency, family socioeconomic status at fall of kindergarten, age of child at each test, gender, and an interaction between gender and current external math test score. Panel B also adds covariates for all prior and current engagement (approaches to learning) and externalizing problem behavior scores. Models are weighted by the Early Childhood Longitudinal Study—Kindergarten Class of 1998-99 sampling weight. Data are described in Study 1 of Robinson-Cimpian et al. (2014) as the third-grade sample; note, however, that we further restricted the sample to include only students with complete external reading test scores (unweighted N = 3,836).
Nevertheless, some of the techniques Card (2014) suggested for modeling measurement could be incorporated into our models, after carefully considering which techniques are likely to have the most impact in addressing modeling concerns. This sort of prioritizing ensures a good balance between validity and parsimony. As a more general lesson, when modeling developmental processes in their own projects, it will be important for researchers to consider what are the most pressing analytic concerns to be addressed regarding validity while also maintaining a parsimonious model.

Can We Expect More From Teachers?

Turning away (slightly) from the issue of modeling and what evidence various models present concerning teacher bias, we now address Penner’s (2014) question of “Can we expect more from teachers?” Penner rightly cautions that by focusing on the role of teachers and schools, we run the risk of taking attention away from the role of larger society in producing gender inequities (see Lubinski, 2003, for similar arguments about social class disparities). Although we appreciate Penner’s arguments and recognize that teachers are in many ways a reflection of society in general, we argue that, within the arena of educating children, we actually should expect more of teachers than we expect of less influential individuals in society at large. Consider a different scenario, one in which a business does not promote its female employees despite equal performance relative to male employees. We would not say that businesses are part of society and therefore are justified in promoting only males. In relation not only to achievement but also to gender issues more broadly (e.g., labor-market discrimination, sexual violence), Weis (1997) wrote: “We cannot afford to take the position that schools simply mirror society and that they, therefore, have no responsibility to address the conditions under which certain forms of gender constructions and relations are formed” (p. 82). We agree with Weis on this point, particularly in this case in which there is evidence that the differential rating of girls reflects teachers’ bias (not actual differences) and affects students’ learning within the classroom. In this case, the evidence suggests that the practices within the school are contributing to gender disparities. Although perhaps it is too much to ask of teachers to reverse disparities that stem from students’ experiences outside of school, we should expect at the very least that teachers should “do no harm.”

An important, related point is the specificity of the bias revealed by our research, which may facilitate developing interventions to eliminate this bias. The bias we uncovered is specific to mathematics classrooms (i.e., not evident in reading) and specific to boys’ and girls’ abilities rather than rates the child’s task-specific proficiency (e.g., “This child divides a 3-digit number by a 1-digit number”; see the supplemental material for the target article). Thus, although we agree with Penner (2014) that we should expect more of society (not just of teachers), it is important to note that in our research, the beliefs about boys’ and girls’ abilities are not merely operating on a global level reflective of societal beliefs that women are less capable, but rather are specific to the mathematics classroom. Although such task-specific stereotypes may indeed originate from more generalized beliefs about males’ and females’ abilities, we have identified a more subtle—and perhaps more targetable—set of implicit beliefs about boys’ and girls’ proficiency. This is helpful because it suggests that, by creating interventions to uncover and disrupt teachers’ beliefs about boys’ and girls’ math abilities, we may be able to reduce the development of the gender gap in math. Though it may seem that we are unfairly putting the onus on teachers, this is clearly a much more targeted and feasible way to intervene; interventions to address larger societal issues with no clear target would be much more difficult to implement.

Another reason to intervene on teachers is that they may not be aware of their bias. We recently spoke with a group of four early elementary schoolteachers while developing an intervention inspired by the studies in the target article. One teacher said, “No offense, but I don’t know why you’re looking at gender. I have looked at the test scores, and my girls are performing just as well as my boys because the girls try hard.” As soon as she said this, she realized that she was attributing the girls’ achievement to hard work and the boys’ achievement to natural ability. The other teachers agreed that they likely have the same tendencies to attribute achievement to different sources for girls and boys (cf. Dweck et al., 1978; Tiedemann, 2000) and that these attributions may affect their expectations of boys and girls.

Conclusion

In sum, our research suggests that despite an overrating of girls on the surface, teachers rate the math proficiency of girls lower than that of similarly performing and behaving boys, and this underrating of girls likely contributes to the early development of the gender gap in math. By modeling this phenomenon in a causal mediation framework, our research improves on recent descriptive analyses and contributes more evidence that the gender gap in math is at least partly socially constructed. Our research provides strong evidence that schools are shortchanging girls by underestimating their math abilities and identifies specific targets for intervention. Although teachers are undoubtedly not alone in holding these gender-related beliefs, they have the potential to influence young minds through the vast amount of time spent with children and their position of authority. Reformers place many demands on teachers’ instructional time and energy. However, what we are envisioning would not add to teachers’ already crowded curriculum, but instead involves teacher-educators finding ways to confront biases and foster more productive beliefs among pre- and in-service teachers. This may be a small step in potentially making larger changes to society as a whole.

References


Received March 24, 2014
Accepted March 26, 2014