AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 10, 2487–2509
https://doi.org/10.1037/xge0001646

Girls Persist More but Divest Less From Ineffective Teaching Than Boys

Mia Radovanovic, Ece Yucer, and Jessica A. Sommerville
Department of Psychology, University of Toronto

Teaching is the primary way children learn about the world. However, successful learning involves recognizing when teaching is ineffective, even in the absence of overt cues, and divesting from ineffective teaching to explore novel solutions. Across three experiments, we investigated 7- to 10-year-old children's ability to recognize ineffective teaching; we tested the hypothesis that girls may be less likely than boys to divest by exploring new solutions, given documented gender differences in socialization toward conformity and obedience. Overall, we demonstrate that children independently tested taught solutions and, upon learning that the solutions were ineffective, rationally traded off between instruction and exploration. Simultaneously, gender differences in divestment emerged. On average, girls demonstrated greater persistence in applying the taught solution, while boys tended to explore their own ideas, leading to differences in solving and learning. Importantly, these differences were observable across both masculine- and feminine-stereotyped tasks. These results have important implications for children's learning and the development of leadership.

Public Significance Statement
Both boys and girls are capable of testing the effectiveness of teaching. However, on average, girls are more likely to persist in a taught solution, while boys are more likely to explore alternatives, creating gender disadvantages in success and learning when teaching is ineffective. These findings suggest interventions to help children flexibly decide when to follow teaching and to prioritize their own ideas.

Keywords: exploration, gender differences, informant reliability, innovation, people-pleasing

Supplemental materials: https://doi.org/10.1037/xge0001646.supp

One of the primary ways that people learn, particularly children, is through teaching. While we often perceive teaching as accurate and efficient, there are many real-world circumstances in which teaching fails students by imparting inaccurate or incomplete knowledge. Importantly, it is not always obvious when teaching is ineffective, leading to widespread biases and inaccuracies in public knowledge. Thus, students are often fed biased content inconsistent with scientific consensus (Fuller et al., 2022; Griffith & Brem, 2004;

This article was published Online First September 5, 2024.
Megan M. Saylor served as action editor.
Mia Radovanovic https://orcid.org/0000-0002-9142-3310
The data and analytic code for this publication are publicly available. These data can be found at the Open Science Framework at https://osf.io/qn3w5/ (Radovanovic et al., 2024).

We have complied with all relevant ethical regulations to conduct this work, which was approved by the Research Ethics Board at the University of Toronto, Protocol Title: "Social Influences on Independent Problem Exploration," ID: 00038355. Written informed consent was obtained from individuals and minors' legal guardians for both participation and the publication of any potentially identifiable images or data included in this article.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Funding for this research came from the Canadian Foundation for Innovation (The Origins of Social Learning, awarded to Jessica A. Sommerville), Graduate Women in Science (awarded to Mia Radovanovic), the Institute for Gender and the Economy at the University of Toronto's Rotman School of Management (Quantifying and understanding gender disadvantages in reactions to incorrect teaching, awarded to Mia Radovanovic), the John Templeton Foundation (Give BIG: Investigating and fostering generosity in young children, awarded to Jessica A. Sommerville), the Ontario Research Fund (The Origins of Social Learning, awarded to Jessica A. Sommerville), and the Society for Research in Child Development (Exploring the influence of people-pleasing socialization and problem-solving context on gender differences in children's adaptations to ineffective teaching, awarded to Mia Radovanovic).

The authors thank the researchers of the Toronto Early Cognition Lab at the University of Toronto for their help in participant recruitment, data collection, and coding. In particular, the authors thank Lydia Altun, Miguel Dominico Alzona, Denise Arefhaghi, Tim Wei-Ting Chao, Christie Lai, Yang (Leona) Liu, Josie Hatfield, Alexander W. McArthur, Annabelle Persaud, Aafiya Somani, Mahima Tirunelveli Santhakumar, Poorvi Sharma, Tiffany Tse, Sofia Westerhoff, Justine Vorvis, Renée Wang, Jesse Whiteman, and Grace Zheng. The authors also thank Arnav Verma for his assistance in creating an online platform to host our video game for data collection.

Mia Radovanovic played a lead role in conceptualization, data curation, formal analysis, methodology, validation, visualization, writing–original draft, and writing–review and editing and an equal role in investigation and project administration. Ece Yucer played a supporting role in data curation and writing–original draft and an equal role in investigation and project administration. Jessica A. Sommerville played a lead role in funding acquisition, resources, and supervision, a supporting role in conceptualization, and an equal role in investigation, methodology, project administration, and writing–review and editing.

Correspondence concerning this article should be addressed to Mia Radovanovic, Department of Psychology, University of Toronto, Toronto, ON M5S 3G3, Canada. Email: m.radovanovic@mail.utoronto.ca

Guttmacher Institute, 2022). One potent example is education on anthropogenic global warming. There is a nearly unanimous scientific consensus (>95%) that global warming is caused by human activities (J. Cook et al., 2016), and international policy recommendations focus on restructuring human practices (Pörtner et al., 2022). On the other hand, many Americans are unaware of the role of human activities in climate change (~40%; Hamilton et al., 2015) or of the degree of scientific consensus on the topic (~80%; Leiserowitz et al., 2020). Consequently, less than half of American schools emphasize human activity as causing climate change (Plutzer et al., 2016), and climate-denial biases slip into the language of even scientists when discussing climate change (Lewandowsky et al., 2015).

While this example is extreme, even when teaching is not misleading (i.e., omitting information about anthropogenic causes), there are often instances in which there are multiple strategies to solve a problem, some more effective than others (Star et al., 2016). Given the more general phenomenon (i.e., that teaching is not always effective), it is important that children are not merely passive recipients of knowledge but that they can actively assess and interrogate the knowledge they are given. Likewise, it may be critical for children to develop this foundational skill early on as the developmental consequences of accepting misinformation cascade with age. Here, we empirically test whether children are able to identify inaccuracies in teaching in the absence of overt cues and, importantly, whether they can compensate for inaccuracies with their own exploration and innovation in order to succeed.

Divesting From Ineffective Instruction

Teaching is central to learning from early in life because it allows for efficient knowledge acquisition through cues that highlight important information, like eye gaze and gestures (Csibra & Gergely, 2006, 2009). This is particularly true when content is difficult or impossible to learn alone, for example, when solving logical problems or computer programing problems (Mayer, 2004). Because of these features, many have argued that instruction is superior to learning on one's own (Debowski et al., 2001; Kirschner et al., 2006; Klahr & Nigam, 2004; Stockard et al., 2018). On the other hand, because teaching is often efficient and useful, humans also make assumptions about teaching. For example, children assume that teaching is complete and will explore a new object less after being taught by a teacher than after observing a naïve learner (Bonawitz et al., 2011). Likewise, humans will faithfully copy even a teacher's arbitrary and unhelpful actions because they assume the actions must have important cultural or social purposes (Hoehl et al., 2019). Thus, while teaching is invaluable in cases when it is complete and accurate, these same assumptions can lead learners to overlook details or imitate arbitrary actions when teaching contains omissions or is inaccurate.

However, teaching is not the only way children acquire information. Both in the presence and absence of teaching, children are capable explorers. Even preschoolers can utilize exploration to learn about cause and effect (L. E. Schulz & Bonawitz, 2007; L. E. Schulz & Gopnik, 2004; Sim & Xu, 2017), and as children get older, they are able to optimize their exploratory actions to maximize information gain (C. Cook et al., 2011; Ruggeri et al., 2019). Critically, when children are provided with explicit cues to evaluate teaching, they skillfully adapt to ineffective teaching. For instance, 4- and 5-year-old children consistently choose to learn names for new objects from an adult that was previously correct (rather than previously incorrect; Corriveau et al., 2011). Importantly, children also use exploration to compensate for pedagogical inaccuracies. For example, when an adult makes a counterintuitive claim (e.g., that the largest of a set of objects is the lightest), elementary-aged children will explore the objects for longer than if provided with an intuitive claim (Ronfard et al., 2018). Similarly, 6- and 7-year-old children will explore an object more after instruction if an adult had previously omitted information than if they had been completely informative (Gweon et al., 2014). Together, this evidence suggests that children are able to use explicit cues to identify ineffective teaching even very early in life, trading off between taught information and self-generated information to optimize learning, and that these abilities become more nuanced with age.

However, as mentioned, explicit cues are often not available to allow children to evaluate teaching. While it is easy to identify blatant inaccuracy (e.g., mislabeling everyday objects), without explicit cues, learners might not even realize teaching is ineffective until they hit a roadblock. For example, if a seemingly reliable teacher says that a particular key opens a door, you would not know the key was ineffective unless you tested it yourself. Therefore, children must be able to compensate for inaccuracy with exploration and to use exploration to *identify* inaccuracy in the first place. Moreover, this ability may be particularly important for children who are perceived as girls because discrimination makes it more likely that their perspectives and experiences will be omitted from teaching and research (e.g., Geller et al., 2018; Mansukhani et al., 2016; Prakash et al., 2018; Scott et al., 2018; Woitowich et al., 2020). For instance, much of existing research on attention-deficit/hyperactivity disorder is based on studies of boys and men (Da Silva et al., 2020; Hinshaw et al., 2022; Quinn & Madhoo, 2014). As a result, diagnostic criteria for attention-deficit/hyperactivity disorder have historically overlooked the symptoms of girls, who now must independently explore and advocate for diagnoses (Williamson & Johnston, 2015). This example illustrates that, while all students benefit from detecting inaccuracies in teaching about general topics (e.g., global warming), girls will have a greater stake than boys in identifying inaccuracies for topics that are heavily affected by gender marginalization.

Gendered People-Pleasing Socialization

Yet, gendered socialization practices may paradoxically make it more difficult, on average, for children perceived as girls to identify and divest from ineffective teaching because girls[1] are oriented toward greater people-pleasing than boys. Here, we use "people-pleasing" to refer to the tendency to deprioritize one's own preferences, experiences, or ideas for fear of repercussions (Psychology Today, 2024). Even as toddlers, girls are encouraged to prioritize others' emotions to a greater extent than boys (Zahn-Waxler et al., 1991). This trend continues throughout early childhood and adolescence, such that girls are more frequently expected to consider the needs of others and told to "be nice" than boys (Bussey & Bandura, 1999; Jordan et al., 1991). Beginning in the elementary years, girls are generally more likely to conceal their own needs and emotions to avoid disrupting social relationships (Crick & Nelson, 2002; Letendre, 2007). This emphasis on attending to others' needs and

[1] While similar gender-socialization practices exist in many cultures, we limit our evidence in the literature review to North America and Europe to match the demographics of our samples. Cross-cultural findings and constraints on generality are reviewed in the General Discussion section.

maintaining positive relationships may translate to gender differences in average agreeableness (the extent to which individuals can be perceived as kind, cooperative, and considerate) and conscientiousness (duty and diligence, including taking obligations to others seriously). While neither of these constructs explicitly or exclusively taps people-pleasing, each shares conceptual similarities with people-pleasing. To this end, work with large-scale samples has found that girls generally score higher in conscientiousness and agreeableness than boys, beginning by 3 years of age (Whalen et al., 2021) and persisting throughout adolescence into young adulthood (De Bolle et al., 2015; Slobodskaya & Kornienko, 2021; Soto, 2016; Van den Akker et al., 2021).

Moreover, gender differences can also be observed in sociotropy, defined as an excessive concern for maintaining positive relationships (Beck, 1983) and considered to be in tension with autonomy. Unlike agreeableness and conscientiousness, sociotropy measures people-pleasing more closely as questionnaires typically incorporate items evaluating excessive concerns for what others think, dependency, and pleasing others (Robins et al., 1994). While investigations of sociotropy are sparse with elementary-aged children, reliable gender differences have been identified in older samples. For instance, Yang and Girgus (2019) conducted a meta-analysis of over 90 articles, concluding that girls and women in individualistic cultures score higher on average on measures of sociotropy than boys and men. Importantly, these differences are evident at the youngest ages assessed, including adolescents (e.g., Baron & Peixoto, 1991; Calvete, 2011; Teppers et al., 2013) and children as young as 8 years of age (Brenning et al., 2011). Thus, a variety of evidence suggests that there are gender differences in the extent to which girls and boys are encouraged to prioritize others' needs and maintain positive relationships, producing differences in boys' and girls' average people-pleasing tendencies from early childhood into adulthood.

In turn, these gender differences may compound within the context of ineffective teaching both because of power dynamics with teachers and because divesting from ineffective teaching requires a willingness to stand out and go on one's own way. Regarding these power dynamics, Mickelson's (1989) sex-role socialization hypothesis proposed that girls may be disproportionately encouraged to obey authority figures that may give girls an "edge" in school, earning higher marks on average than boys (e.g., Duckworth & Seligman, 2006; Matthews et al., 2009). Indeed, feminine-typed activities that condition obedience have been found to contribute to gender differences in academic performance in a large, nationally representative sample of 5-year-olds (Orr, 2011). Accordingly, gendered expectations of conformity and obedience are particularly evident in educational contexts. For instance, teachers have rated elementary-aged boys as more nonconforming on average than girls (Gralewski & Karwowski, 2013), and teachers have generally indicated that they perceive girls to understand classroom expectations better than boys (Åhslund & Boström, 2018). Even when praising students, teachers tend to emphasize elementary-aged girls' obedience (Dweck et al., 1978; Jones & Myhill, 2004), while tending to praise nonconformity and independence in boys (Gralewski, 2019). In turn, girls may demonstrate greater reluctance on average to move beyond ineffective teaching not just for fear of hurting others' feelings but also for fear of appearing disobedient or failing to meet authority figures' expectations.

Likewise, divesting from ineffective teaching to some extent requires self-promotion, as children must believe in their own capabilities and prioritize their own ideas over teaching. To this end, work with adults suggests that women are more likely to avoid offering their opinions to avoid backlash (Brescoll, 2011), to publicly undersell their achievement (Daubman et al., 1992), and to nominate themselves for promotions less often on average than men (Moss-Racusin & Rudman, 2010). Relatedly, women are more likely than men to be discouraged from participating in fields where brilliance is heavily emphasized (Bian et al., 2018). These tendencies appear to emerge early in childhood; Bian et al. (2017) observed that by 6 years old, girls were more likely to endorse boys and men as "really, really smart" than girls and women, while more frequently endorsing girls and women as "really, really nice." These tendencies translated to girls' likelihood to nominate themselves and others for challenging games. While 6-year-old girls were just as likely to nominate their own gender to play a game for "really, really hardworking" children as boys were, they were significantly less likely than boys to nominate their own gender to play a game for children "who are really, really smart."

Taken together, a variety of evidence suggests that girls are encouraged to people-please more than boys on average: From early in life, girls are encouraged to consider others' feelings and prioritize relationships more than boys. In turn, these differences in socialization manifest in greater agreeableness and sociotropy, as well as differences in self-promotion, which can be observed beginning in the elementary years and persist into adulthood. This evidence suggests that girls may generally persist more in teaching than boys. While this strategy is adaptive in contexts in which teaching is effective, it may disproportionately disadvantage girls in the context of ineffective teaching as they may not feel allowed to explore their own ideas.

The Present Work

As such, the present work had two overarching goals. First, to add a layer of real-world nuance to psychological research on children's exploratory skills by acknowledging that children are sometimes provided with inaccurate or ineffective teaching, which is difficult to identify. While past work has established that children can reason flexibly about teaching effectiveness when provided with overt cues, the inherent ambiguity of inferring ineffectiveness using only self-generated evidence creates unique complexities that merit thorough investigation. Thus, we sought to investigate whether children could evaluate and adapt to ineffective teaching in the absence of overt cues. Second, we sought to highlight how ineffective teaching may, on average, have disproportionate harms for girls relative to boys. Thus, the second goal of this work was to investigate differences in how boys and girls broadly engage in evaluating ineffective teaching in the absence of overt cues and whether these differences relate to differences in people-pleasing.

Across three experiments, 7- to 10-year-olds were recruited to play self-contained problem-solving games comparable to tasks previously used in research on exploration. This choice allowed us to strip away some of the inherent complexity in real-world topics to first understand the basic phenomenon. This age range was recruited because elementary-aged children can use exploration to assess incomplete and counterintuitive teaching (e.g., Gweon et al., 2014; Ronfard et al., 2018) and because gender stereotypes about science and brilliance are both observable at these ages (Bian et al., 2017; Lei et al., 2019). Across experiments, children first received instructions on how to solve a game and then played a test game in

which the taught solution did not work. Importantly, children were not given any overt indications that the solution would not work or that the teacher was unreliable. Instead, children had to self-generate evidence of the solution's ineffectiveness by testing the taught solution. To understand whether observed differences generalized across contexts, a masculine-stereotyped search task was used in Experiments 1 and 2, while a feminine-stereotyped rule-learning task was used in Experiment 3. Further, to begin to understand whether differences in people-pleasing are related to children's divestment from ineffective teaching, we collected a measure of sociotropy in Experiment 3.

Children's engagement with teaching was evaluated in several ways: whether they successfully tested the taught solution, the time they spent imitating and exploring (before and after testing), children's success during the games, and how much they learned about the games. We expected that children would generally be able to test the taught solutions and that they would increase their exploration in response to evidence that the taught solutions were ineffective. Further, because girls are not less competent than boys (Duckworth & Seligman, 2006; Matthews et al., 2009; Orr, 2011) and because children are overall very sensitive to pedagogical cues (e.g., Bonawitz et al., 2011; Buchsbaum et al., 2011; Carr et al., 2015; Csibra & Gergely, 2006, 2009; Hoehl et al., 2019), we did not expect that girls' average performance and strategies would initially differ from boys'. Instead, we hypothesized that children would engage similarly with instruction before they tested the taught solution but that gender differences in exploration would emerge at the point at which children were required to confidently infer teaching was ineffective. Specifically, we hypothesized that once the taught solution failed, girls would generally hesitate to explore their own ideas because of pressure to people-please, and this would translate to differences in average solving and learning.

Experiment 1

Method

Participants

We recruited 7- to 10-year-olds ($n = 62$, $M_{AGE} = 8.80$ years, 33 girls, 28 boys, 1 non-binary participant) from a database of families from Canada and the United States that had previously volunteered to participate in research. Children's gender was measured through caregiver report. This experiment was intended to be preliminary and exploratory; thus, no statistical methods were used to predetermine sample size, and a sample of $n = 62$ was selected to ensure sufficient normality in each age group (i.e., at least $n = 30$ for each half of the sample). Caregivers identified their children as either White ($n = 23$), multiracial ($n = 15$), South Asian ($n = 7$), East Asian ($n = 5$), Southeast Asian ($n = 3$), Arab ($n = 1$), a race/ethnicity not indicated ($n = 4$), or did not report ($n = 4$). For each child, at least one caregiver held at least a bachelor's degree ($n = 25$), master's degree ($n = 24$), doctoral or professional degree ($n = 9$), or did not report ($n = 4$). Thus, our sample was somewhat ethnically diverse but not educationally diverse.

Children were individually tested in a single session hosted and recorded using Zoom video conferencing, lasting approximately 30 min. We complied with all relevant ethical regulations to conduct this work, which was approved by the Research Ethics Board at the University of Toronto, Protocol Title: "Social Influences on Independent Problem Exploration," ID: 00038355. Caregivers signed a consent form before participation, and consent to publish images of research participants was obtained. Children provided their assent prior to beginning the experiment. Caregivers received a $5 gift card for their participation. Data were excluded from 10 additional participants due to failure to independently complete a practice trial ($n = 4$), technological difficulties ($n = 3$), refusal to follow directions ($n = 2$), and experimenter error ($n = 1$).

Procedure

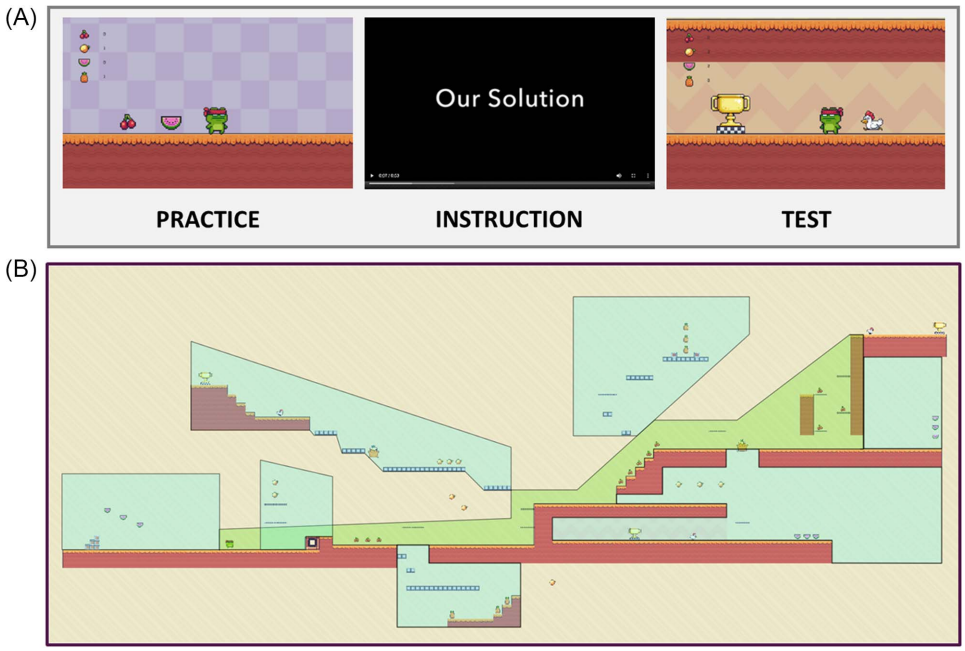
To understand children's ability to detect ineffective teaching and its influence on exploratory behavior, children played a platforming video game. In the game, participants navigated as a frog character, and the goal was to find a trophy. To encourage exploration, collectibles and enemies were spread across the map, and different textures of platforms were used. While these elements did not systematically indicate the presence of trophies, they did introduce multiple possibilities that participants could test. The game had three phases: practice, instruction, and test. Children learned about the game in the practice and instruction phases, and their problem solving and reaction to instruction were measured during the test phase.

Practice. To ensure that all children included in the final sample had sufficient knowledge of the game controls, participants first completed a short practice trial in which they interacted with all the basic elements of the game: collectibles, enemies, jumping, platform types, and trophies. Participants were given feedback from the experimenter while playing to help them learn the game controls, and caregivers were able to provide instructions but not complete the practice for their children. Data were excluded from participants that were unable to independently complete the practice within 10 min ($n = 4$). Importantly, once they succeeded in the practice game, it was emphasized to participants that the main goal was to find a trophy.

Instruction. Following practice, children were shown an instructional video featuring a seemingly knowledgeable adult navigating the test game to find a trophy. Importantly, the game map in the practice game was different in several ways (e.g., background color, map structure) from the test game, while the instructional video and test game were ostensibly the same. Prior literature indicates that children are very sensitive to pedagogical instruction (e.g., Bonawitz et al., 2011; Buchsbaum et al., 2011; Csibra & Gergely, 2006, 2009; Hoehl et al., 2019); thus, our instructional video balanced pedagogical and nonpedagogical elements. All experimenter directions emphasized that children would be shown a solution, but the instructions were not deterministic. The adult providing instructions was not visible in the video but rather demonstrated themselves playing the frog game over video, providing narration only at the beginning and end of the video.

Before demonstrating, the teacher said: "Now it's time for the real game! But before you play, we'll show you how we find the trophy to win." In the demonstration that followed, the teacher moved the frog character on a linear path from the left of the game map to the right. Eventually, the teacher navigated to a set of platforms and demonstrated that the platforms could be used to jump and find a trophy. Critically, the platforms that the teacher demonstrated using were missing in the test game, rendering the teaching ineffective. Thus, children were not told that the solution in the instructions was ineffective but had to test the taught solution themselves to generate evidence of its ineffectiveness. The teacher concluded instruction by saying, "Great! Now it's your turn. You must play the game until

Figure 1
Experimental Procedure for Experiments 1 and 2



Note. (A) Participants began with a practice, viewed an instructional video, and then were given 7 min to solve a test game. (B) For the purposes of behavioral coding, the game map was divided into exploratory regions, depicted in blue (light), and imitative regions, depicted in green (dark). See the online article for the color version of this figure.

you find a trophy or until 7 minutes have passed. You can win any way that you can find."

Test Game. After instruction, participants completed the test game, which appeared identical to the instructional video except for the missing platforms. While the solution shown in the video did not work, the participants could solve the game by finding either of two alternate trophies in areas branching from the main path. Before participants played the game, they were told by the experimenter,

> For this part of the game, we want to see what you do on your own, so I'm going to do my best not to answer any questions, and we ask that parents do the same. I'll set a timer for 7 minutes and let you know when 7 minutes pass. Remember, the main goal of the game is to find a trophy, but you can win with any trophy that you find!

Participants were given up to 7 min to solve the game, which began once the experimenter finished giving directions. During the test game, the experimenter's video was visible, but they did not provide feedback and maintained a neutral affect.

Behavioral Coding

Coding was performed by human raters to classify testing behavior, time spent exploring and imitating, and whether participants solved the game. A primary coder was randomly assigned to each participant. Data generated by primary coders were used in the analyses. In addition, a secondary coder was randomly assigned such that 50% of all data were double-coded, allowing for the estimation of interrater reliability. Agreement between primary and secondary coders was high across both dichotomous measures (i.e., solving and testing; ≥94% agreement) and continuous measures (i.e., imitation and exploration; all intraclass correlation coefficients [ICCs] ≥ .94, all *p*s < .001).

Testing the Taught Solution. Participants were considered to have successfully tested the taught solution if they followed the path in the instructional video and completed one full jump where the platforms were missing. At this point, participants should have evidence that the solution will not work as shown.

Coding Imitation and Exploration. Classifications were made using the physical location of the frog on the game map (see Figure 1). If the frog was in the portion of the game map that was used in the path of the instructional video, participants' behavior was coded as imitation, starting from the frame they entered the area until the frame before they exited. If the frog was in areas branching off from this path, exploration was instead coded from the frame they entered the area until the last frame before they exited. Thus, these definitions generate two mutually exclusive codes. As the absolute time children played the game varied, proportions were used in analyses. Importantly, the proportion of time engaging in exploration and imitation (respectively) was calculated both before and after participants tested the taught solution so that behavior could be compared once participants gained evidence of the solution's ineffectiveness.

Solving. Participants were considered to have solved the game if they found a trophy within 7 min.

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and we follow

Journal Article Reporting Standards (Kazak, 2018). The analyses presented here were not preregistered. All data, analysis code, and files for the frog game are available at the Open Science Framework at https://osf.io/qn3w5/. Data were analyzed using RStudio, Version 2023.12.0 (R Core Team, 2023), using the functions binom.test, t.test, lm, glm, and cor.test. All reported tests were two-tailed, and data met the assumptions of the statistical tests used. Points detected at $SD = 3$ above/below the mean were considered outliers and would have been removed from analyses using pairwise deletion (see Supplemental Table S1 for mean and standard deviation for each variable). However, no outliers were detected following this definition. Likewise, because we did not have statistical power to evaluate nonbinary participants, they were excluded only from analyses of gender.

Results

Overall Analyses

To determine how children reacted to ineffective teaching, we first sought to understand whether children were engaged with the instruction. Participants were not told that the instruction was ineffective and instead had to utilize their own exploratory actions to discover its ineffectiveness. Thus, given that instruction is often efficient and that children readily adopt the solutions of teachers (e.g., Bonawitz et al., 2011; Buchsbaum et al., 2011; Carr et al., 2015; Csibra & Gergely, 2006, 2009; Hoehl et al., 2019), we expected that all children would first exploit instruction to try to solve the game quickly, then increase exploration after evidence of failure. Therefore, we hypothesized that children would initially favor imitation, increasing exploration after testing the taught solution.

Thus, we first performed a one-sample t test to compare the proportion of time children spent imitating before testing the taught solution to chance (i.e., 0.5). Indeed, children initially imitated significantly more than they explored, $M = 0.75$, $t(61) = 8.92$, $p < .001$, $d = 1.13$. Likewise, we hypothesized that children would generally succeed in testing the taught solution given children's broad exploratory skills. We performed a binomial test comparing the proportion of children who tested to 0.5 to see if the majority of children tested the taught solution; 74% of children tested the taught solution, a rate greater than chance ($p < .001$). Finally, a paired-sample t test was performed to compare children's rates of exploration before and after testing the taught solution. As a group, children explored more and imitated less after testing than before testing, $M_{Diff} = 0.14$, $t(45) = 4.14$, $p < .001$, $d = 0.61$. Thus, children generally generated evidence that the taught solution was ineffective using only their own exploration, and on average, adjusted their strategies based on this information.

We further sought to understand how exploratory behavior after testing was related to success (i.e., solving). We hypothesized that children who spent a greater proportion of time exploring after testing the taught solution would be more likely to solve the game. A two-sample t test revealed that children who solved the game spent a significantly greater proportion of time exploring after testing ($M = 0.44$) than children who did not solve, $M = 0.25$; $t(44) = 3.77$, $p < .001$, $d = 1.12$. Taken together, these results suggested an overall calibrated response from children: first, attempting to make use of the instructions until they self-generated evidence that the taught solution was ineffective and then increasing rates of exploration, with those who succeeded adjusting at the greatest rates.

Analyses of Gender Differences

Having established general trends, we sought to explore gender-based trends in our data. Recall that our specific hypothesis is that girls would be disadvantaged when they needed to divest from teaching, rather than that girls would be generally less exploratory or successful at the game than boys. Thus, we hypothesized that girls would not differ substantially from boys initially and may even test the solution at higher rates given that girls tend to outperform boys academically when teaching is generally correct (Duckworth & Seligman, 2006; Matthews et al., 2009; Orr, 2011). However, our sample size was limited due to the exploratory nature of the experiment, so we did not have sufficient power to investigate our specific hypothesis that girls would explore less than boys *after* observing that the solution was ineffective, as a limited portion of children tested the taught solution ($n = 46$; achieved power = .26). Thus, we instead focused our analyses on the total time spent exploring and overall solving rates. However, it is worth noting that, on average, girls explored less ($M = .30$) than boys after testing the taught solution ($M = 0.37$). Girls also tended to explore less on average before testing the taught solution ($M = 0.14$) than boys ($M = 0.26$; see Supplemental Table S1 for all descriptive statistics).

Consistent with our predictions, we did not find a significant difference in rates of testing the taught solution between girls and boys ($p = .57$). As we are cautious to overinterpret a null result, we additionally qualified this analysis with post hoc binomial tests to see whether the majority of girls and the majority of boys, respectively, were able to test the taught solution and generate evidence of its ineffectiveness. Indeed, the majority of children in both groups were able to effectively engage with the instruction (both $ps \leq .01$). However, differences emerged when looking at rates of exploration and solving the game. To this end, we used a linear regression to predict the overall proportion of time spent exploring and a logistic regression predicting solving using the main effects of age and gender as predictors. These analyses revealed that, on average, girls explored for a smaller overall proportion of time and imitated for a greater proportion of time than boys, $\beta = -0.10$, $SE = .04$, $t(58) = -2.36$, $p = .02$, and they tended to solve the game at lower rates, $\beta = -1.35$, $SE = 0.59$, $z(58) = -2.28$, $p = .02$, $OR = 0.26{:}1$.

Discussion

Taken together, our first experiment suggests that children as a group were able to engage with ineffective instruction in a nuanced way: first attempting to use instruction to solve the game efficiently and then shifting their strategies to increased exploration. While there was significant variability in girls' and boys' performance (i.e., while many girls solved the game, and many boys did not), we also found initial evidence that girls tended to imitate more and explore less than boys on average. Likewise, we observed that girls were less likely to solve the game. However, given the limited power and exploratory nature of this initial experiment, we were unable to make conclusive statements as to whether this decreased exploration *caused* solving differences. As such, we collected an adequately powered sample for mediational analyses in Experiment 2.

Experiment 2

Method

Participants

We recruited 7- to 10-year-olds ($n = 150$, $M_{AGE} = 9.02$ years, 76 girls, 74 boys) from a database of families from Canada and the United States that had previously volunteered to participate in research. Children's gender was measured through caregiver report. Sample size was determined utilizing a priori power analyses to achieve at least 80% power for all hypothesized effects. Caregivers identified their children as White ($n = 57$); multiracial ($n = 33$); East Asian ($n = 24$); South Asian ($n = 10$); Arab ($n = 4$); Southeast Asian ($n = 3$); Latin, Central, or South American ($n = 2$); Native American ($n = 1$); a race not listed ($n = 3$); or did not report ($n = 13$). For each child, at least one caregiver held at least a high school diploma ($n = 8$), college degree ($n = 10$), bachelor's degree ($n = 34$), master's degree ($n = 52$), doctoral or professional degree ($n = 33$), or did not report ($n = 13$). Thus, our sample was somewhat ethnically diverse but not educationally diverse.

Children were individually tested in a single session using Zoom videoconferencing, lasting roughly 45 min. We complied with all relevant ethical regulations to conduct this work, which was approved by the Research Ethics Board at the University of Toronto, Protocol Title: "Social Influences on Independent Problem Exploration," ID: 00038355. Caregivers signed a consent form before participation, and consent to publish images of research participants was obtained. Children provided their assent prior to beginning the experiment. Caregivers received a $5 gift card for their participation. Data were excluded from 14 additional children due to caregiver or sibling interference ($n = 6$), failure to independently complete the practice ($n = 4$), technological issues ($n = 2$), refusal to comply ($n = 1$), and experimenter error ($n = 1$).

Procedure

The procedure for Experiment 2 was identical to Experiment 1 other than the exceptions outlined here. First, because the teacher was always a woman in Experiment 1, we matched the gender of the teacher to the participant in Experiment 2 to control for potential matching effects. Second, we added a set of posttest questions. Participants first completed a learning assessment, allowing us to evaluate the extent to which differences in immediate success may be consequential for downstream outcomes. Specifically, short-term differences in exploration may create later disadvantages if they affect the extent to which children understand the game map more broadly, as this could affect not just their ability to find trophies but also their ability to succeed in future objectives (e.g., collecting the most pineapples, finding the tallest point). Thus, participants were shown a series of four images portraying various game map areas (some that were actually in the game and some that were not) and asked whether they were part of the game. This resulted in a final score out of 4. To ensure that this measure reflected learning and not simply memory or attention to instruction, the majority of game areas assessed were not featured in the teacher's demonstration (except for one item, which was included to sustain children's self-efficacy). After the learning assessment, children were asked to rate how often they played video games on a 5-point Likert scale ranging from 1 (*never*) to 5 (*daily*), generating a control measure of video game skill/interest for analyses.

In addition to these key measures, we assessed children's confidence by asking children who tested the taught solution if they believed they had done something wrong or if the game was broken when the solution failed. Finally, as previous work has documented gender differences in self-regulation (e.g., Duckworth & Seligman, 2006; Matthews et al., 2009; Montroy et al., 2016; Størksen et al., 2015), we sought to explore the possibility that gender differences in exploration may be accounted for by differences in executive function. Thus, children completed three executive function tasks before concluding the experiment. Analyses and further details on the procedure for these measures can be found in Supplemental Materials, as they are not directly relevant to the findings presented here.

Behavioral Coding

The same procedures and variables were coded as in Experiment 1, with high interrater reliability on dichotomous measures (i.e., solving & testing; 100% agreement) and continuous measures (i.e., exploration and imitation; all ICC's $\geq .95$, all p's $< .001$). In addition, coders performed coding to quantify persistence in attempting the taught solution by counting the number of jumps participants performed. As before, a primary coder was assigned to each participant, and 50% of data were double-coded for assessment of reliability. Agreement was extremely high (ICC = .99, $p < .001$).

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and we follow Journal Article Reporting Standards (Kazak, 2018). The analyses presented here were not preregistered. All data, analysis code, and files for the frog game are available at the Open Science Framework at https://osf.io/qn3w5/. Data were analyzed using RStudio, Version 2023.12.0 (R Core Team, 2023), using the functions binom.test, t.test, lm, glm, and cor.test. All reported tests were two-tailed, and data met the assumptions of the statistical tests used. Mediational hypotheses were tested with a bootstrap procedure to determine the significance of the indirect effect (Preacher & Hayes, 2004); 5,000 bootstrap resamples and a random seed of 65,336 were used to estimate the direct, indirect, and total effects using the PROCESS v4.0 macro (Hayes, 2022); 95% confidence intervals were determined from the bootstrap resamples, and any interval that did not include 0 was considered significantly different from 0. Outliers detected at $SD = 3$ above/below the mean were be removed from the analyses using pairwise deletion (see Supplemental Table S2 for mean and standard deviation for each variable). Under this definition, outliers were only detected for the persistence measure ($n = 2$). Thus, outliers were only excluded from analyses of persistence, rather than from all analyses.

Results

Overall Analyses

Once again, we first sought to understand how children engaged with instruction when it was unclear that it was ineffective. Thus, we performed a one-sample t test comparing the proportion of time spent imitating to chance. This analysis indicated that children initially

prioritized imitation over exploration, $M = 0.72$, $t(149) = 12.00$, $p < .001$, $d = 0.98$. Using a binomial test to compare the proportion of children that tested to chance (0.5), we also found that the majority of children tested the taught solution (74% of children, $p < .001$). Thus, we replicated the finding that children as a group initially prioritized imitation and engaged with instruction rather than exploring their own ideas. Likewise, as a group, children explored more and imitated less after testing than before, $t(110) = 6.03$, $p < .001$, $d = 0.57$, and children who solved the game explored more and imitated less than children who did not solve the game, $t(109) = 8.87$, $p < .001$, $d = 1.76$. As with solving, a Pearson correlation suggested that the proportion of time that children spent exploring after testing the taught solution was correlated with learning ($r = .32$, $p < .001$). Thus, patterns found regarding differences in solving were also reflected in differences in children's learning. Overall, children who explored their own solutions after generating evidence of the teachings' ineffectiveness were advantaged in several ways relative to their peers.

Analyses of Gender Differences

Having elaborated general trends, our next goal was to understand gender differences in children's performance (see Supplemental Table S2 for all descriptive statistics). Recall that our prediction is not that there are general differences in learning and success by gender but rather that girls will feel less able to divest from teaching and prioritize their own ideas once they test the taught solution. To this end, we first began by constructing linear regressions predicting exploration, using the main effects of age and gender as predictors, as well as a logistic regression with the same specifications predicting whether children tested the taught solution. While we did not find a significant difference between girls' ($M = 3.86$) and boys' ($M = 4.20$) average video game experience, we did find a nonsignificant trend suggesting that boys played video games more often, $t(148) = 1.65$, $p = .10$. Thus, we chose to control for video game experience in all models, though it is worth noting that inclusion of the variable did not affect patterns of significance (see Supplemental Table S4).

Consistent with our prediction, there were no gender differences observed in children's ability to test the taught solution ($p = .25$). As we are cautious to overinterpret a null result, we additionally qualified this analysis with post hoc binomial tests to see whether the majority of girls and the majority of boys, respectively, were able to test the solution and generate evidence of its ineffectiveness. Indeed, the majority of children in both groups were able to effectively engage with the instruction (both ps $\leq .002$). On the other hand, differences emerged between girls and boys when divesting from teaching became necessary. We found that, overall, girls explored less and imitated more on average than boys, $\beta = -0.13$, $SE = .03$, $t(146) = -4.47$, $p < .001$. These differences in exploration and imitation were driven by behavior after testing the taught solution, $\beta = -0.16$, $SE = .04$, $t(107) = -4.24$, $p < .001$ (see Figure 2), as average differences between girls' and boys' exploration and imitation before testing the taught solution were not significant ($p = .09$; see Supplemental Table S5 for further analyses of timing). As a result, we proceed with analyses of the proportion of time children spent exploring after testing the taught solution.

Thus, as expected, differences emerged between girls and boys when divesting from teaching became necessary. While on average, boys explored more and imitated less than girls after testing the taught solution, girls tended to be more persistent,[2] attempting the taught solution significantly more times on average than boys, $\beta = 0.68$, $SE = 0.33$, $t(105) = 2.07$, $p = .04$. Ultimately this persistence came at a cost in the context of ineffective teaching, as girls were less likely to solve the game, $\beta = -1.08$, $SE = 0.44$, $z(107) = -2.45$, $p = .01$, $OR = 0.34{:}1$, and girls answered significantly fewer learning questions correctly on average than boys, $\beta = -0.36$, $SE = 0.11$, $t(107) = -3.34$, $p = .001$ (see Table 1), even after controlling for video game experience. Altogether, these results suggest that girls were not merely worse at the game, as they did not differ from boys in their ability to test the taught solution and persisted in the taught solution in the face of difficulty. Likewise, results were not driven by differences in video game experience. Instead, disadvantages began for girls when they needed to divest from the ineffective instruction, as we observed differences in their subsequent behaviors: exploration/imitation, solving, and learning.

Having demonstrated gender differences in several key variables, we finally sought to elaborate on the pathways creating these differences. Specifically, we hypothesized that girls exhibited lower levels of solving and learning because they explored less. Thus, we hypothesized that girls' lower learning and solving rates were mediated by the proportion of time spent exploring after testing (see Figure 3). These analyses revealed that gender affected solving as a function of its relationship with exploration ($ab = -1.46$, $SE = 0.52$, 95% confidence interval, CI, [−2.76, −0.75]). When the indirect path through exploration was taken into account, the direct effect of gender on solving was no longer significant ($p = .71$). Thus, solving rates were parsimoniously explained through the indirect pathway, and girls' decreased rates of solving could specifically be attributed to differences in exploration. Likewise, gender affected learning as a function of its relationship with exploration ($ab = -0.10$, $SE = 0.05$, 95% CI [−0.20, −0.01]). Unlike solving, there was still a direct effect of gender on learning when the indirect path through exploration was taken into account ($c' = -0.25$, $SE = 0.12$, 95% CI [−0.48, −0.02]), suggesting learning was only partially explained by the co-occurrence of gender and lower exploration after testing the taught solution. Nevertheless, the relationship between gender and learning was mediated by exploration.

Discussion

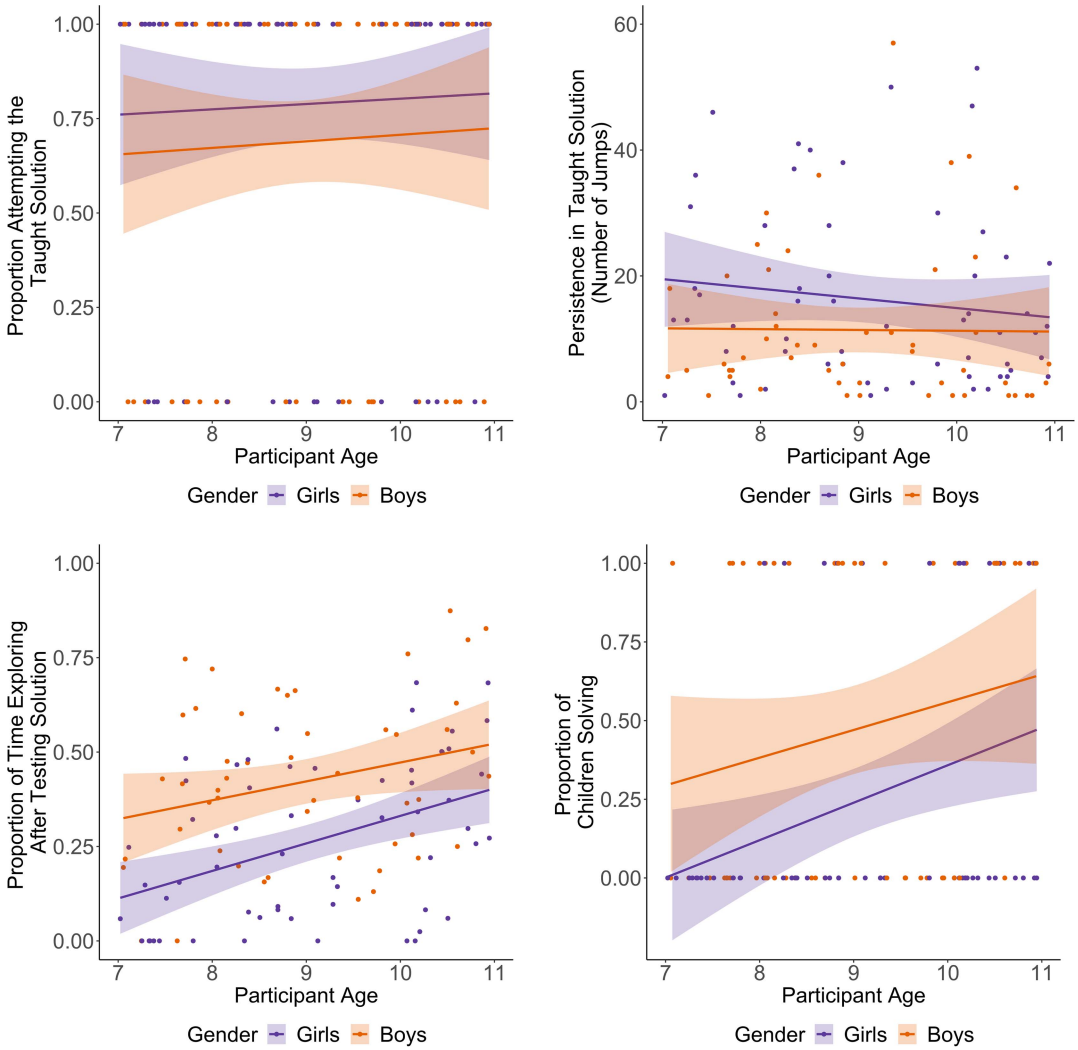
Taken together, these findings suggested that children were capable of independently verifying teaching and adjusting to self-generated evidence, with those who ultimately succeeded and learned the most making the largest adjustments. However, these results also corroborated the gender differences we observed in Experiment 1. Although girls tested the taught solution at similar rates as boys, they spent less time on average exploring their own solutions than boys. In turn, these differences in exploration caused gender differences in average success and learning. However, it is important to note that rather than giving up, girls instead tended to persist in the taught solution, making more attempts on average to refine their jumping in the area where the taught solution failed.

Simultaneously, several open questions remained that we sought to address in Experiment 3. First, while models in Experiment 2 controlled for video game experience, the game could have been perceived as stereotypically "masculine." For this reason, girls may

[2] Because we found skewness = 1.27 for our measure of persistence, a square-root transformation was applied.

Figure 2

Key Variables in Experiment 2 Plotted Against Participant Age and Gender



Note. Differences in testing behavior were not observed as a function of gender. However, differences in solving were observed as a function of participant gender, as well as in persistence and exploration after testing the taught solution. Individual observations are plotted, along with lines of best fit estimated using linear regressions. As can be seen, while girls' and boys' performance differed on average, there was also substantial overlap in performance. For example, while girls explored less on average than boys, several girls explored at high rates, and several boys explored at low rates. Shaded regions represent 95% confidence intervals. See the online article for the color version of this figure.

have been sensitive to expectations that they would underperform on the task relative to boys (e.g., Kaye & Pennington, 2016). Thus, the extent to which the observed gender differences generalize to different settings remains unclear, especially those which are considered stereotypically "feminine." Indeed, prior work has documented that gender stereotypes can be stronger in masculine-stereotyped domains (e.g., science and technology; Bian et al., 2018). As such, Experiment 3 was designed to understand whether gender differences in divestment from ineffective teaching replicate across contexts by introducing a problem-solving task that appeared to be feminine-stereotyped. To ensure our manipulations of gender stereotyping were effective, children were directly asked to rate the gender typing of the tasks from both Experiments 2 and 3.

In addition to assessing domain specificity, the problem-solving game for Experiment 3 investigated the role of problem structure in shaping gender differences. In Experiments 1 and 2, we employed a search task in which children needed to locate rewards in the game environment (i.e., trophies) to succeed. This choice of task allowed us to efficiently classify children's behavior into mutually exclusive categories, as exploration and imitation could be classified from children's position in the game. Search tasks have also previously been effective in evoking and classifying differences in children's exploratory strategies (e.g., E. Schulz et al., 2019). However, although search tasks have several methodological advantages in the study of exploration, they also do not represent the full range of problems that children need to learn to solve.

Table 1
Participant Gender Predicts Key Measures in Experiment 2

Predictor	Proportion of time exploring (after testing)		Square root (persistence)	
	β (*SE*)	95% CI	β (*SE*)	95% CI
Intercept	−0.20 (0.15)	[−0.50, 0.09]	4.79 (1.31)***	[2.23, 7.36]
Age	0.06 (0.02)***	[0.03, 0.09]	−0.13 (0.14)	[−0.41, 0.15]
Video game experience	0.02 (0.01)	[−0.005, 0.05]	−0.17 (0.13)	[−0.42, 0.09]
Gender (girl)	−0.16 (0.04)***	[−0.23, −0.09]	0.68 (0.33)*	[0.04, 1.33]

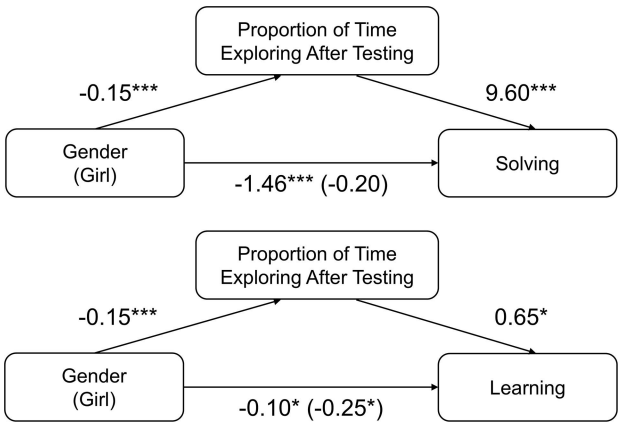
Predictor	Solving		Learning	
	β (*SE*)	*OR* (95% CI)	β (*SE*)	95% CI
Intercept	−6.01 (1.91)**	0.002 [<.001, 0.10]	1.33 (0.43)**	[0.49, 2.17]
Age	0.47 (0.19)*	1.61 [1.10, 2.34]	0.13 (0.05)**	[0.04, 0.22]
Video game experience	0.39 (0.20)†	1.48 [0.99, 2.21]	0.03 (0.04)	[−0.06, 0.11]
Gender (girl)	−1.08 (0.44)*	0.34 [0.14, 0.81]	−0.36 (0.11)**	[−0.57, −0.15]

Note. *SE* = standard error; CI = confidence interval; *OR* = odds ratio.
† $p < .08$. * $p < .05$. ** $p < .01$. *** $p < .001$.

For example, as children could only find an alternate trophy through sustained exploration of the game map, their success was closely connected to their exploration. However, in real-world contexts, children often learn new information from sparse evidence, especially if the evidence they generate is highly salient. As such, it is also important to consider whether gender differences are observed when learning is less dependent on exploration. To this end, rule learning presented a rich ground for investigation in Experiment 3, as children must infer many rules in daily life. For instance, children can learn that heavier objects gain more momentum than lighter objects (a rule) from a small set of observations.

In addition to being less reliant on exploration, employing a rule-learning task allowed us to assess learning through generalization to new items. This presented two methodological advantages. First, generalization to new items is generally regarded as a hallmark of deep learning as it requires children to think conceptually (Fiorella & Mayer, 2016). Second, because learning was assessed in Experiment 2 via children's knowledge of the game map, the measure only reflected children's knowledge of the game they themselves played. While this learning is consequential for future experiences in the same environment, it may not have far-reaching consequences in new environments. On the other hand, the rules that children learn influence their interactions with novel items and form the foundation for future learning (e.g., understanding how momentum relates to impulse). Thus, differences in generalization provide more compelling evidence that differences in reactions to ineffective teaching create consequential disadvantages beyond the immediate context.

Figure 3
Gender Differences in Success and Learning Are Mediated by Differences in Exploration



Note. Standardized path estimates are presented. The parenthetical value is for the direct effect once the indirect pathway is accounted for.
* $p < .05$. *** $p < .001$.

Moreover, while the video game in Experiments 1 and 2 allowed us to cleanly classify children's behavior through mutually exclusive categories of imitation and exploration, this choice ultimately leaves ambiguity regarding the precise nature of children's strategies. That is, there are three possibilities regarding the observed gender differences in strategy: (1) it may be the case that spending less time imitating is particularly influential for learning and success and that boys tend to imitate less on average than girls, (2) it may be the case that spending more time exploring is particularly influential for learning and success and that boys tend to explore more on average than girls, or (3) exploration and imitation each contribute to variability in success. As such, we sought to develop a task in which children's exploration and imitation could vary independently so that we could interrogate the relative contributions of each behavior.

Finally, while we propose that people-pleasing socialization creates greater pressure for girls to obey instruction and avoid upsetting teachers, our previous experiments did not directly assess this possibility. Thus, in Experiment 3, we sought to incorporate a measure of sociotropy, as validated empirical measures of people-pleasing are not available. Sociotropy does not measure all components of our definition of people-pleasing, focusing on concern for maintaining positive relationships but not necessarily appearing obedient or avoiding self-promotion. However, there is robust evidence of gender differences in sociotropy among adults (Yang & Girgus, 2019), and differences have been found in children between 8 and 14 years old (Brenning et al., 2011). Thus, by

incorporating a measure of sociotropy, we could assess whether girls reported greater average levels of sociotropy than boys, as well as whether sociotropy related to children's strategies and success in the context of ineffective teaching.

Experiment 3

Method

Participants

We recruited 7- to 10-year-old Canadians ($n = 100$, $M_{AGE} = 8.94$ years, 46 girls, 54 boys). Children's gender was measured through caregiver report. Children were recruited from a database of families that had previously volunteered to participate in research and tested in the laboratory ($n = 40$). In addition, participants were recruited from and tested at local museums ($n = 60$). Sample size was determined utilizing a priori power analyses to achieve at least 80% power for main effects in models involving children who tested the taught solution ($n = 93$ testers). Caregivers identified their children as White ($n = 35$); East Asian ($n = 20$); multiracial ($n = 16$); Southeast Asian ($n = 6$); South Asian ($n = 4$); Black ($n = 2$); Arab ($n = 1$); Latin, Central, or South American ($n = 1$); a race/ethnicity not indicated ($n = 3$); or did not report ($n = 12$). For each child, at least one caregiver held at least a high school diploma ($n = 2$), college degree ($n = 13$), bachelor's degree ($n = 37$), master's degree ($n = 17$), doctoral or professional degree ($n = 20$), or did not report ($n = 11$). In addition, caregivers reported their annual household income as below $100,000 ($n = 22$), above $100,000 ($n = 61$), or did not report ($n = 17$).[3] Thus, our sample was somewhat ethnically diverse but not economically diverse.

Children were individually tested in a single session lasting roughly 30 min. We complied with all relevant ethical regulations to conduct this work, which was approved by the Research Ethics Board at the University of Toronto, Protocol Title: "Social Influences on Independent Problem Exploration," ID: 00038355. Caregivers signed a consent form before participation, and consent to publish images of research participants was obtained. Children provided their assent before beginning the experiment. For children who participated in the lab, caregivers received a $5 gift card for their participation. Children who participated in museums instead selected a small toy to take home. Data were excluded from eight additional children due to technical issues with the locks/keys ($n = 3$), developmental disability ($n = 2$), lack of video recording ($n = 2$), and sibling interference ($n = 1$).

Materials

The game was shaped like a dollhouse (9.5 cm × 29 cm × 38.5 cm; see Figure 4) and contained five locked boxes (11 cm × 11.9 cm × 10.6 cm). The boxes were made to be pulled out of the frame, allowing them to be manipulated more easily by children. The exterior of the dollhouse was painted in lilac and pink, with small accents such as flowers and butterflies. In addition, two Barbie dolls were placed in the house to emphasize that children were "opening the doors for the dolls." Each box was painted a solid color and featured a different number of symbols painted in gold. For example, the red box was red on all sides and had a moon on one face labeled with "1," while the pink box was pink on all sides and had clouds on two sides (one cloud labeled "1," and the other labeled "2"). There were 13 metal keys with colorful fobs. Each fob featured a solid color, plus either a shape *or* number (e.g., red "1" key, gray cloud key). The boxes opened using a number rule, such that each box was opened by the key whose fob matched the highest number on the box (e.g., the gray "2" key for the pink box). This number also matched the number of symbols on the box (e.g., 2 clouds on the pink box).

Procedure

The procedure for Experiment 3 had four phases, closely following the previous procedure: (1) practice, (2) instruction, (3) test, and (4) poststudy questions.

Practice. To ensure that all children included in the final sample had sufficient knowledge of the keys, participants began by trying them on a practice box. The practice box was unpainted, and the practice keys did not have fobs, distinguishing them from the materials for the test game. The experimenter first demonstrated the key that worked, "To use the keys, you have to put them all the way in the lock and then turn to the left, and then if the key is the right one, it will open easily." The box was closed, and children were asked to use the key to open it themselves. Next, the experimenter demonstrated the key that did not work,

> If you have the wrong key sometimes it won't fit in the lock at all, but other times it will go all the way inside and it just won't turn. So, if you think you have the right key, double check that it's all the way in, and if it's still not going, don't force it.

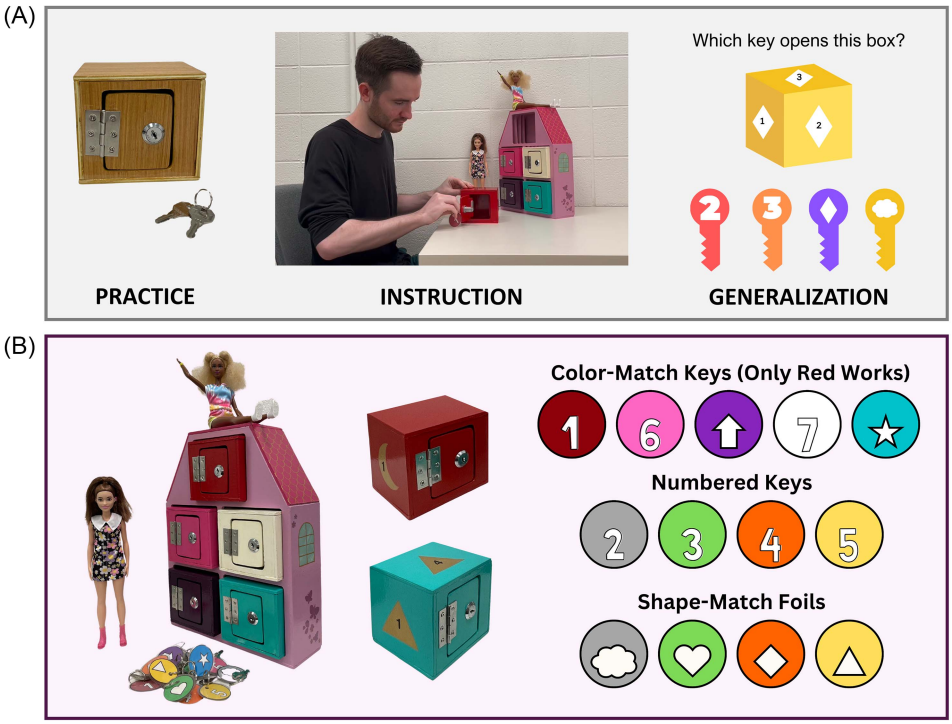
Children were then asked to try the key themselves, allowing them to see how it felt for a key to fail and practice checking that the key was fully inside the lock. Finally, the children were asked to use the working key to independently open the box one last time. Thus, all participants opened the practice box twice.

Instruction. Next, children were shown an instructional video in which a seemingly knowledgeable teacher showed them an incorrect rule to open the boxes. To build on previous experiments, we randomly assigned children to receive instruction from either a woman or a man. In this way, we could assess whether the gender of the teacher affected children's tendency to persist in the taught solution, as well as whether children's own gender interacted with the teacher's gender (see Supplemental Table S7 and Figure S2). As before, the teacher did not provide any cues that the teaching was ineffective. Thus, children had to test the solution themselves to discover evidence of its ineffectiveness. The teacher began by communicating pedagogical intent, "Now you're almost ready to start playing with the real dollhouse, but first, I'm going to show you the right way to unlock the doors for the dolls," then took out the red box from the frame. They explained that a color rule was used to open the boxes, using a key with a red fob to successfully open the red box: "To open the doors, you have to use a key that matches the color of the box. So, to open this red box, I'm going to use this red key. Great, now you can open all the doors!"

Unlike Experiments 1 and 2, the teacher taught about the same dollhouse as was used for the test. Rather than modifying the test boxes to render teaching entirely ineffective, the teacher instead demonstrated with a confounded key: for only the red box, the

[3] Here, we report annual household income relative to the median household income in Toronto, Ontario (i.e., approximately $100,000). For complete information on household income, see Supplemental Table S6.

Figure 4
Experiment Procedure for Experiment 3



Note. (A) Participants began with a practice, viewed an instructional video in which a teacher taught an ineffective color rule, and then were given 5 min to open all five boxes and discover the number rule. Learning was assessed through generalization to novel boxes. (B) The problem-solving game was designed to be feminine-stereotyped. Boxes were stored in a dollhouse adorned with decorations (e.g., butterflies) and dolls. Each box was a different solid color and featured shapes and numbers on its sides. Participants were given 13 accompanying keys whose fobs varied in color and symbol. See the online article for the color version of this figure.

correct key had a fob that matched both the color and the highest number on the box. For the other boxes, keys matched only one feature (i.e., if a key fob matched in color, it did not match the highest number). As such, the teacher correctly taught how to open the red box but taught an ineffective rule.

Test Game. After instruction, the experimenter pulled out the five boxes from the dollhouse frame and arranged them in a row in front of the participants. Children were told:

> So, for this part, we want to see what you do all on your own, so I can't answer any questions or help, but I will get you set up. The goal is to open all five boxes in 5 min. To make the boxes easier to pick up, turn around, look at. (demonstrates rotating box), I'll take them out of the frame. I'll line them up, but you don't have to keep them in this order when you start.

Because the true unlocking rule was based on number, the order of the boxes could influence test difficulty. As such, boxes were always presented in a quasisequential order based on the key used to unlock them (i.e., 1, 2, 4, 3, 5). Once the boxes were arranged, the experimenter poured the test keys out of a small box into one pile directly in front of the participant. To ensure that keys were presented randomly to participants, the box storing the keys was shaken between sessions. As they provided the children with the keys, the experimenter concluded: "And these are all the keys (pours keys on table). Remember, the goal is to open all five boxes in 5 minutes. I'll set a timer for 5 minutes, and you can go ahead and start now." During the test game, the experimenter sat beside the children, so children could not see them watching unless they turned to the experimenter. The experimenter did not provide feedback and maintained a neutral affect. Test terminated after 5 min, or after participants opened all the boxes, whichever occurred first.

Learning Assessment. After the test, the experimenter did not provide children any information about the unlocking rule or their performance. Learning was assessed through four generalization questions in which children were shown images of novel boxes. Each box featured a solid color, a shape, and numbers. For each box, children were shown four keys which only matched one feature on the box: (1) a color-matched key, (2) a shape-matched key, (3) a number-matched key, and (4) a number foil that did *not* match the box. Including both a number-matched key and a number foil allowed us to assess whether children truly understood the number rule, or if they simply inferred that number was somehow relevant to the task. For each box, children were told: "This is a new box that you've never seen before. If you had to guess, which key opens this box?" From these questions, we were able to assess children's learning (i.e., correct choices out of four), as well as their adherence to the color rule.

Poststudy Questions. Next, the children were asked a final set of questions. Most importantly, children reported how often they played

with dolls or dollhouses on a 5-point Likert scale ranging from 1 (*never*) to 5 (*daily*), and they were shown pictures of both the video game from Experiments 1 and 2 and the dollhouse from Experiment 3 and asked to rate them on a 5-point Likert scale ranging from 1 (*a lot more for boys*), 3 (*for boys and girls equally*), to 5 (*a lot more for girls*). For additional questions and analyses, see Supplemental Materials.

Sociotropy Questionnaire. Finally, we sought to collect an explicit measure to evaluate the extent to which observed differences between boys and girls could be explained by differences in people-pleasing socialization. To this end, children were shown a modified version of the Personal Style Inventory (Robins et al., 1994). To reduce the length of the assessment, children were only shown the 24 items relating to sociotropy (see Supplemental Table S8), and the 24 items relating to autonomy were omitted. The language of each item was edited to be more child-friendly, and children completed the questionnaire with the assistance of an experimenter. For 7- and 8-year-old participants, the experimenter read all questions aloud, while 9- and 10-year-olds could choose to read the questions on their own. The experimenter told participants that they could offer definitions and examples for any words children did not understand and followed a standardized script when providing clarification. Participants provided their responses using a 6-point Likert scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*). Higher scores indicated higher levels of sociotropy, and the questionnaire contained three subscales: Concern About What Others Think, Dependency, and Pleasing Others. Because this questionnaire has almost exclusively been used with adolescents and adults, children were allowed to skip questions if they did not understand them or if they had never encountered the circumstance assessed (e.g., a friend never canceled their plans). As a result, it was common for children to skip at least one question ($n = 45$, $M = 1.15$ questions). Thus, children's responses were averaged across questions to produce final sociotropy scores for analyses.

Behavioral Coding

Coding was performed by human raters to classify copying, testing, time spent exploring and imitating, participants' success, and unlocking skill. A primary coder was randomly assigned to each participant. Data generated by primary coders were used in the analyses. In addition, a secondary coder was randomly assigned such that 100% of data were double-coded for copying, testing, imitation, solving, unlocking time, and number unlocked. For our measure of exploration, persistence, and two measures of unlocking skill, 35% of data were double-coded. Agreement between primary and secondary coders was high across all dichotomous measures (i.e., copying, testing, solving; ≥98% agreement) and continuous measures (i.e., imitation, exploration, unlocking time, persistence, and measures of skill; all ICCs ≥ .93, all ps < .001).

Testing and Copying the Taught Solution. A participant was considered to have successfully tested the taught solution if they fully inserted a color-matched key into a lock corresponding to any box other than the red box (e.g., the pink key on the pink box) at any point during the test. At this point, participants should have evidence to suggest the color rule is ineffective for the novel boxes. In addition, we coded whether children began the session by copying the taught solution. Children were considered to have copied if their first attempt involved trying the red key on the red box.

Time Spent Imitating and Exploring. Our preregistration outlined that exploration and imitation would be coded based on the amount of time that children spent holding color-matched (vs. non-color-matched keys). However, given the nature of the task, this definition did not allow for sufficient construct validity (i.e., children could hold keys without using them, or they could use a color-matched key on a non-color-matched box). Thus, we instead coded imitation as the time taken to abandon the color rule and exploration as the time taken to begin exclusively exploring. Children were considered to have abandoned the color rule when they made their first unlocking attempt with a non-color-matched key (in seconds). Children were considered to have begun exclusively exploring at the moment they attempted their last color-matched key (in seconds). In this way, our measure of imitation reflected the amount of time children spent *only* imitating, while our measure of exploration reflected the amount of time children spent *only* exploring. Using these definitions also allowed for independent classification. That is, because children could reattempt color-matched keys after initially abandoning the color rule, the time taken to begin exclusively exploring was not deterministically related to our measure of imitation.

Persistence. We also coded a measure of persistence. To measure persistence before abandoning the color rule, a score was constructed factoring in the total number of color-matched keys that children attempted and the number of boxes they attempted color-matched keys on. The persistence score was weighted by the number of boxes children attempted color-matched keys on, as attempting color-matched keys on a greater number of boxes would provide better evidence that the color-matched keys were ineffective, relative to simply trying the same color-matched key multiple times on the same box. In this way, the persistence score reflected how thoroughly children attempted the taught solution before abandoning the color rule:

$$\text{Persistence} = (\text{Color Matched Attempts}) \times \frac{\text{Colored Matched Boxes Attempted}}{5}. \quad (1)$$

Solving. A participant was considered to have solved the game if they opened all five boxes in 5 min or less. However, children solved the puzzle at higher rates (66%) than in previous experiments (49%), and they unlocked most of the locks ($M = 4.24$, $SD = 1.21$). Thus, to generate a score with more variability for analyses, we constructed a solving score based on the time taken to unlock. For children who did not unlock all five locks, 300 s was entered (as this was the maximum time the test trial could last).

Unlocking Skill. Finally, we wanted to ensure that any differences we observed did not merely reflect faster unlocking or greater skill with the keys. Thus, we also coded two exploratory measures: time taken to unlock the first lock and attempt rate. Attempts were coded whenever the metal of the key contacted the metal of the lock. Therefore, attempts did not include times when children were merely holding keys. Once attempts were coded, they were divided by unlocking time (in minutes) to generate an attempt rate (allowing us to explore whether there were differences in the speed at which children attempted the keys).

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and we follow Journal Article Reporting Standards (Kazak, 2018). The analyses

presented here were preregistered at https://aspredicted.org/qw5hx.pdf. All data, analysis code, and dollhouse files are available at https://osf.io/qn3w5/. Data were analyzed using RStudio, Version 2023.12.0 (R Core Team, 2023), using the functions binom.test, t.test, lm, glm, and cor.test. All reported tests were two-tailed, and the data met the assumptions of the statistical tests used. Outliers were preregistered as points detected at $SD = 3$ above/below the mean and were removed from analyses using pairwise deletion (see Supplemental Table S9 for mean and standard variation for each variable). Under this definition, outliers were detected for the attempt rate ($n = 1$), the time taken to unlock the first lock ($n = 2$), and persistence ($n = 1$). Thus, participants were only excluded from analyses that utilized variables with outlying values rather than from all analyses.

Results

Manipulation Check

To ensure that the problem-solving games children played were indeed perceived as stereotypically feminine and masculine, we began by analyzing children's gender typicality ratings for the frog game and dollhouse. We first compared the ratings for the two games using a paired-sample *t* test. Indeed, children perceived the dollhouse game as significantly more "for girls" than the frog video game, $t(99) = 9.14$, $p < .001$, $d = 0.91$. Likewise, one-sample *t* tests comparing children's ratings to the neutral baseline of the scale, 3 (*for boys and girls equally*), revealed that children rated the frog game as significantly more masculine than baseline, $M = 2.45$, $t(99) = -7.84$, $p < .001$, $d = -0.78$, and the dollhouse game as significantly more feminine than baseline, $M = 3.62$, $t(99) = 7.99$, $p < .001$, $d = 0.80$. Thus, our stimuli were effective for inducing variability in gender typing between experiments, and children generally perceived the dollhouse problem-solving game as being of interest to girls.

We also sought to verify that our measures of exploration and imitation varied independently. A Pearson correlation failed to provide evidence that exploration and imitation were significantly related ($p = .21$). Thus, unlike in previous experiments, our measures of imitation (i.e., time to abandon the color rule) and exploration (i.e., time to begin exclusively exploring) did not appear to be inherently related to one another.

Overall Analyses

As before, we first sought to understand whether children overall were engaged with the instruction. Thus, we utilized binomial tests to understand how children broadly engaged with the taught solution. The majority of children began by first attempting the red key on the red box (67%, $p < .001$), and nearly all children tested the taught solution (93%, $p < .001$). This tendency to test the taught solution was significantly higher than in Experiment 2 (74%, $p < .001$), perhaps due to the fact that less effort and skill were required for children to test the taught solution in Experiment 3. Taken together, this evidence suggested that children displayed a general tendency to engage with the taught solution.

Our next goal was to understand how children's divestment from teaching related to their overall performance. Once again, we found that imitation and exploration predicted children's performance: Imitation negatively predicted success and learning, such that the longer it took children to abandon the color rule, the slower they were to solve, $\beta = 0.95$, $SE = 0.44$, $t(97) = 2.17$, $p = .03$, and the fewer correct keys they chose during generalization, $\beta = -0.02$, $SE = 0.01$, $z(97) = -2.40$, $p = .02$. Simultaneously, exploration positively predicted success and learning, such that the faster children were to begin exclusively exploring, the faster they were to solve, $\beta = 0.66$, $SE = 0.05$, $t(97) = 13.93$, $p < .001$, and the more correct keys they chose during generalization, $\beta = -0.004$, $SE = 0.001$, $z(97) = -2.97$, $p = .003$. Thus, children were able to engage with the taught solution to evaluate its effectiveness, and exploration and imitation each predicted success and learning. As such, children that both abandoned the taught solution more quickly and focused on exploring their own solutions more quickly were advantaged in several ways relative to their peers (Figures 5 and 6).

Analyses of Gender Differences

Having elaborated general trends, our next goal was to investigate the role of gender in children's performance (see Supplemental Table S9 for all descriptive statistics). Before constructing our final models, we first looked to see whether doll experience was a significant predictor in any models of gender. Based on our preregistered plan, we did not control for doll experience in any models because it did not predict any dependent variable (all $ps \geq .11$). Given the extremely high rate of testing, we performed analyses of our entire sample rather than just children who tested the taught solution. To this end, we constructed linear regressions predicting time to abandon the color rule, time to begin exclusively exploring, unlocking skill, and solving, as well as logistic models predicting testing, and Poisson models of learning. In each model, we entered the main effects of age and gender as our only predictors. As we did not preregister analyses controlling for unlocking skill, we do not control for unlocking skill in our final models. However, it should be noted that the inclusion of the variable did not affect patterns of significance (see Supplemental Table S10).

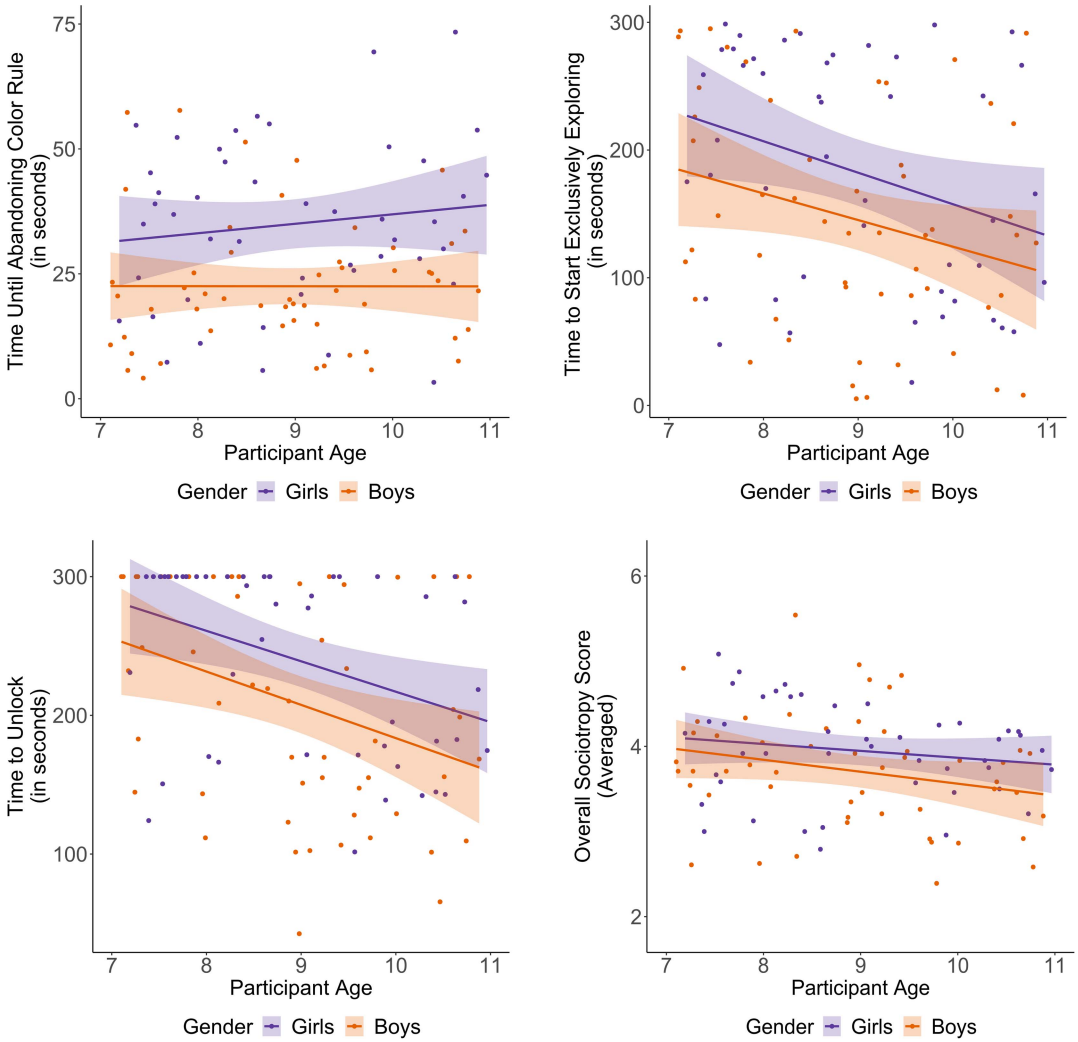
As before, we did not find any evidence that boys and girls differed in their ability. We did not observe gender differences in children's ability to test the taught solution ($p = .99$). As we are cautious to overinterpret a null result, we additionally qualified this analysis with post hoc binomial tests to see whether the majority of girls and the majority of boys, respectively, were able to test the solution and generate evidence of its ineffectiveness. Indeed, the majority of children in both groups were able to effectively engage with the instruction (both $ps \leq .001$). In addition, we did not observe differences in the time children needed to unlock their first lock ($p = .85$) or their overall rate of unlocking attempts ($p = .26$). Thus, we did not find evidence to suggest there were average differences in skill between girls and boys (Table 2).

However, we did find that girls were more likely to engage deeply with the taught solution than boys, as they took longer on average to abandon the color rule, $\beta = 12.39$, $SE = 2.97$, $t(97) = 4.17$, $p < .001$. Consistent with this finding, we also found that girls demonstrated greater persistence in the color rule prior to abandoning it, $\beta = 0.26$, $SE = 0.10$, $t(96) = 2.68$, $p = .009$, than boys.[4] Thus, on average, girls appeared to demonstrate greater persistence, as they tended to spend more time applying the color rule before considering alternatives.

[4] See Supplemental Materials for additional analyses of persistence.

Figure 5

Key Variables in Experiment 3 Plotted Against Participant Age and Gender



Note. Differences were observed in children's tendency to begin by adopting the taught rule (color), as well as the time taken to abandon the color rule and unlock the locks during the test. Similarly, marginal differences were found in sociotropy such that girls scored higher on sociotropy on average than boys. Individual observations are plotted, along with lines of best fit, estimated using linear regressions. As can be seen, while girls' and boys' performance differed on average, there was also substantial overlap in performance. For example, while girls explored less on average than boys, several girls explored at high rates, and several boys explored at low rates. Shaded regions represent 95% confidence intervals. See the online article for the color version of this figure.

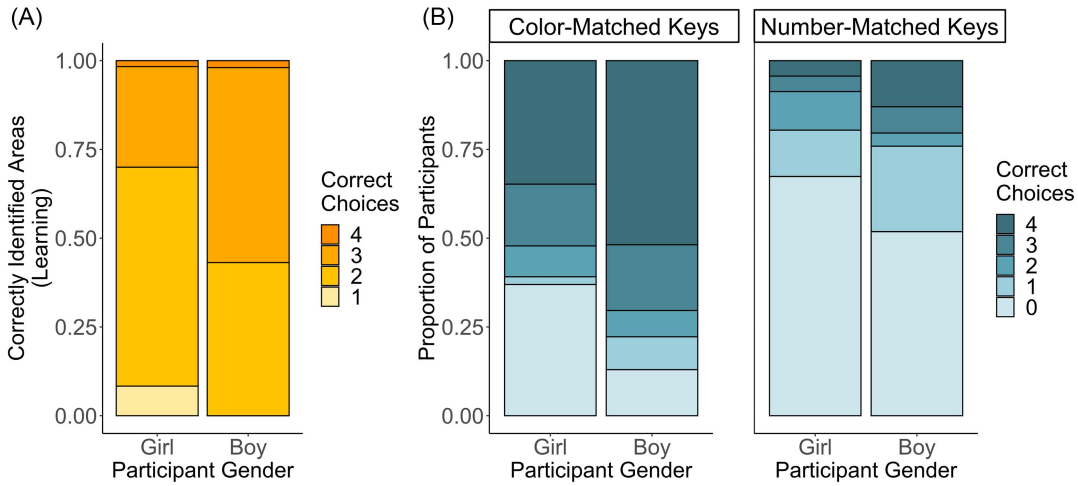
Conversely, we found that girls were slower on average to begin exclusively exploring than boys, $\beta = 37.31$, $SE = 17.33$, $t(97) = 2.15$, $p = .03$. Thus, consistent with our hypotheses and prior experiments, girls tended to demonstrate greater persistence but lower exploration than boys.

In turn, we found that girls solved the problem more slowly on average than boys, $\beta = 31.30$, $SE = 13.89$, $t(97) = 2.25$, $p = .03$, and they tended to answer fewer of the learning questions correctly than boys, $\beta = -0.49$, $SE = 0.23$, $z(97) = -2.16$, $p = .03$. Importantly, these differences in learning appeared to be driven by the fact that girls selected more color-matched keys on average than boys during generalization, $\beta = 0.52$, $SE = 0.17$, $z(97) = 3.12$, $p = .002$. Thus, gender differences appeared to manifest not just in children's immediate performance but also to generalize to novel problems.

Sociotropy Analyses

Using models following the same specifications as the other gender models, we sought to understand whether girls reported greater average sociotropy than boys. Looking at overall scores, we found that girls scored marginally higher in sociotropy than boys on average, $\beta = 0.24$, $SE = 0.13$, $t(94) = 1.91$, $p = .06$, and this appeared to be driven by the Concern For What Others Think subscale, $\beta = 0.36$, $SE = 0.17$, $t(94) = 2.10$, $p = .04$. We further sought to assess how

Figure 6
Performance on Learning Questions



Note. Differences in learning were observed as a function of participant gender such that boys answered significantly more learning questions correctly on average in (A) Experiment 2 and (B) Experiment 3. Because learning scores can only take discrete values between 0 and 4, they are displayed through stacked bar graphs. On all graphs, darker colors represent a greater number of correct answers. Thus, in the case of color-matched selections, the darkest bar corresponds to the proportion of participants that selected color-matched keys on none of the trials, and the lightest bar corresponds to the proportion of participants that selected color-matched keys on all four trials. See the online article for the color version of this figure.

sociotropy related to performance. While we did not find that children's overall sociotropy scores related to their immediate performance during the game (i.e., exploration, imitation, or solving; all $ps \geq .47$), children with higher sociotropy scores tended to select fewer correct keys, $\beta = -0.36$, $SE = 0.18$, $t(94) = -1.97$, $p = .049$, and to select more color-matched keys during generalization, $\beta = 0.37$, $SE = 0.13$, $t(94) = 2.76$, $p = .006$. Because these effect sizes were relatively small, we were underpowered to explore whether gender differences in learning were mediated by sociotropy.

Discussion

Altogether, these results suggest that gender differences in divestment from ineffective teaching emerged across contexts, such

Table 2
Participant Gender Predicts Key Measures in Experiment 3

	Time to abandon taught rule		Time to exclusively explore	
Predictor	β (*SE*)	95% CI	β (*SE*)	95% CI
Intercept	14.79 (11.65)	[−8.04, 37.62]	348.35 (68.03)***	[215.01, 481.69]
Age	0.87 (1.29)	[−1.65, 3.39]	−22.59 (7.51)**	[−37.31, −7.87]
Gender (girl)	12.39 (2.97)***	[6.57, 18.20]	37.31 (17.33)*	[3.34, 71.28]
	Unlocking time		Learning (correct keys)	
Predictor	β (*SE*)	95% CI	β (*SE*)	95% CI
Intercept	415.09 (54.51)***	[308.26, 521.92]	−1.38 (0.86)	[−3.07, 0.32]
Age	−23.05 (6.02)***	[−34.85, −11.26]	0.16 (0.09)	[−0.03, 0.34]
Gender (girl)	31.30 (13.89)*	[4.08, 58.51]	−0.49 (0.23)*	[−0.93, −0.05]
	Sociotropy (overall)		Sociotropy (concern)	
Predictor	β (*SE*)	95% CI	β (*SE*)	95% CI
Intercept	4.72 (0.49)***	[3.76, 5.68]	4.87 (0.66)***	[3.57, 6.17]
Age	−0.11 (0.05)*	[−0.22, −0.006]	−0.16 (0.07)*	[−0.30, −0.01]
Gender (girl)	0.24 (0.13)†	[−0.006, 0.49]	0.36 (0.17)*	[0.02, 0.69]

Note. *SE* = standard error; CI = confidence interval.
† $p < .08$. * $p < .05$. ** $p < .01$. *** $p < .001$.

that girls were more likely to adopt and persist in the taught solution and less likely to explore novel solutions than boys, even during a feminine-stereotyped task that differed in its problem structure from Experiments 1 and 2. However, as before, girls did not appear to merely be worse at the game, as they did not differ from boys in their ability to test the taught solution nor in the speed of their unlocking attempts. Further, on average, girls showed greater persistence in the taught solution. Although this persistence would likely be beneficial in the context of effective teaching, provided ineffective teaching, girls' average tendency to adopt taught solutions created disadvantages, translating to slower solving. Likewise, these gender differences appear to translate beyond the immediate problem-solving context, as girls were less likely, on average, to correctly answer learning questions, suggesting they are more likely to generalize ineffective teaching to new contexts.

General Discussion

Learning through others' teaching is the primary way children learn about the world, though teaching is sometimes inaccurate. However, in real-world settings, teachers may not provide overt cues (e.g., prior mistakes) to aid children in identifying these inaccuracies. As such, the present experiments had four primary objectives: (1) understand how elementary-aged children engaged with ineffective teaching when explicit cues were not provided to ascertain effectiveness; (2) evaluate whether gender differences emerged in divestment from ineffective teaching and whether these differences created disparities in success; (3) evaluate the robustness of these patterns across contexts varying in gender typing, problem structure, and measures of learning; and (4) explore the extent to which girls and boys differed on average in sociotropy (a proxy for people-pleasing). To this end, we designed a video game and a dollhouse problem-solving game that required participants to test the effectiveness of a taught solution using only their own exploratory actions. Our findings revealed that, overall, children calibrated to the evidence they generated and balanced between learning from instruction and exploration. However, significant, replicable gender differences emerged, such that girls were generally more hesitant to prioritize exploring their own solutions than boys, leading to differences in success and learning even in feminine-stereotyped contexts. In addition, we observed gender differences in sociotropy and found sociotropy related to children's tendency to generalize the ineffective solution to new contexts, providing initial evidence that people-pleasing may play a role in shaping gender differences.

Overall Patterns in Divestment in the Absence of Overt Cues

By creating a context in which children had to utilize their own actions to ascertain the quality of teaching, we observed several optimized behaviors. As instruction is generally correct and efficient (Csibra & Gergely, 2006, 2009; Kirschner et al., 2006; Stockard et al., 2018), it was reasonable for children to first capitalize on instruction to try to win the games. This was reflected in high initial rates of copying and high rates of testing the taught solution. Once children tested the solution, they generated evidence that it was ineffective, which ought to have prompted behavioral change. Indeed, children generally increased their exploration after successfully testing the taught solution. Likewise, children who adjusted to this evidence, divesting from teaching more quickly and spending more time exploring their own ideas, experienced more success and learned more.

These findings are consistent with prior work on children's resource rationality. Humans seek to minimize the effort they exert when possible and expect other agents to do so as well (Jara-Ettinger et al., 2016). Children balance the effort they exert from very early in life, prioritizing exploration over imitation when a solution does not work (Radovanovic et al., 2021; Solby et al., 2021), balancing problem-solving strategies (Lucca et al., 2020), and persisting less when they know teaching is available (Rett & Walker, 2020). The results elaborated here are consistent with these trends, demonstrating that children first exploited teaching presumably under the assumption that the taught solution was correct and efficient but tended to divest from this teaching when it garnered diminishing returns and continued to fail. Importantly, children were able to navigate these tradeoffs even when relying on self-generated evidence to evaluate teaching effectiveness despite the inherent ambiguity of relying only on their own experiences of failure with the taught solution.

Gender Differences in Divestment

Critically, we observed that girls' and boys' average tendencies to divest from instruction differed. Specifically, girls were more hesitant, on average, to explore their own ideas than boys: exploring for a smaller proportion of time and imitating for a greater proportion of time than boys in Experiments 1 and 2 and taking longer to abandon the taught solution and begin exclusively exploring in Experiment 3. In turn, these differences in divestment were also reflected in achievement differences. Girls were less successful during the immediate problem-solving games and answered fewer learning questions correctly. Importantly, the observed gender differences were not attributable to differences in skill. We did not observe evidence of skill differences between girls and boys (e.g., unlocking skill), and across all experiments, we found that both girls and boys successfully tested the taught solutions at high rates. Further, controlling for measures of experience (i.e., video game experience) or skill (i.e., unlocking skill) did not alter patterns of significance in our findings.

On the other hand, several patterns were consistent with the notion that girls are more likely to be encouraged to people-please than boys (i.e., prioritize others' feelings, obey authority, and avoid self-promotion; Mickelson, 1989; Orr, 2011). First, in Experiment 2, we observed that girls' lower average levels of solving and learning were mediated by differences in exploration. This was especially true for solving, as the direct effect of gender on solving was no longer significant once average differences in exploration after testing the taught solution were accounted for. Thus, while we failed to find evidence of skill differences, we found direct evidence that girls' tendency to divest less from teaching than boys parsimoniously explained achievement differences. In addition, Experiment 3 provided initial evidence that people-pleasing socialization may shape gender differences in divestment. In Experiment 3, we observed that girls scored higher on average in sociotropy than boys, replicating past work done with older kids (e.g., Baron & Peixoto, 1991; Brenning et al., 2011; Calvete, 2011; Teppers et al., 2013) in the youngest sample to our knowledge (i.e., 7- to 10-year-olds). In turn, we observed that sociotropy related to differences in learning, such that children higher in sociotropy were more likely to generalize the

ineffective teaching to a new context. Taken together, these findings suggest that average differences in children's concern for maintaining positive relationships are a more compelling explanation than the notion that girls merely underperformed.

Importantly, prior work has suggested that gender disparities in North America are often starker in masculine-stereotyped contexts than in feminine-stereotyped contexts due to differences in the extent to which girls and women feel they belong (e.g., video gaming; Kaye & Pennington, 2016) or differences in the extent to which brilliance is emphasized (e.g., Bian et al., 2018). Thus, we also sought to understand how the gender differences we observed may vary across contexts. To this end, we observed that girls demonstrated greater average hesitation to move away from teaching even in an overtly feminine context. Likewise, as we varied the type of problem-solving task (i.e., search task vs. rule learning), our work also suggests that gender differences may exist across problem types and not just merely for problems in which sustained exploration is inherently necessary for success. As such, this work provides initial evidence for domain-general differences in responses to ineffective teaching. On the other hand, children faced a high degree of ambiguity during our problem-solving games because they had to rely on self-generated evidence to infer that the taught solution was ineffective. Given this ambiguity, children may be more sensitive to normative expectations for how they ought to behave compared to contexts in which they have high certainty that the teaching is ineffective. As such, these gender differences may not be present in less ambiguous contexts.

Along this theme, it is also critical to highlight the patterns observed here as different problem-solving strategies rather than simply a failure by girls. Of course, there are many situations where focusing deeply on instruction and persisting through difficulty would be beneficial. Particularly when information is difficult or impossible for children to extract independently, expert instruction can be more effective for learning than children's own exploration (Mayer, 2004). In these contexts, girls may be advantaged for the same reasons they are disadvantaged in the context of our experiments. Indeed, girls generally outperform boys academically (Duckworth & Seligman, 2006; Matthews et al., 2009; Orr, 2011). Exploring entirely novel solutions can also be time-consuming and costly. Thus, while populations benefit from having explorers, individuals who spend a large proportion of their time exploring may not themselves experience success (Wisdom & Goldstone, 2011). As such, boys may instead be disadvantaged, on average, in contexts in which taught solutions are effective but confusing or difficult to apply, as they may be less likely to persist until the taught solution succeeds and instead invest time in alternative solutions that take a long time to be developed. Thus, it is important to understand how differences in socialization may affect adaptation to evidence across contexts for both girls and boys.

Limitations of the Present Work

On the other hand, although we demonstrated that gender differences, on average, reported sociotropy and relations between sociotropy and learning, we did not find sociotropy related to measures of immediate performance, such as the time to abandon the taught solution or immediate success. There are several possibilities for these results. First, as we sought to use a validated, preexisting measure, our ability to directly assess people-pleasing with elementary-aged children was limited. We incorporated a measure of sociotropy given that it was conceptually most similar to people-pleasing (i.e., concern for maintaining positive relationships) and because of robust evidence of gender differences in sociotropy (Yang & Girgus, 2019). However, research on sociotropy has almost exclusively been done in older populations, such that even investigations with adolescents are limited. To this end, we took several measures to ensure that the scale was as valid as possible for young children: adapting the language in questions to be more developmentally appropriate, using a standardized script to provide examples, and allowing children to skip questions as needed. However, despite these efforts, the validity of the scale was likely limited in our sample compared to investigations with adolescents and adults.

Second, sociotropy likely does not fully capture people-pleasing. Here, we propose that people-pleasing manifests in a multifaceted set of behaviors revolving around prioritizing others' ideas, thoughts, and emotions above one's own and fearing repercussions for failing to do so. While sociotropy captures an overlapping construct, it revolves mostly around maintaining positive relationships and does not tap into all components, such as self-promotion or obedience to authority figures. Thus, it may be that different facets of people-pleasing drive differences in immediate performance (e.g., time taken to abandon the taught solution) from those that drive differences in generalization because different subprocesses may be implicated in different facets of children's responses to ineffective teaching. For instance, children provided their responses verbally to the experimenter during generalization, while the experimenter simply sat next to them and did not interact with them during the test. As such, it may be the case that concern for maintaining positive relationships plays a stronger role in shaping differences during generalization, while a different facet (e.g., sensitivity to authority) may shape test performance since the role of the experimenter is much more ambiguous during the test than generalization. Understanding these dynamics and mechanisms provides rich ground for future investigation.

Likewise, our measurement of gender is limited. As we primarily hypothesized that gender differences would be driven by differences in socialization rather than by inherent differences between children, we measured gender through caregiver report. In this way, our measure could reflect how children are perceived by their close family and, as a result, relate better to the socialization pressures they may be exposed to in their daily lives. However, caregiver-reported gender may not reflect children's actual gender identity, both because caregivers may discourage nonconformity (D'Augelli et al., 2006; Spivey et al., 2018) and because caregivers may be unaware of how their children identify (Pew Research Center, 2013). Our measure of gender identity was also limited by its categorical nature, as gender is multidimensional and encompasses not just labeling, but also other constructs (e.g., typicality, contentedness, other-gender identification; Egan & Perry, 2001; Martin et al., 2017). As such, it is important to note that gender is not an essentialist predetermined category (Cikara et al., 2022) and that there was considerable variability among participants grouped within the same gender category in our experiments. Further, the categorical nature of our measure required us to exclude a nonbinary participant from analyses of gender. However, many people do not identify within the binary (Ainsworth, 2015), so nonbinary children's experiences are important to consider both in their own right and for a more complete theoretical understanding of gender (Dunham & Olson, 2016). As such, it is important that future work considers a more multifaceted

view of gender to avoid essentializing claims and to better understand the diversity of children's experiences.

Constraints on Generality

Finally, we wish to elaborate constraints on the generalizability of the current experiments. While families in our samples had higher than average levels of education and wealth, it is important to acknowledge that the type of ineffective teaching we aimed to target (i.e., wherein teachers are unaware of ineffectiveness) typically pertains to structural knowledge relating to historical oppression (King, 2017; Solomon, 2002; Yosso, 2005). Information about oppression is particularly likely to be omitted from curricula or incorrectly taught because it stands in contrast to systems of oppression (e.g., "color-blind" narratives where in racism is not acknowledged, meritocratic myths), and North American teachers are unlikely to be aware of these inaccuracies as most are White middle-class women and not trained to understand systems of oppression (Bartolomé, 2004). Given this reality, it is important to understand how gender differences in divestment may differ for children with less cultural capital (e.g., Anyon, 1980; Lareau & Weininger, 2003).

It is also important to consider how gendered socialization practices may vary, both within North America and cross-culturally. The gendered expectations and messages children are exposed to are racialized. For example, while children tend to associate brilliance with White men and boys (relative to women and girls; Bian et al., 2017), children are less likely to associate brilliance with Black men (Jaxon et al., 2019) and East Asian men (Shu et al., 2022) when compared to Black and East Asian women. Likewise, girls' racial and ethnic experiences can shape their ideas of leadership (e.g., Hernández-Matías et al., 2023; Mims & Kaler-Jones, 2020), likely informing their responses to ineffective teaching and optimal strategies for fostering girls' comfort divesting from ineffective teaching. Finally, while higher rates of sociotropy are generally seen in women across cultures, these differences are less pronounced in cultures that are higher in relatedness and collectivism (Yang & Girgus, 2019), perhaps because overall cultural ideals are more closely aligned with concern for maintaining positive relationships (Cuddy et al., 2015). As such, gender differences in divestment from teaching and success may also be less pronounced among children in cultures that prioritize relatedness. Future work should systematically evaluate how socioeconomic status, race, and cultural background shape gender differences in divestment, as well as how these dynamics manifest in real-world settings. Until such work is conducted, these results should be generalized with caution.

Overall, however, the trends articulated here in regard to gender have the potential to explain broad societal disparities between men and women. A key focus in psychology and education has been to leverage interventions early in development to reduce gender disparities in particular professions, especially science, technology, engineering, and mathematics (Boston & Cimpian, 2018; Master et al., 2021; Smyth & Nosek, 2015). However, while these disparities have largely been framed as an issue of gender stereotypes in particular professions, women are underrepresented at leadership levels in most fields relative to their participation at entry levels, and many fields have pipeline problems (Clay, 2017; Kenan Institute, 2021; Sutton, 2014). These trends hold true even in professions traditionally viewed as feminine and in which women are broadly overrepresented. For example, while 76% of public school teachers in the United States identify as women, women comprise only 27% of superintendents (Tienken & Domenech, 2021). Of course, while there are many workplace biases that influence women's pursuit of leadership positions (Stamarski & Son Hing, 2015), the results here elucidate a potential overarching disparity for future research: Women likely do not feel the same power to explore and advocate for their own ideas as men. While entry-level success often relies on mastering the "tools of the trade" (i.e., learning existing knowledge and practices), advancement and leadership often require one to abandon conformity and innovate, introducing novel ideas. Socialization toward people-pleasing may allow women to initially prosper, but women simultaneously likely struggle to advance when they receive backlash for advancing their own ideas and innovations.

Conclusion

Taken together, our results suggested that gendered socialization practices shape children's responses to ineffective teaching such that girls were more likely, on average, to persist in the taught solution, while boys were more likely, on average, to explore alternatives. In turn, these average differences in exploration translate to differences in success and learning that can be seen in both masculine-stereotyped and feminine-stereotyped contexts. Likewise, we provide the first evidence that gender differences in sociotropy can be observed in children as young as 7 to 10 years old. However, our contention is not that children should always divest from teaching and explore their own solutions. In fact, it is likely that girls' tendency to persist in taught solutions enables them to achieve greater academic success, on average, relative to boys. Rather, the results here demonstrate that socialization practices may limit children of all genders from flexibly switching between strategies (i.e., persisting and divesting), as they gain evidence of teaching (in)effectiveness, and point to the importance of understanding how socialization practices shape children's ability to gain accurate information and succeed across contexts.

References

Åhslund, I., & Boström, L. (2018). Teachers' perceptions of gender differences -What about boys and girls in the classroom? *International Journal of Learning, Teaching and Educational Research*, *17*(4), 28–44. https://doi.org/10.26803/ijlter.17.4.2

Ainsworth, C. (2015). Sex redefined. *Nature*, *518*(7539), 288–291. https://doi.org/10.1038/518288a

Anyon, J. (1980). Social class and the hidden curriculum of work. *Journal of Education*, *162*(1), 67–92. https://doi.org/10.1177/002205748016200106

Baron, P., & Peixoto, N. (1991). Depressive symptoms in adolescents as a function of personality factors. *Journal of Youth and Adolescence*, *20*(5), 493–500. https://doi.org/10.1007/BF01540633

Bartolomé, L. I. (2004). Critical pedagogy and teacher education: Radicalizing prospective teachers. *Teacher Education Quarterly*, *31*(1), 97–122.

Beck, A. T. (1983). Cognitive therapy of depression: New perspectives. In P. J. Clayton & J. E. Barrett (Eds.), *Treatment of depression: Old controversies and new approaches* (pp. 265–290). Raven Press.

Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389–391. https://doi.org/10.1126/science.aah6524

Bian, L., Leslie, S.-J., Murphy, M. C., & Cimpian, A. (2018). Messages about brilliance undermine women's interest in educational and professional

opportunities. *Journal of Experimental Social Psychology*, *76*, 404–420. https://doi.org/10.1016/j.jesp.2017.11.006

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330. https://doi.org/10.1016/j.cognition.2010.10.001

Boston, J. S., & Cimpian, A. (2018). How do we encourage gifted girls to pursue and succeed in science and engineering? *Gifted Child Today*, *41*(4), 196–207. https://doi.org/10.1177/1076217518786955

Brenning, K., Soenens, B., Braet, C., & Bosmans, G. (2011). The role of depressogenic personality and attachment in the intergenerational similarity of depressive symptoms: A study with early adolescents and their mothers. *Personality and Social Psychology Bulletin*, *37*(2), 284–297. https://doi.org/10.1177/0146167210393533

Brescoll, V. L. (2011). Who takes the floor and why. *Administrative Science Quarterly*, *56*(4), 622–641. https://doi.org/10.1177/0001839212439994

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–340. https://doi.org/10.1016/j.cognition.2010.12.001

Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, *106*(4), 676–713. https://doi.org/10.1037/0033-295X.106.4.676

Calvete, E. (2011). Integrating sociotropy, negative inferences and social stressors as explanations for the development of depression in adolescence: Interactive and mediational mechanisms. *Cognitive Therapy and Research*, *35*(5), 477–490. https://doi.org/10.1007/s10608-010-9320-4

Carr, K., Kendal, R. L., & Flynn, E. G. (2015). Imitate or innovate? Children's innovation is influenced by the efficacy of observed behaviour. *Cognition*, *142*, 322–332. https://doi.org/10.1016/j.cognition.2015.05.005

Cikara, M., Martinez, J. E., & Lewis, N. A., Jr. (2022). Moving beyond social categories by incorporating context in social psychological theory. *Nature Reviews Psychology*, *1*(9), 537–549. https://doi.org/10.1038/s44159-022-00079-3

Clay, R. A. (2017). Women outnumber men in psychology, but not in the field's top echelons. *Monitor on Psychology*, *48*(7). https://www.apa.org/monitor/2017/07-08/women-psychology

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349. https://doi.org/10.1016/j.cognition.2011.03.003

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R. L., Verheggen, B., Maibach, E. W., Carlton, J. S., Lewandowsky, S., Skuce, A. G., Green, S. A., Nuccitelli, D., Jacobs, P., Richardson, M., Winkler, B., Painting, R., & Rice, K. (2016). Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, *11*(4), Article 048002. https://doi.org/10.1088/1748-9326/11/4/048002

Corriveau, K. H., Pickard, K., & Harris, P. L. (2011). Preschoolers trust particular informants when learning new names and new morphological forms. *British Journal of Developmental Psychology*, *29*(1), 46–63. https://doi.org/10.1348/2044-835X.002009

Crick, N. R., & Nelson, D. A. (2002). Relational and physical victimization within friendships: Nobody told me there'd be friends like these. *Journal of Abnormal Child Psychology*, *30*(6), 599–607. https://doi.org/10.1023/A:1020811714064

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development* (Vol. XXI, pp. 249–274). Oxford University Press. https://doi.org/10.1093/oso/9780198568742.003.0011

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Cuddy, A. J. C., Wolf, E. B., Glick, P., Crotty, S., Chong, J., & Norton, M. I. (2015). Men as cultural ideals: Cultural values moderate gender stereotype content. *Journal of Personality and Social Psychology*, *109*(4), 622–635. https://doi.org/10.1037/pspi0000027

D'Augelli, A. R., Grossman, A. H., & Starks, M. T. (2006). Childhood gender atypicality, victimization, and PTSD among lesbian, gay, and bisexual youth. *Journal of Interpersonal Violence*, *21*(11), 1462–1482. https://doi.org/10.1177/0886260506293482

Da Silva, A. G., Malloy-Diniz, L. F., Garcia, M. S., & Rocha, R. (2020). Attention-deficit/hyperactivity disorder and women. In J. Rennó, Jr., G. Valadares, A. Cantilino, J. Mendes-Ribeiro, R. Rocha, & A. Geraldo da Silva (Eds.), *Women's mental health* (pp. 215–219). Springer. https://doi.org/10.1007/978-3-030-29081-8_15

Daubman, K. A., Heatherington, L., & Ahn, A. (1992). Gender and the self-presentation of academic achievement. *Sex Roles*, *27*(3–4), 187–204. https://doi.org/10.1007/BF00290017

De Bolle, M., De Fruyt, F., McCrae, R. R., Löckenhoff, C. E., Costa, P. T., Aguilar-Vafaie, M. E., Ahn, C.-K., Ahn, H.-N., Alcalay, L., Allik, J., Avdeyeva, T. V., Bratko, D., Brunner-Sciarra, M., Cain, T. R., Chan, W., Chittcharat, N., Crawford, J. T., Fehr, R., Ficková, E., … Terracciano, A. (2015). The emergence of sex differences in personality traits in early adolescence: A cross-sectional, cross-cultural study. *Journal of Personality and Social Psychology*, *108*(1), 171–185. https://doi.org/10.1037/a0038497

Debowski, S., Wood, R. E., & Bandura, A. (2001). Impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic search. *Journal of Applied Psychology*, *86*(6), 1129–1141. https://doi.org/10.1037/0021-9010.86.6.1129

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. https://doi.org/10.1037/0022-0663.98.1.198

Dunham, Y., & Olson, K. R. (2016). Beyond discrete categories: Studying multiracial, intersex, and transgender children will strengthen basic developmental science. *Journal of Cognition and Development*, *17*(4), 642–665. https://doi.org/10.1080/15248372.2016.1195388

Dweck, C. S., Davidson, W., Nelson, S., & Enna, B. (1978). Sex differences in learned helplessness: II. The contingencies of evaluative feedback in the classroom and III. An experimental analysis. *Developmental Psychology*, *14*(3), 268–276. https://doi.org/10.1037/0012-1649.14.3.268

Egan, S. K., & Perry, D. G. (2001). Gender identity: A multidimensional analysis with implications for psychosocial adjustment. *Developmental Psychology*, *37*(4), 451–463. https://doi.org/10.1037/0012-1649.37.4.451

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Fuller, K., Clonan-Roy, K., Goncy, E., & Naser, S. (2022). The omission and minimisation of sexual decision-making skills in US sex education textbooks. *Sex Education*, *22*(4), 409–423. https://doi.org/10.1080/14681811.2021.1949974

Geller, S. E., Koch, A. R., Roesch, P., Filut, A., Hallgren, E., & Carnes, M. (2018). The more things change, the more they stay the same. *Academic Medicine*, *93*(4), 630–635. https://doi.org/10.1097/ACM.0000000000002027

Gralewski, J. (2019). Teachers' beliefs about creative students' characteristics: A qualitative study. *Thinking Skills and Creativity*, *31*, 138–155. https://doi.org/10.1016/j.tsc.2018.11.008

Gralewski, J., & Karwowski, M. (2013). Polite girls and creative boys? Students' gender moderates accuracy of teachers' ratings of creativity. *The Journal of Creative Behavior*, *47*(4), 290–304. https://doi.org/10.1002/jocb.36

Griffith, J. A., & Brem, S. K. (2004). Teaching evolutionary biology: Pressures, stress, and coping. *Journal of Research in Science Teaching*, *41*(8), 791–809. https://doi.org/10.1002/tea.20027

Guttmacher Institute. (2022, January 1). *Sex and HIV education*. https://www.guttmacher.org/state-policy/explore/sex-and-hiv-education

Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are under-informative. *Cognition*, *132*(3), 335–341. https://doi.org/10.1016/j.cognition.2014.04.013

Hamilton, L. C., Hartter, J., Lemcke-Stampone, M., Moore, D. W., & Safford, T. G. (2015). Tracking public beliefs about anthropogenic climate change. *PLOS ONE*, *10*(9), Article e0138208. https://doi.org/10.1371/journal.pone.0138208

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.

Hernández-Matías, L., Díaz-Muñoz, G., & Guerrero-Medina, G. (2023). Seeds of success: Empowering latina STEM girl ambassadors through role models, leadership, and stem-related experiences. *Journal of STEM Outreach*, *6*(2), 1–14. https://doi.org/10.15695/jstem/v6i2.03

Hinshaw, S. P., Nguyen, P. T., O'Grady, S. M., & Rosenthal, E. A. (2022). Annual research review: Attention-deficit/hyperactivity disorder in girls and women: Underrepresentation, longitudinal processes, and key directions. *Journal of Child Psychology and Psychiatry*, *63*(4), 484–496. https://doi.org/10.1111/jcpp.13480

Hoehl, S., Keupp, S., Schleihauf, H., Mcguigan, N., Buttelmann, D., & Whiten, A. (2019). "Over-imitation": A review and appraisal of a decade of research. *Developmental Review*, *51*, 90–108. https://doi.org/10.1016/j.dr.2018.12.002

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Jaxon, J., Lei, R. F., Shachnai, R., Chestnut, E. K., & Cimpian, A. (2019). The acquisition of gender stereotypes about intellectual ability: Intersections with race. *Journal of Social Issues*, *75*(4), 1192–1215. https://doi.org/10.1111/josi.12352

Jones, S., & Myhill, D. (2004). "Troublesome boys" and "compliant girls": Gender identity and perceptions of achievement and underachievement. *British Journal of Sociology of Education*, *25*(5), 547–561. https://doi.org/10.1080/0142569042000252044

Jordan, J. V., Kaplan, A. G., Miller, J. B., Stiver, I. P., & Surrey, J. L. (1991). *Women's growth in connection: Writings from the stone center*. Guilford Press.

Kaye, L. K., & Pennington, C. R. (2016). "Girls can't play": The effects of stereotype threat on females' gaming performance. *Computers in Human Behavior*, *59*, 202–209. https://doi.org/10.1016/j.chb.2016.02.020

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kenan Institute. (2021, May 12). *Fixing the leaky gender equality pipeline*. https://kenaninstitute.unc.edu/kenan-insight/fixing-the-leaky-gender-equality-pipeline/

King, L. J. (2017). The status of Black history in U.S. schools and society. *Social Education*, *8*(1), 14–18.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effect of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667. https://doi.org/10.1111/j.0956-7976.2004.00737.x

Lareau, A., & Weininger, E. B. (2003). Cultural capital in educational research: A critical assessment. *Theory and Society*, *32*(5/6), 567–606. https://doi.org/10.1023/B:RYSO.0000004951.04408.b0

Lei, R. F., Green, E. R., Leslie, S. J., & Rhodes, M. (2019). Children lose confidence in their potential to "be scientists," but not in their capacity to "do science". *Developmental Science*, *22*(6), Article e12837. https://doi.org/10.1111/desc.12837

Leiserowitz, A., Maibach, E., Rosenthal, S., Kotcher, J., Bergquist, P., Ballew, M., Goldberg, M., Gustafson, A., & Wang, X. (2020). *Climate change in the American mind: April 2020*. Yale University and George Mason University, Yale Program on Climate Change Communication.

Letendre, J. (2007). "Sugar and spice but not always nice": Gender socialization and its impact on development and maintenance of aggression in adolescent girls. *Child & Adolescent Social Work Journal*, *24*(4), 353–368. https://doi.org/10.1007/s10560-007-0088-7

Lewandowsky, S., Oreskes, N., Risbey, J. S., Newell, B. R., & Smithson, M. (2015). Seepage: Climate change denial and its effect on the scientific community. *Global Environmental Change*, *33*, 1–13. https://doi.org/10.1016/j.gloenvcha.2015.02.013

Lucca, K., Horton, R., & Sommerville, J. A. (2020). Infants rationally decide when and how to deploy effort. *Nature Human Behaviour*, *4*(4), 372–379. https://doi.org/10.1038/s41562-019-0814-0

Mansukhani, N. A., Yoon, D. Y., Teter, K. A., Stubbs, V. C., Helenowski, I. B., Woodruff, T. K., & Kibbe, M. R. (2016). Determining if sex bias exists in human surgical clinical research. *JAMA Surgery*, *151*(11), 1022–1030. https://doi.org/10.1001/jamasurg.2016.2032

Martin, C. L., Andrews, N. C. Z., England, D. E., Zosuls, K., & Ruble, D. N. (2017). A dual identity approach for conceptualizing and measuring children's gender identity. *Child Development*, *88*(1), 167–182. https://doi.org/10.1111/cdev.12568

Master, A., Meltzoff, A. N., & Cheryan, S. (2021). Gender stereotypes about interests start early and cause gender disparities in computer science and engineering. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(48), Article e2100030118. https://doi.org/10.1073/pnas.2100030118

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology*, *101*(3), 689–704. https://doi.org/10.1037/a0014240

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist*, *59*(1), 14–19. https://doi.org/10.1037/0003-066X.59.1.14

Mickelson, R. A. (1989). Why does Jane read and write so well? The anomaly of women's achievement. *Sociology of Education*, *62*(1), 47–63. https://doi.org/10.2307/2112823

Mims, L. C., & Kaler-Jones, C. (2020). Running, running the show: Supporting the leadership development of black girls in middle school. *Middle School Journal*, *51*(2), 16–24. https://doi.org/10.1080/00940771.2019.1707342

Montroy, J. J., Bowles, R. P., Skibbe, L. E., McClelland, M. M., & Morrison, F. J. (2016). The development of self-regulation across early childhood. *Developmental Psychology*, *52*(11), 1744–1762. https://doi.org/10.1037/dev0000159

Moss-Racusin, C. A., & Rudman, L. A. (2010). Disruptions in women's self-promotion: The backlash avoidance model. *Psychology of Women Quarterly*, *34*(2), 186–202. https://doi.org/10.1111/j.1471-6402.2010.01561.x

Orr, A. J. (2011). Gendered capital: Childhood socialization and the "Boy crisis" in education. *Sex Roles*, *65*(3–4), 271–284. https://doi.org/10.1007/s11199-011-0016-3

Pew Research Center. (2013). *A survey of LGBT Americans: Attitudes, experiences and values in changing times.*

Plutzer, E., Lee, H. A., Rosenau, J., McCaffrey, M. S., Berbeco, M., & Reid, A. H. (2016). *Mixed messages: How climate is taught in America's schools*. National Center for Science Education. http://ncse.com/files/MixedMessages.pdf

Pörtner, H.-O., Roberts, D., Tignor, M., Poloczanska, E., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., & Rama, B. (Eds.). (2022, February 27). *Climate change 2022: Impacts, adaptation, and vulnerability*. Intergovernmental Panel on Climate Change.

Prakash, V. S., Mansukhani, N. A., Helenowski, I. B., Woodruff, T. K., & Kibbe, M. R. (2018). Sex bias in interventional clinical trials. *Journal of Women's Health*, *27*(11), 1342–1348. https://doi.org/10.1089/jwh.2017.6873

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717–731. https://doi.org/10.3758/BF03206553

Psychology Today. (2024). *People-pleasing*. https://www.psychologytoday.com/ca/basics/people-pleasing

Quinn, P. O., & Madhoo, M. (2014). A review of attention-deficit/hyperactivity disorder in women and girls: Uncovering this hidden diagnosis. *The Primary Care Companion for CNS Disorders*, *16*(3). https://doi.org/10.4088/PCC.13r01596

R Core Team. (2023). *RStudio* (Version 2023.12.0) [Computer software]. https://www.R-project.org

Radovanovic, M., Solby, H., Soldovieri, A., & Sommerville, J. A. (2021). *Try smarter, not harder: Exploration and strategy diversity predict infant persistence* [Conference session]. Proceedings of the 43rd Annual Meeting of the Cognitive Science Society. https://escholarship.org/uc/cognitivesciencesociety/43/43

Radovanovic, M., Yucer, E., & Sommerville, J. A. (2024, March 14). *Girls persist more but divest less from ineffective teaching than boys*. https://osf.io/qn3w5/

Rett, A., & Walker, C. M. (2020). *Knowing when to quit: Children consider access to solutions when deciding whether to persist* [Conference session]. Proceedings of the 42st Annual Conference of the Cognitive Science Society.

Robins, C. J., Ladd, J., Welkowitz, J., Blaney, P. H., Diaz, R., & Kutcher, G. (1994). The Personal Style Inventory: Preliminary validation studies of new measures of sociotropy and autonomy. *Journal of Psychopathology and Behavioral Assessment*, *16*(4), 277–300. https://doi.org/10.1007/BF02239408

Ronfard, S., Chen, E. E., & Harris, P. L. (2018). The emergence of the empirical stance: Children's testing of counterintuitive claims. *Developmental Psychology*, *54*(3), 482–493. https://doi.org/10.1037/dev0000455

Ruggeri, A., Swaboda, N., Sim, Z. L., & Gopnik, A. (2019). Shake it baby, but only when needed: Preschoolers adapt their exploratory strategies to the information structure of the task. *Cognition*, *193*, Article 104013. https://doi.org/10.1016/j.cognition.2019.104013

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for rewards like a child means less generalization and more directed exploration. *Psychological Science*, *30*(11), 1561–1572. https://doi.org/10.1177/0956797619863663

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045–1050. https://doi.org/10.1037/0012-1649.43.4.1045

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162–176. https://doi.org/10.1037/0012-1649.40.2.162

Scott, P. E., Unger, E. F., Jenkins, M. R., Southworth, M. R., McDowell, T.-Y., Geller, R. J., Elahi, M., Temple, R. J., & Woodcock, J. (2018). Participation of women in clinical trials supporting FDA approval of cardiovascular drugs. *Journal of the American College of Cardiology*, *71*(18), 1960–1969. https://doi.org/10.1016/j.jacc.2018.02.070

Shu, Y., Hu, Q., Xu, F., & Bian, L. (2022). Gender stereotypes are racialized: A cross-cultural investigation of gender stereotypes about intellectual talents. *Developmental Psychology*, *58*(7), 1345–1359. https://doi.org/10.1037/dev0001356

Sim, Z. L., & Xu, F. (2017). Learning higher-order generalizations through free play: Evidence from 2- and 3-year-old children. *Developmental Psychology*, *53*(4), 642–651. https://doi.org/10.1037/dev0000278

Slobodskaya, H. R., & Kornienko, O. S. (2021). Age and gender differences in personality traits from early childhood through adolescence. *Journal of Personality*, *89*(5), 933–950. https://doi.org/10.1111/jopy.12624

Smyth, F. L., & Nosek, B. A. (2015). On the gender-science stereotypes held by scientists: Explicit accord with gender-ratios, implicit accord with scientific identity. *Frontiers in Psychology*, *6*, Article 415. https://doi.org/10.3389/fpsyg.2015.00415

Solby, H., Radovanovic, M., & Sommerville, J. A. (2021). A new look at infant problem-solving: Using deeplabcut to investigate exploratory problem-solving approaches. *Frontiers in Psychology*, *12*, Article 705108. https://doi.org/10.3389/fpsyg.2021.705108

Solomon, R. P. (2002). School leaders and antiracism: Overcoming pedagogical and political obstacles. *Journal of School Leadership*, *12*(2), 174–197. https://doi.org/10.1177/105268460201200205

Soto, C. J. (2016). The Little Six personality dimensions from early childhood to early adulthood: Mean-level age and gender differences in parents' reports. *Journal of Personality*, *84*(4), 409–422. https://doi.org/10.1111/jopy.12168

Spivey, L. A., Huebner, D. M., & Diamond, L. M. (2018). Parent responses to childhood gender nonconformity: Effects of parent and child characteristics. *Psychology of Sexual Orientation and Gender Diversity*, *5*(3), 360–370. https://doi.org/10.1037/sgd0000279

Stamarski, C. S., & Son Hing, L. S. (2015). Gender inequalities in the workplace: The effects of organizational structures, processes, practices, and decision makers' sexism. *Frontiers in Psychology*, *6*, Article 1400. https://doi.org/10.3389/fpsyg.2015.01400

Star, J. R., Rittle-Johnson, B., & Durkin, K. (2016). Comparison and explanation of multiple strategies. *Policy Insights From the Behavioral and Brain Sciences*, *3*(2), 151–159. https://doi.org/10.1177/2372732216655543

Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research*, *88*(4), 479–507. https://doi.org/10.3102/0034654317751919

Størksen, I., Ellingsen, I. T., Wanless, S. B., & McClelland, M. M. (2015). The influence of parental socioeconomic background and gender on self-regulation among 5-year-old children in norway. *Early Education and Development*, *26*(5–6), 663–684. https://doi.org/10.1080/10409289.2014.932238

Sutton, R. (2014, March 6). Women everywhere in food empires but no head chefs. *Bloomberg*. https://www.bloomberg.com/news/articles/2014-03-06/women-everywhere-in-chang-colicchio-empires-but-no-head-chefs

Teppers, E., Klimstra, T. A., Van Damme, C., Luyckx, K., Vanhalst, J., & Goossens, L. (2013). Personality traits, loneliness, and attitudes toward aloneness in adolescence. *Journal of Social and Personal Relationships*, *30*(8), 1045–1063. https://doi.org/10.1177/0265407513481445

Tienken, C., & Domenech, D. A. (2021). *The American superintendent 2020 decennial study*. Rowman & Littlefield.

Van den Akker, A. L., Briley, D. A., Grotzinger, A. D., Tackett, J. L., Tucker-Drob, E. M., & Harden, K. P. (2021). Adolescent Big Five personality and pubertal development: Pubertal hormone concentrations and self-reported pubertal status. *Developmental Psychology*, *57*(1), 60–72. https://doi.org/10.1037/dev0001135

Whalen, D. J., Gilbert, K. E., Jackson, J. J., Barch, D. M., & Luby, J. L. (2021). Using a thin slice coding approach to assess preschool personality dimensions. *Journal of Personality Assessment*, *103*(2), 214–223. https://doi.org/10.1080/00223891.2020.1722140

Williamson, D., & Johnston, C. (2015). Gender differences in adults with attention-deficit/hyperactivity disorder: A narrative review. *Clinical Psychology Review*, *40*, 15–27. https://doi.org/10.1016/j.cpr.2015.05.005

Wisdom, T. N., & Goldstone, R. L. (2011). Innovation, imitation, and problem-solving in a networked group. *Nonlinear Dynamics Psychology and Life Sciences*, *15*(2), 229–252. https://pcl.sitehost.iu.edu/rgoldsto/pdfs/innovationimitation.pdf

Woitowich, N. C., Beery, A., & Woodruff, T. (2020). A 10-year follow-up study of sex inclusion in the biological sciences. *eLife*, *9*, Article e56344. https://doi.org/10.7554/eLife.56344

Yang, K., & Girgus, J. S. (2019). Are women more likely than men are to care excessively about maintaining positive social relationships? A meta-

analytic review of the gender difference in sociotropy. *Sex Roles*, *81*(3–4), 157–172. https://doi.org/10.1007/s11199-018-0980-y

Yosso, T. J. (2005). Whose culture has capital? A critical race theory discussion of community cultural wealth. *Race, Ethnicity and Education*, *8*(1), 69–91. https://doi.org/10.1080/1361332052000341006

Zahn-Waxler, C., Cole, P. M., & Barrett, K. C. (1991). Guilt and empathy: Sex differences and implications for the development of depression. In J. Garber & K. A. Dodge (Eds.), *The development of emotion regulation and dysregulation* (pp. 243–272). Cambridge University Press. https://doi.org/10.1017/CBO9780511663963.012

Received January 9, 2023
Revision received June 19, 2024
Accepted June 25, 2024 ■