Learning and Individual Differences 26 (2013) 141–152

Contents lists available at SciVerse ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Can spatial training improve long-term outcomes for gifted STEM undergraduates?

David I. Miller [a,*], Diane F. Halpern [b]

[a] Graduate School of Education, University of California, Berkeley, United States
[b] Department of Psychology, Claremont McKenna College, Claremont, United States

ARTICLE INFO

Article history:
Received 3 May 2011
Received in revised form 12 December 2011
Accepted 17 March 2012

Keywords:
Spatial training
STEM education
Spatial skills
Gender differences
Gifted education

ABSTRACT

This one-year longitudinal study investigated the benefits of spatial training among highly gifted science, technology, engineering and mathematics (STEM) undergraduates (28 female, 49 male). Compared to a randomized control condition, 12 h of spatial training (1) improved the skills to mentally rotate and visualize cross-sections of 3-D objects shortly after training, (2) narrowed gender differences in spatial skills shortly after training, and (3) improved examination scores in introductory physics ($d = .38$) but not for other STEM courses. After eight months, however, there were no training differences for spatial skills, STEM course grades, physics self-efficacy, or declared majors. Large gender differences, favoring males, persisted for some spatial skills, physics grades, and physics self-efficacy eight months after training. These results suggest that sustained exposure to spatially enriching activities over several semesters or years may be necessary to address gender gaps in spatial skills among highly gifted STEM undergraduates.

1. Introduction

Many spatially intensive scientific discoveries such as Kepler's laws of planetary motion or the helical structure of DNA illustrate the importance of spatial skills in science, technology, engineering, and mathematics (STEM) fields. Affirming these anecdotal examples, decades of educational research have demonstrated that measures of spatial skills are robust predictors of students' interest and success in STEM fields (Lohman, 1988; Smith, 1964; Wai, Lubinski, & Benbow, 2009). A recent National Science Board (2010) argued that individuals skilled in spatial thinking are "an untapped pool of talent critical for our highly technological society" (p. 20). However, despite its fundamental value, spatial thinking is "underrecognized, undervalued, underappreciated, and therefore, underinstructed" in traditional education (National Research Council, 2006, p. 14–15). Past research has suggested that students less skilled in spatial thinking may find undergraduate STEM courses to be particularly challenging (for a review see Hegarty, 2010b). Men substantially outperform women on several spatial tasks such as mental rotation (Peters, Manning, & Reimers, 2007; Voyer, Voyer, & Bryden, 1995), suggesting that these learning difficulties may be particularly acute for women (Halpern et al., 2007). Such spatially intensive STEM courses include engineering graphics (Sorby, 2009) or Newtonian physics (Kozhevnikov, Motes, & Hegarty, 2007).

We investigated: can spatial training improve long-term outcomes for gifted STEM undergraduates? The phrase "long-term" here means a time scale of about a year which is long in comparison to most spatial training studies. The phrase "gifted STEM undergraduates" here means students enrolled in a highly selective science and engineering college who were also in the top 3% of academic aptitude (e.g., as measured by SAT scores). This interest in gifted STEM undergraduates is justified since such undergraduates are disproportionately more likely to pursue advanced educational degrees and occupational positions in STEM fields (National Science Board, 2010; see also Reis & Renzulli, 2010 for the importance of gifted education). For instance, Wai et al. (2009) found that 45% of all STEM PhDs in their longitudinal study ($n = 400{,}000$) were within the top 4% of spatial skills in high school. "Spatial skills" here refers to competencies in internally representing and transforming (e.g., rotating) single objects; we found this commonly used definition useful, although perhaps limited. Recent theoretical frameworks have critiqued this narrow within-object definition and expanded it by including the role of between-object skills (Newcombe & Shipley, in press), external visualizations (Hegarty, 2010a) and domain specificity (NRC, 2006). However, these issues are not addressed here. Within the realm of within-object spatial transformational skills, we focus on "mental rotation" and "spatial visualization" as described by Linn and Petersen (1985).

Recent meta-analytic evidence found that spatial experience such as playing action video games (Feng, Spence, & Pratt, 2007), sketching 3-D objects (Sorby, 2009) and even practicing spatial tests (Lohman & Nichols, 1990; Wright, Thompson, Ganis, Newcombe, & Kosslyn, 2008) can robustly improve spatial skills (Uttal et al., in press). Across 217 research studies, Uttal et al. found an average effect size of $d = .62$ for training improvements in spatial skills and suggested that most people's lack of experience with non-navigational spatial tasks such as mental rotation may explain these large improvements. Despite

* Corresponding author at: University of California, Berkeley, 4407 Tolman Hall, Berkeley, CA 94607, United States. Tel.: +1 2064915537.
E-mail address: David.Miller@Berkeley.edu (D.I. Miller).

1041-6080/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.lindif.2012.03.012

these promising findings, most prior research has failed to investigate (1) how spatial training can improve STEM learning outcomes and (2) how long these effects last.

1.1. How can spatial training improve STEM learning outcomes?

With some exceptions, most outcome measures of training studies have been limited to psychometric tests of spatial skills rather than improved STEM course performance. The most compelling evidence that spatial training enhances STEM learning comes from a set of studies with several self-selected cohorts of undergraduate engineering students (Sorby, 2009). Sorby's findings suggested that spatial training increased engineering retention rates for women and improved grades in future STEM courses for both genders (see also Blasko & Holliday-Darr, 2010). However, since Sorby's studies were mostly quasi-experimental, they confounded self-selection effects and probably other factors such as motivation or help-seeking attitudes. Similar differences in GPA and retention rates were found for her randomized studies, although as Sorby (2009) noted, "sample sizes for the randomly selected groups were generally too small to infer statistical significance" (p. 476). A few other spatial training studies have found improved learning outcomes in calculus (Ferrini-Mundy, 1987), chemistry (Small & Morton, 1983; Tuckey, Selvaratnam, & Bradley, 1991), engineering (Alias, Black, & Gray, 2003; Hsi, Linn, & Bell, 1997), geoscience (Piburn et al., 2005), physics (Pallrand & Seeber, 1984) and surgical training (Stransky, Wilcox, & Dubrowski, 2010). However, none of these studies have investigated the benefits of spatial training among gifted STEM undergraduates. For instance, Sorby's (2009) studies excluded students who scored above a low threshold on a spatial test (e.g., 60% on a test of mental rotation). As mentioned previously, we focused on gifted STEM undergraduates since they are disproportionately more likely to become STEM innovators (NSB, 2010). Finding that even extremely gifted STEM students can improve their spatial skills would also support the hypothesis that students of *all* ranges of initial skills can benefit from spatial instruction, thereby substantially extending the applicability of Sorby's promising results.

1.2. How long do training effects last?

Most research has not investigated the longitudinal effects of spatial training although this is critically important for education. Out of the 217 research studies reviewed by Uttal et al., only four studies measured spatial skills more than one month after training (Feng et al., 2007; Hedley, 2008; Pallrand & Seeber, 1984; Terlecki, Newcombe, & Little, 2008). These four studies found large, durable training improvements (average $d = .67$) for two to five months after training. Most encouragingly, Terlecki et al. (2008) found durable *transfer* to untrained spatial tasks despite earlier research suggesting such far transfer can sometimes be rare (Sims & Mayer, 2002). We investigated how this longitudinal research would generalize to gifted STEM undergraduates by measuring spatial skills eight months after training and STEM learning outcomes ten months after training.

1.3. Can spatial training narrow gender differences in spatial skills?

Another open research question regards whether spatial training can narrow or even eliminate the robust gender differences, favoring males, in spatial skills (Baenninger & Newcombe, 1995; Terlecki et al., 2008). As mentioned previously, men often outperform women on many spatial tasks (for an extensive review, see Halpern, 2012). For instance, gender differences in mental rotation emerge as early as 3–4 months of age (Moore & Johnson, 2008; Quinn & Liben, 2008), persist throughout the life span (Peters et al., 2007), have slightly increased during the years 1947–1992 (Voyer et al., 1995) and exist in at least 53 nations (Lippa, Collaer, & Peters, 2010). The consensus across 217 studies indicates that spatial training does not narrow gender gaps in spatial skills in general (Uttal et al., in press). However, Uttal et al. argued it is unclear whether *extensive* instruction can close the spatial gender gap because (1) most studies gave relatively little training and (2) women may show larger gains later in training (see Terlecki et al., 2008). Accordingly, in this study, students completed 12 h of spatial instruction so that we could at least attempt to meaningfully analyze whether training can narrow gender differences in spatial skills.

Given the importance of spatial thinking in STEM fields (National Research Council, 2006), narrowing the gender differences in spatial skills could have societal relevance for increasing the representation of women in STEM fields (Halpern et al., 2007). However, scholars disagree whether these cognitive skills differences play a *primary* role in the underrepresentation of women in STEM fields, especially when compared to other factors such as work–family conflicts or career interests (for an extensive review, see Ceci, Williams, & Barnett, 2009).

1.4. Current study

This study extended previous research by using random assignment to investigate four related research questions: among gifted STEM undergraduates, can 12 h of spatial training (1) improve spatial skills, (2) narrow gender differences in spatial skills, (3) improve STEM learning outcomes, and (4) improve spatial skills and STEM learning outcomes eight to ten months after training? We hypothesized "yes" to all four questions. For average populations, prior research strongly supported hypothesis 1 (Uttal et al., in press), weakly supported hypothesis 2 (Terlecki et al., 2008; Uttal et al., in press), and moderately supported hypothesis 4 (Feng et al., 2007; Hedley, 2008; Pallrand & Seeber, 1984; Terlecki et al., 2008); the longitudinal research for hypothesis 4, however, is limited both in the number of studies ($n = 4$) and retention intervals studied (at most, 2–5 months after training). For more average STEM undergraduates, prior research moderately supported hypotheses 3 and 4 (Sorby, 2009) although such conclusions are tentative because of self-selection effects in some quasi-experimental studies. Also, as noted earlier, it was unclear how previous research would generalize to gifted STEM students, highlighting the need for this study.

2. Method

2.1. Participants

STEM undergraduate majors (28 women, 49 men) were recruited during their first year at a small, highly selective liberal arts college with a strong STEM focus. They represented 37% of their overall first-year class (72 women, 134 men). Notice that the percentage of women for the sample (36.4%) is similar to that of the first-year class (35.0%). Self-reported SAT—Mathematics scores ($M = 761$, $SD = 37$, $n = 67$), SAT—Critical Reading scores ($M = 732$, $SD = 51$, $n = 65$), and SAT—Writing scores ($M = 707$, $SD = 61$, $n = 66$) indicated exceptionally high academic aptitude. All students were either 18 years old ($n = 64$) or 19 years old ($n = 13$) at the time of recruitment/pre-testing. Forty-nine percent of students' mothers and 60% of students' fathers had received an advanced graduate or professional degree. For the overall first-year class, 35% of students were their high schools' valedictorian or salutatorian. Also for the overall first-year class, 61.3% were Caucasian, 17.5% were Asian/Asian-American, 8.5% were multi-ethnic, 5.7% were Hispanic/Latino, .5% were African-American, .5% were Native American, and 6.2% were other or unknown (high school class rank and racial demographics were not available for the subset of first-year students who were participants).

These students demonstrated strong *interests* in pursuing STEM fields by *pre-selecting* into this highly specialized college. This college only offers bachelor's degrees from six STEM departments (biology, chemistry, computer science, engineering, mathematics, and physics) and joint degrees from these different departments. Participants declared majors at the end of their sophomore year: 41.6% declared majors in engineering, 23.4% in computer science, 13% in physics, 9.1% in mathematics, 6.5% in chemistry, and 6.5% in biology. This study's women comprise an atypical population in that they selected into this historically male-dominated college (men typically outnumbered women 2:1).

2.2. Research design

Students were randomly assigned to a training condition (14 women, 25 men) in which they completed six two-hour spatial training sessions distributed over six weeks, or a control condition (14 female, 24 male) that did not participate in any other type of training (e.g., learning vocabulary). Students completed measures of spatial skills prior to training (pretest), one week after the last spatial training session (immediate posttest), and eight months after training (delayed posttest). For completing the spatial pretest, students in both conditions received $5 cash (or $5 gift certificate) and a raffle ticket for six randomly distributed $50 cash prizes; compensation was equivalent for completing the immediate and delayed spatial posttests. Students in the training condition received $90 more to complete the entire spatial training program; students in control condition did not receive that additional $90 but received compensation for completing the spatial tests. All students assigned to the training condition completed all spatial sessions and immediate posttest. One male student in the control condition did not complete the immediate posttest. A longitudinal subsample of students ($n = 55$ overall) in the training condition (12 women, 18 men) and control condition (10 women, 15 men) completed the delayed spatial posttests eight months after training. This longitudinal subsample represented 71% of the original pre-test sample ($n = 77$) and missing data analyses indicated that retention rates did not significantly differ in terms of experimental assignment ($\chi^2(1) = 1.17$, $p = .280$), gender ($\chi^2(1) = 1.10$, $p = .294$), initial spatial skills (all Fs<1), or SAT scores (all Fs<1).

We also obtained students' STEM course grades up to ten months after the last spatial training session. These grades were available for the entire pretest sample ($n = 77$). Hence, we analyzed training differences in STEM course grades for the entire pretest sample even though we did not have delayed posttest data for all students. With a fixed significance level of .05, power analyses showed that 38 students per condition yield a statistical power of 80% for detecting an effect size of $d = .58$ for a one-tailed independent samples t-test.

2.3. Instructional materials

Sorby and Wysocki (2003) published workbook exercises and multimedia software for nine modules on specific spatial topics. Sorby (2009) summarized these materials' development and application. This study used six training modules: Isometric Drawings and Coded Plans, Orthographic Drawings, Rotation of Objects about a Single Axis, Rotation of Objects about Two or More Axes, Object Reflections and Symmetry, and Cutting Planes and Cross Sections (see Section 2.5 for description of what spatial skills these modules cover). The workbook exercises typically required students to visualize spatial transformations such as rotations of abstract 3-D objects and then sketch the transformed 3-D objects with paper and pencil (see Fig. 1). In contrast to simply practicing multiple-choice spatial tests (e.g., Wright et al., 2008), this study required students to *generate* as opposed to *select* correct spatial transformations. In a randomized controlled study, Sorby (2009) found the sketching exercises improved spatial skills substantially more than the 3-D computer activities which only required watching or selecting correct spatial transformations. These results mirrored other studies finding that generation activities, as opposed to passive instruction or selection activities, can introduce desirable difficulties that can improve learning in skill acquisition (Bjork, 1999), verbal declarative memory (Kornell, Hays, & Bjork, 2009; Richland, Linn, & Bjork, 2007), and even conceptual science knowledge (Linn, Chang, Chiu, Zhang, & McElhaney, 2010). At multiple universities, Sorby (2009) and Veurink et al. (2009) found large within-subjects improvements in spatial skills with these materials.

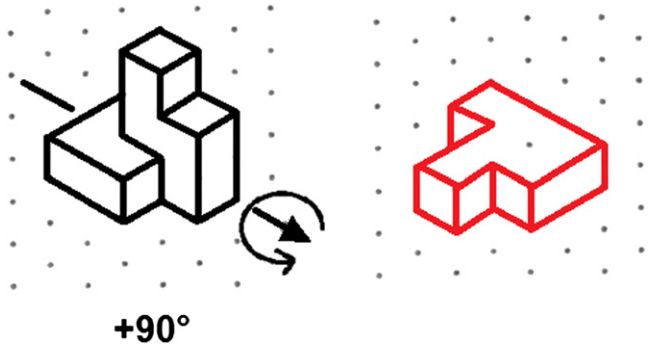

Fig. 1. Sample workbook problem from the spatial training. On 2-D sketch paper, students are asked to mentally rotate the left 3-D object 90° around the indicated axis and then sketch the correct rotation (shown in red) on the dot paper to the right. Adapted with permission.
From Sorby and Wysocki (2003).

2.4. Procedure

Students completed spatial training at one of the college's computer laboratories outside of normal class time. They were divided into four sections that each had a different instructor.

Each training session began with a ~15-minute PowerPoint lecture/demonstration. Veurink et al. (2009) used the same lecture slides. Instructors presented different spatial strategies for approaching the workbook problems. Example strategies include using physical hand gestures to determine the direction of rotational motion or applying a step-by-step approach for translating orthographic projections (2-D representations) into isometric sketches (3-D representations). Prior to each session, instructors reviewed together the lecture content to ensure consistent instruction. At the end of the lecture, students attempted a sample sketching exercise and then watched the instructor solve that problem. Students then interacted with the software materials which reinforced these spatial strategies and illustrated various spatial transformations.

After the software materials, students spent most of the session completing the sketching workbook problems. To help complete the workbook problems, students were given fifteen physical snap-cube blocks at the beginning of each session. These blocks allowed students to physically build the three-dimensional structures that they were asked to sketch in 2-D form. Hence, the physical cube blocks served as scaffolds for externally visualizing spatial transformations such as mental rotation. During the following week, instructors graded and returned workbook pages. Instructors encouraged students to review problems that students got incorrect or did not finish; however, students were not obligated to do so.

During the last 10 min of every session, students completed online surveys that asked about training module's difficulty level, quality, length, etc. To encourage transfer, the surveys also asked students to identify three specific topics in their current STEM courses relevant to the module's spatial lessons. At the start of next week's session, another survey asked students to recall those specific course topics

and indicate how successful students were in applying those spatial lessons.

2.5. Measures

We distinguish between wave 1 data which were collected at any point before training and up to three months after training, and wave 2 data which were any data collected between eight to ten months after training. The wave 1 and wave 2 spatial tests were not the same because of ceiling effects found after analyzing wave 1 tests. Hence, since the analysis of wave 1 data informed the selection of wave 2 measures, we first present the analysis of wave 1 data in Section 3 before describing wave 2 measures in Section 4. Wave 1 data included (1) four measures of spatial skills at pre-testing and immediate post-testing, (2) online evaluation surveys collected during training, and (3) grades for STEM courses taken concurrently with training. We choose the spatial measures because they test skills covered in instruction and also, in part, because they matched tests used in Veurink et al. (2009)'s study of these training materials' efficacy.

2.5.1. Mental Cutting Test

The Mental Cutting Test (MCT) measured students' skills to visualize cross-sections of 3-D objects cut by a specified 2-D plane (CEEB, 1939). The MCT measured skills taught on the training Module 7 (Cutting Planes and Cross Sections). The original scale consisted of 25 items; this study used the same subset of 10 items that were by Veurink et al. (2009). Students were given 8 min to complete these 10 problems. Internal consistency was acceptable at pre-testing ($\alpha = .74$) and immediate post-testing ($\alpha = .72$).

2.5.2. Mental Rotation Test

The Mental Rotation Test (MRT) measured students' skills to mentally rotate 3-D objects (Peters et al., 1995; Vandenberg & Kuse, 1978). This test required students to select two correct rotations (given four possibilities) of a specified 3-D object (see Fig. 2). The MRT tests skills covered on the training Module 4 (Rotation of Objects about a Single Axis) and Module 5 (Rotation of Objects about Two or More Axes). For each item, students received two points for identifying both correct rotations and 1 point for identifying one correct rotation. Students were given 20 min to complete 24 items of Form A described by Peters et al. (1995). Internal consistency was excellent at pre-testing ($\alpha = .84$) and acceptable at post-testing ($\alpha = .67$). The timing limit (20 min) differs substantially from the limit originally specified (6 min) by Peters et al. (1995). Hence, the test was essentially untimed (95% of students completed at least 21 of the 24 MRT problems at pre-testing) and therefore measured student's accuracy, not speed. We adjusted the timing for two reasons: (1) to ensure that any possible measured gender differences are caused by accuracy not time pressure (see Voyer, 2011 for a meta-analysis of this issue), and (2) to cross-validate the test with STEM course examinations where accuracy, not speed, is generally the central factor (e.g. this college's course examinations are typically designed to give students ample time).

2.5.3. Lappan Test

The Lappan Test (Veurink et al., 2009) measured students' skills to visualize different views of 3-D objects using both isometric sketches (3-D representations) and orthographic views (2-D representations). The Lappan test measured skills taught on the training Module 1 (Isometric Drawings and Coded Plans) and Module 2 (Orthographic Drawings). This study used the same subset of 10 items that were used by Veurink et al. (2009). Students were given 8 min to complete 10 problems. Internal consistency was somewhat poor at pre-testing ($\alpha = .62$) and poor at post-testing ($\alpha = .48$). Hence, this measure's scores should be interpreted cautiously.

2.5.4. Paper Folding Test

The Paper Folding Test (PFT) showed students a folded and hole-punched 2-D piece of paper and asked students to identify how the paper would look when unfolded. The PFT measured skills taught on the training Module 6 (Object Reflections and Symmetry). The test consisted of 20 problems that were divided into two ten-problem sections, each section lasting 3 min. Consistent with the test instructions specified on the PFT, every incorrect answer penalized students' scores by one-fourth of a point. Internal consistency was acceptable at pre-testing ($\alpha = .72$) and at post-testing ($\alpha = .70$).

2.5.5. Online surveys

Students answered questions about each training module's length, enjoyabilitiy, quality, etc. at the end of every training session. We report two types of surveys responses: (1) ratings of which training components (lecture/software, workbook exercises, and 3-D cube blocks) were most helpful to learning and (2) ratings of the module's difficulty. The survey was adapted from Veurink et al. (2009). We also report on survey data on how successful students were in applying last week's spatial lessons into their current STEM courses.

2.5.6. Concurrent STEM course grades

Final GPA grades were available for courses that most students took concurrently with spatial training during the second semester of their first year. These courses were introductory calculus-based Newtonian physics ($n = 62$), introductory engineering design ($n = 37$), introductory chemistry ($n = 77$), differential equations I ($n = 75$), and multi-variable calculus I ($n = 70$). Grades were converted to numerical scores by assigning "A" = 4.0, "A−" = 3.667, and so on. Students completed training during the first six weeks of the four-month semester. Hence, final grades reflect student work completed during training and up to 2.5 months after training.

2.5.7. Physics-specific learning outcomes

For all first-year students (including students not enrolled in this study), we had access to scores on all physics examinations/quizzes taken throughout the semester. We created a composite examination score based on scores from the final (31.25%), two midterms (25% each), and three quizzes (6.25% each); the physics course syllabus specified these weightings. Examination scores contributed 80% of students' final physics GPA and homework contributed 20%. Hence, composite examination scores were nearly equivalent to final physics

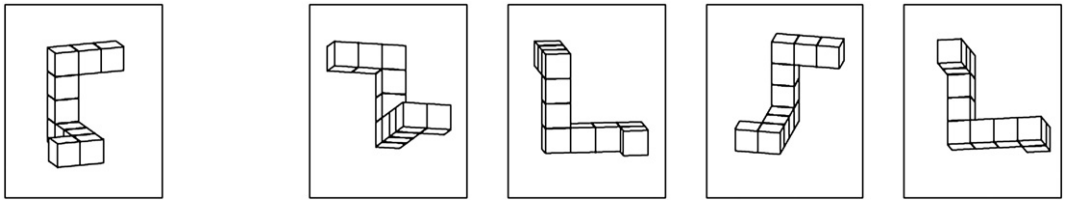

Fig. 2. Sample problem from the Mental Rotation Test. Students are asked to identify which objects on the right hand side can be rotated to match the far left object. Correct answers are the second and third figures on the right hand side.
From Peters et al. (1995). Reprinted with permission.

grades ($r(60) = .98$), but the composite score was a finer-grained continuous measure of course performance compared to final grades (which are graduated by discrete values of "A", "A−", etc.). Directly prior to training and two months after training, students also completed the Force Concept Inventory—a widely used test of conceptual understanding of Newtonian mechanics in introductory physics (Hestenes, Wells, & Swackhamer, 1992). This test did not contribute to the students' final GPA or the composite examination score. The Force Concept Inventory and physics course examinations require substantially different processes for applying physics knowledge. For instance, the course examination questions required extensive application of mathematical problem solving skills including calculus. In contrast, the conceptual questions assessed students' qualitative, not mathematical, understanding of physics principles. With regards to spatial thinking, the most important differences between the two physics assessments were: the course examination problems often provided no visual–spatial diagrams of the physical situations (students generated their own diagrams or attempted problems without such diagrams) and typically involved more complex motion (including rotational and 3-D motion).

3. Results

3.1. Descriptive statistics and pretest gender differences

Table 1 contains descriptive statistics, intercorrelations, and pretest gender differences for all spatial measures and SAT scores. Note that mean scores on most spatial measures were within one standard deviation of the maximum score (100), demonstrating ceiling effects. In terms of the number of standard deviations from ceiling performance, the Mental Rotation Test and Lappan Test demonstrated particularly large pretest ceiling effects, suggesting less opportunity for improvement on these specific measures. These ceiling effects became larger at post-testing (see Table 1 above the diagonal). All posttest spatial measures exceeded skewness magnitudes of .8. These ceiling effects may have explained the somewhat low, but acceptable, internal consistencies of this study's spatial measures, noted in Section 2.5.

Table 1 also shows that, at pretesting, males substantially outperformed females on the Mental Cutting Test, Mental Rotation Test, Lappan Test, and SAT—Mathematics ($ds = .60 \ldots 1.04$) and no gender differences on the Paper Folding Test, SAT—Critical Reading, and SAT—Writing. Past research aligns with the null gender differences on the Paper Folding Test (Voyer et al., 1995) and aligns with the large differences on the other spatial measures (Sorby, 2009; Veurink et al., 2009; Voyer et al., 1995). This study's SAT—Mathematics gender difference ($d = .60$) exceeds the magnitude for average populations (The College Board, 2010, $d = .29$) but matches the magnitude for gifted populations (e.g., Casey, Nuttall, Pezaris, & Benbow, 1995, $d = .70$). Hence, we found large gender differences where we would expect them.

Table 1
Intercorrelations, descriptive statistics, and pretest gender differences for spatial measures and SAT scores.

Measure	1	2	3	4	5	6	7	*M*	*SD*
1. MCT	–	.36**	.24*	.27**	.25*	.11	−.01	83.7	18.8
2. MRT	.52**	–	.36**	.34**	.36**	.12	−.13	97.3	3.8
3. Lappan	.37**	.44**	–	.42**	.43**	.30**	.09	92.1	11.1
4. PFT	.44**	.52**	.18**	–	.18	.19	−.04	88.7	11.9
5. SAT-M	.37**	.43**	.37*	.33**	–	.16	.20	761	37
6. SAT-CR	.21	.31**	.22	.21**	.16**	–	.40**	732	51
7. SAT-W	.20	.28*	.04	.12**	.20**	.40**	–	707	61
M	70.8	94.7	86.2	82.9	761	732	707		
SD	23.8	8.1	16.0	14.0	37	51	61		
d (gender)	1.04**	.60*	.77**	.01	.60*	.28	.00		

Note. Pretest data are presented below the diagonal, and posttest data are presented above the diagonal. All spatial scores have been normalized to a maximum score of 100. The row "*d* (gender)" presents pretest gender differences (Cohen's *d*); positive values indicate an advantage for males.
MCT = Mental Cutting Test, MRT = Mental Rotation Test, PFT = Paper Folding Test, SAT-M = SAT—Mathematics, SAT-CR = SAT—Critical Reading, SAT-W = SAT—Writing.
*$p < .05$ (one-tailed), **$p < .01$ (one-tailed).

3.2. Online surveys: Learning components and module difficulty

To characterize students' training experiences, Table 2 presents students' self-reported ratings of (1) the relative benefits of the lectures/software, workbook problems, and physical cube blocks and (2) the training modules' difficulties. For most modules, students found the workbook problems to be most beneficial, aligning with past research that highlights the importance of this workbook/sketching component (Sorby, 2009). Use of the physical cube blocks varied greatly between modules, and these blocks were most helpful for the mental rotation modules. Students found most training modules to be challenging, particularly the modules on mental rotation and orthographic projections (scores higher than "3" indicates students judged the module to be "somewhat advanced" or "too advanced"). This result was particularly surprising considering these materials were developed for students with large deficits in spatial skills, not extremely gifted STEM undergraduates (Sorby, 2009). The modules on isometric drawings and reflections/symmetry were somewhat less challenging which would predict smaller improvements on the Paper Folding Test.

3.3. Immediate posttest improvements of spatial skills

We first analyzed spatial pretest/posttest data with a repeated measures multiple analysis of variance (MANOVA) with two between-subjects factors (experimental assignment and gender) and one within-subjects effect (Time [pretest, posttest]). Since we found large ceiling effects on our spatial measures, we also analyzed data with three different data transformations (reflect and inverse, reflect and square root, reflect and logarithm) and, in general, these alternative statistical procedures confirmed the results found with untransformed data.

Table 3 shows descriptive statistics for the spatial measures by time of testing, gender, and experimental assignment. The main effect of time ($F(1, 72) = 79.78$, $p < .001$, $\eta_p^2 = .53$), and the interactions for Time × Assignment ($F(1, 72) = 10.16$, $p = .001$, $\eta_p^2 = .12$), Time × Gender ($F(1, 72) = 6.11$, $p = .008$, $\eta_p^2 = .08$), and Time × Gender × Assignment ($F(1, 72) = 5.95$, $p = .008$, $\eta_p^2 = .08$) were statistically significant. The main effect of time suggested improvements in spatial skills due to perhaps concurrent STEM course enrollment or practice

Table 2
Summary of module evaluation survey responses: learning components and difficulty.

Module	Learning components			Mean difficulty
	Lecture/ software	Workbook	Physical blocks	
Isometric drawings	3.60 (.83)	4.53 (.56)	3.48 (1.03)	2.51 (.76)[a]
Orthographic drawing	3.62 (.98)	4.45 (.69)	2.82 (1.07)	3.21 (.66)[b]
Single axis rotation	3.72 (.82)	4.63 (.49)	4.35 (.80)	3.15 (.49)[b]
Multiple axis rotation	3.64 (1.00)	4.49 (.51)	4.77 (.43)	3.40 (.55)[a]
Reflections and symmetry	3.61 (.79)	4.39 (.60)	2.93 (.92)	2.56 (.65)[a]
Cross sections	3.99 (.93)	3.89 (.83)	[c]	3.00 (.90)

Note. Entries represent mean (standard deviation). The "learning components" columns summarize responses to the questions: "The _____ were/was beneficial to understanding this module's material: strongly disagree (1), …, strongly agree (5)." The "mean difficulty" column are ratings of the modules' difficulties (1 = "too simple", 3 = "appropriate", 5 = "too advanced").
[a] Mean difficulty was significantly different from 3 = "appropriate" ($p < .05$).
[b] $p < .10$.
[c] Physical blocks were irrelevant to this module.

effects of taking multiple spatial tests; the Time × Assignment interactions indicated training-specific improvements in spatial skills. To investigate which specific spatial skills were most responsive to training, we conducted four separate ANOVAs for each spatial measure. The main effect of time was significant for all four tests (see Table 3). Table 3 also shows that for the Mental Cutting Test, all interactions with time were significant. Note that for time by group interactions (such as Time × Assignment), the Cohen's *d* is calculated as the effect size difference in *gain scores* (posttest score minus pretest scores) between two groups (e.g., training and control groups). A positive effect size indicates either greater gain scores for training group (Time × Assignment) or for women (Time × Gender). For the Mental Rotation Test, the interactions Time × Assignment and Time × Gender were significant although the Time × Gender × Assignment interaction was not. For the Lappan test, only the interaction of Time × Gender was significant. For the Paper Folding Test, no interactions were significant. Hence, students in both conditions improved over time on all four spatial skills measures, students in the training group improved more on the Mental Cutting and Mental Rotation measures, and gender differences were narrowed at posttesting for the Mental Cutting, Mental Rotation, and Lappan tests.

3.4. Predictions about STEM course grade improvements

On the "Application to Classes" surveys, students often found the training to be somewhat abstract and expressed difficulty in connecting the training to their current courses in just one week's time. Despite these difficulties, students found meaningful, but infrequent, connections between their STEM courses and spatial training. Students selected the course they thought was most applicable to spatial training; the two most frequently selected courses were engineering (44%) and physics (30%), followed by chemistry (8%), differential equations (4%), and other/not specified (14%). Note that multivariable calculus I occurred during the second half of the semester, after spatial training, and therefore was not included in this list. These survey responses suggest the most benefits for students' engineering and physics courses.

We also correlated initial spatial skills with course achievement. We created a composite spatial skills measure by summing each pre-test's standardized *z*-scores. Because of its low reliability, the Lappan test was not included in this composite measure. The correlations between initial spatial skills and GPA were: physics ($r(60) = .51$, $p < .001$), differential equations I ($r(73) = .44$, $p < .001$), multi-variable calculus I ($r(68) = .31$, $p < .01$), chemistry ($r(75) = .27$, $p < .01$), engineering ($r(35) = -.03$, n.s.). After controlling for SAT scores, only the correlations with physics grades and multivariable calculus I remained significant (both $p < .05$). From this survey and correlational data, the only consistent prediction is improved *physics* grades.

3.5. STEM course grades improvements

We first analyzed STEM course grades for gender differences. We found that compared to women, men achieved higher GPAs for Newtonian physics ($p = .008$, $d = .73$), multivariable calculus I ($p = .04$, $d = .53$), and introductory chemistry ($p = .04$, $d = .49$). Men also achieved higher GPAs in differential equations ($p = .09$, $d = .42$) and women achieved higher GPAs in introductory engineering design ($p = .179$, $d = -.47$), although these two differences were not significant. To investigate the effects of spatial training on STEM course performance, we analyzed STEM course grades with an Assignment × Gender analysis of covariance (ANCOVA) using composite initial spatial skills as the covariate. ANCOVA results indicated that training group outperformed the control group for final Newtonian physics grades ($F(1, 57) = 4.06$, $p = .024$, $d = .32$). For all other courses, the main effects of assignment were not significant (all $F < 1$, all $p > .2$, all $|d| < .2$). For all courses, the interaction of Assignment × Gender was not significant (all $F < 1$, all $p > .2$).

3.6. Physics-specific learning outcomes

To investigate physics improvements further, we analyzed scores on individual physics examinations throughout the semester for all first-year students (see Section 2.5.7 for further description). Students not enrolled in this study (but enrolled in the Newtonian physics course) served as a second non-randomized comparison group, designated as the "rest of the class group"; see Fig. 3. The training group consistently outperformed both comparison groups on physics examinations throughout the semester, which includes the physics final which was 2.5 months after the last spatial training session. As discussed in Section 2.5.7, composite examination scores are nearly equivalent to final grades. Hence, as expected, the training group outperformed the control group on composite physics

Table 3
Descriptive statistics and effect sizes for spatial pretest to posttest differences.

			Descriptive statistics		Effect sizes			
Test	Assign (A)	Gender (G)	Pre *M* (*SD*)	Post *M* (*SD*)	Time (*d*)	Time × A (*d*)	Time × G (*d*)	Time × A × G (η_p^2)
Mental Cutting Test	Control	Female	61.4 (25.4)	65.1 (23.5)	.63**	.79**	.46*	.10**
		Male	77.4 (19.1)	83.9 (17.3)				
	Training	Female	52.1 (26.4)	87.2 (12.1)				
		Male	79.2 (18.0)	92.1 (13.2)				
Mental Rotation Test	Control	Female	93.5 (7.0)	95.8 (4.7)	.43**	.35*	.51*	.02
		Male	96.5 (6.0)	97.6 (2.4)				
	Training	Female	90.0 (11.3)	97.0 (5.0)				
		Male	96.2 (7.8)	98.2 (3.5)				
Lappan	Control	Female	80.7 (18.2)	87.9 (15.3)	.44**	−.02	.44*	.03
		Male	87.4 (14.2)	93.0 (10.6)				
	Training	Female	77.1 (22.0)	90.7 (12.1)				
		Male	92.8 (7.9)	94.4 (7.7)				
Paper Folding Test	Control	Female	83.4 (12.8)	89.8 (8.4)	.46**	−.04	−.28	.02
		Male	83.4 (12.6)	89.4 (10.4)				
	Training	Female	82.1 (14.8)	84.1 (16.7)				
		Male	82.1 (15.8)	90.0 (12.0)				

Note. All effect sizes and significance values are based on Time × Assignment × Gender repeated-measures ANOVAs. Positive *d* effect sizes indicate either greater gain scores for the training group (Time × Assignment) or for women (Time × Gender). A = Assignment, G = Gender.
*$p < .05$. **$p < .01$.

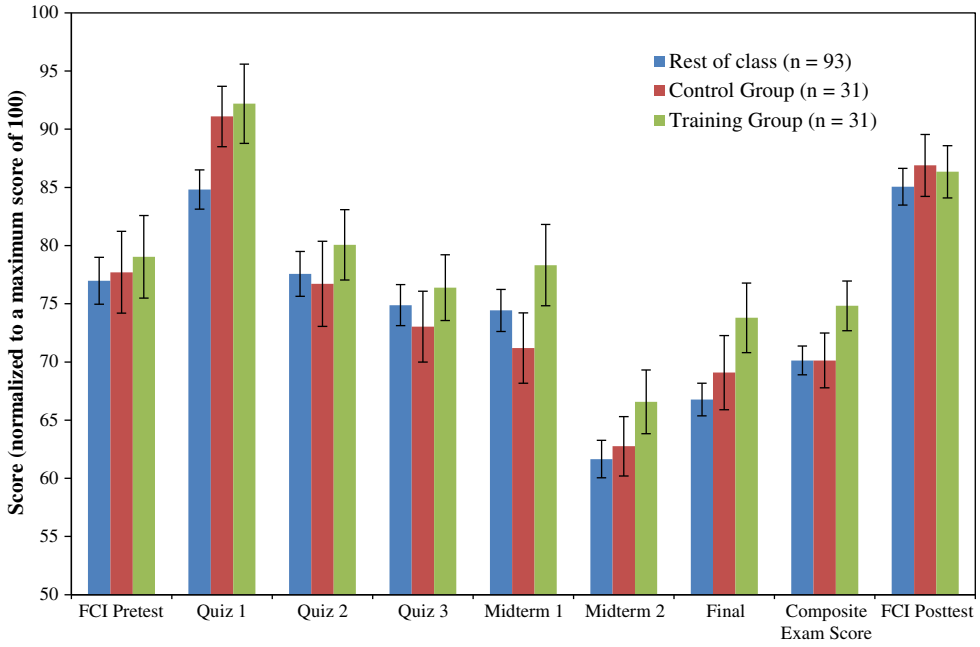

Fig. 3. Scores on Newtonian physics examinations throughout the first year, second semester. FCI = Force Concept Inventory. The FCI pretest was completed before training, Quiz 1 and Midterm 1 were completed during training, and all other exams were completed after training.

examination scores ($F(1, 57) = 6.14$, $p = .008$, $d = .38$) like before. Men outperformed women on this physics composite score ($p = .004$, $d = .80$) and but the Assignment × Gender interaction was not significant ($F < 1$).

Despite the training improvements on examination scores, training and control groups did not differ on the Force Concept Inventory posttest which measured physics conceptual understanding ($F < 1$, $d = -.02$). Men substantially outperformed women on the Force Concept Inventory pretest ($p < .001$, $d = 1.45$) and posttest ($p < .001$, $d = 1.12$). Despite the lack of training differences on physics conceptual understanding, initial spatial skills predicted both Force Concept Inventory pre-test ($r(55) = .45$, $p < .001$) and post-test ($r(55) = .52$, $p < .001$) scores, even with SAT scores partialled out (both $p < .01$).

4. Longitudinal follow-up measures

We followed up on wave 1 data by measuring students' spatial skills eight months after the last training session and collecting STEM course grades ten months after training. Longitudinal spatial skills data were available for a longitudinal subsample ($n = 55$) and STEM course grades were available for the entire pretest sample ($n = 77$); see Section 2.1. We describe how the analysis of wave 1 data informed the selection of each wave 2 measure.

4.1. Mental Cutting Test

Since we found larger improvements for the training group on the Mental Cutting Test (CEEB, 1939), we included this measure again for wave 2 data. For this longitudinal assessment, internal consistency was somewhat poor ($\alpha = .57$).

4.2. Novel Cross-Sections Test

Also since we found larger improvements on the Mental Cutting Test, we also used the Novel Cross-Sections Test (Hegarty, Keehner, Khooshabeh, & Montello, 2009) which also measured students' skills to visualize cross-sections of 3-D objects cut by a specified 2-D plane. We included the Novel Cross-Sections Test to test whether spatial training improved the *construct* of mental cutting, not just test performance on one specific test. Students were given 5 min to complete 10 problems. Internal consistency was somewhat poor ($\alpha = .61$).

4.3. Mental Rotation Test

Since we found training improvements on the Mental Rotation Test, we included this measure again for wave 2. Since we found large ceiling effects at wave 1, we changed to Form C which was designed to be more challenging than Form A (see Peters et al., 1995). To also help avoid ceiling performance, we considered decreasing timing limits. However, changing timing limits could have affected gender differences (see Voyer, 2011 for a meta-analysis) or training differences (e.g., if training improved students' accuracy, not speed, in solving mental rotation problems). As a compromise, students completed 12 problems with 10 min time limits (consistent with wave 1 timing limits) and 12 problems completed with 3 min time limits (consistent with other researchers). Item content was randomly counterbalanced between students, and students always completed problems first under the 10 min time limits. As expected, scores were higher under 10 min time limits compared to 3 min time limits ($p < .001$, $d = 2.39$). However, since the timing condition neither interacted with gender ($F(1, 53) = .02$, n.s.) nor assignment ($F(1, 53) = .75$, $p = .39$), we combined scores across all problems. Internal consistency was good ($\alpha = .81$) for this composite scale.

4.4. Spatial working memory

Because some researchers have suggested that spatial working memory may explain the mechanism for improved spatial skills (Chein & Morrison, 2010) and may underlie the connection between spatial skills and physics learning (Kozhevnikov et al., 2007), we included a measure of spatial working memory . This study's spatial working memory test (Kane et al., 2004, rotation span; adapted from Shah & Miyake, 1996) measured students' capacity to simultaneously process and store novel spatial information. On computers, students judged whether a set of individually presented letters were

normal or mirror-imaged while simultaneously remembering the locations of a sequence of short and long arrows radiating from the center of a computer screen. At the end of a trial, the students recalled the positions of the arrows in the order they were presented. Set sizes ranged from two to six letter-arrow displays per trial (with 3 trials per set size for 15 trials total). We scored the recall data using the partial credit procedure advocated by Conway et al. (2005). Internal consistency was good ($\alpha = .78$).

4.5. Sophomore STEM course grades

Final GPA grades were available for courses that most students took during their first semester of sophomore year (one semester after training). These courses were differential equations II ($n = 76$), introductory calculus-based electricity & magnetism ($n = 75$), introductory statistics ($n = 55$), introductory engineering systems ($n = 57$), and introductory biology ($n = 45$). Final sophomore semester grades reflect student work completed six to ten months after training.

4.6. Physics problem-solving self-efficacy

Improved self-efficacy could have also moderated or mediated the improvements found in physics course performance (Bandura, 1997; Cooke-Simpson & Voyer, 2007). Hence, we measured student's self-efficacy for solving physics problems with three Likert scales. Two scales asked for students' strength of agreement with statements such as, "If I get stuck on a physics problem on my first try, I usually try to figure out a different way that works" (Adams, Perkins, Dubson, Finkelstein, & Wieman, 2006, Solving Confidence Subscale, 4 items; Çalişkan, Selçuk, & Erol, 2007, Solving Physics Problems Subscale, 10 items). Using recommendations by Bandura (1997), we constructed a third scale by presenting students with four math-intensive physics problems (similar to the ones found on their Newtonian physics examinations) and then asking students to rate on a 1–10 scale their confidence in correctly solving such problems during a physics examination. All scales showed good internal consistency (αs = .74 to .93), highly correlated with one another (rs = .64 to .87) demonstrating convergent validity, and highly correlated with the wave 1 physics outcomes described in Section 2.5.7 (rs = .49 to .71) demonstrating criterion validity. Because of these solid psychometric properties, we computed a composite scale by summing each scale's standardized z-scores.

5. Longitudinal follow-up results

5.1. Descriptive statistics and gender differences

Table 4 contains intercorrelations and descriptive statistics for all four longitudinal spatial measures. Notice that the two mental cutting measures only modestly correlated with one another ($r = .30$) although they aim to measure the same construct, suggesting concerns about construct validity. The low internal consistencies of these scales may also explain the low bivariate correlation (Keppel, 1991). Either way, we interpret results with those two mental cutting measures cautiously. Regarding gender differences, men outperformed women on the Mental Rotation and Mental Cutting tests but not on the Novel Cross-Sections or Spatial Working Memory tests (see Table 4). For sophomore course grades, men outperformed women on introductory electricity and magnetism grades ($p < .001$, $d = .97$) although not on differential equations II ($p = .383$, $d = .21$), introductory statistics ($p = .287$, $d = -.30$), introductory engineering systems ($p = .298$, $d = .29$), or introductory biology grades ($p = .673$, $d = .05$). Men also had higher physics self-efficacy compared to women ($p = .022$, $d = .70$).

Table 4
Summary of intercorrelations and descriptive statistics for spatial longitudinal measures.

Measure	1	2	3	4
1. Mental cutting	–			
2. Novel cross-sections	.30*	–		
3. Mental rotation	.41**	.41**	–	
4. Spatial working memory	.10	.44**	.48***	–
M	86.0	70.2	72.1	57.7
SD	15.3	21.3	13.0	13.9
d (gender)	.82**	.34	.71*	.12

Note. All scores have been normalized to a maximum score of 100. The row "*d* (gender)" presents gender differences (Cohen's *d*); positive values indicate an advantage for males.
*$p < .05$. **$p < .01$. ***$p < .001$.

5.2. Training differences in spatial skills after eight months

We analyzed the four longitudinal spatial measures with an Assignment × Gender between-subjects multiple analysis of covariance (MANCOVA) using composite initial spatial skills as the covariate. Results indicated no main effect of assignment ($F(4, 47) = 1.07$, $p = .192$) and no interaction Assignment × Gender ($F(4, 47) = .78$, $p = .272$). For individual measures, the effect sizes for training differences were generally small: Mental Rotation Test ($d = -.09$), Mental Cutting Test ($d = .08$), Novel Cross-Sections Test ($d = .05$), and Spatial Working Memory test ($d = -.37$). Thus, training and control groups did not differ in spatial skills eight months after training.

5.3. Correlations between initial spatial skills and STEM course performance

The correlations between composite initial spatial skills (see Section 3.4) and sophomore course GPA were: electricity and magnetism ($r(73) = .38$, $p < .001$), biology ($r(43) = .30$, $p = .024$), engineering systems ($r(68) = .23$, $p = .045$), differential equations II ($r(74) = .22$, $p = .026$), and statistics ($r(53) = .02$, n.s.). After controlling for SAT scores, only correlations with electricity and magnetism grades (r-partial $(55) = .32$, $p = .012$) remained significant (all other $p > .20$). Hence, like before (Section 3.4), we would predict only improvements in physics grades.

5.4. Results for sophomore STEM Courses, physics self-efficacy, and declared major

We analyzed STEM course grades with an Assignment × Gender analysis of covariance (ANCOVA) using composite initial spatial skills as the covariate. Except for biology grades, neither the main effects of assignment (all $F < 1$, all $p > .2$, all $|d| < .22$) nor interaction of Assignment × Gender was significant (all $F < 2.2$, all $p > .1$). However, for biology grades, the training group outperformed the control group ($F(1, 40) = 5.84$, $p = .01$, $\eta_p^2 = .13$, $d = .45$) and these training improvements were larger for women compared to men ($F(1, 40) = 4.13$, $p = .049$, $\eta_p^2 = .09$). Given the inflation of type I error because of multiple tests, we cannot make strong conclusions from this biology finding especially considering we found no training differences on four spatial tests or four other course grades. We also have no a priori reason to expect improvements in biology (e.g., biology grades did not correlate with initial spatial skills after controlling for SAT scores). Finally, we found no main effect of assignment or interaction of Assignment × Gender for physics self-efficacy (Fs < 1) and there were no significant differences in the majors that students declared at the end of the sophomore year ($\chi^2(5) = 3.46$, $p = .629$).

5.5. Does spatial working memory correlate with physics success?

Because some researchers have suggested that spatial working memory may underlie the connection between spatial skills and physics learning (Kozhevnikov et al., 2007), we correlated physics outcomes with spatial working memory. Contrary to this hypothesis, spatial working memory correlated with none of the physics outcomes from either the first-year or sophomore year (all $ps > .10$). These physics outcomes were: Force Concept Inventory pretest ($r = .16$), Force Concept Inventory posttest ($r = .02$), first-year Newtonian examination scores ($r = .17$), physics self-efficacy ($r = .11$), and sophomore electricity & magnetism grades ($r = .05$).

6. Discussion

This study investigated the benefits of 12 h of spatial training among highly gifted STEM undergraduates. From these investigations, four major findings emerged. First, the training group improved more over time in the skills needed to mentally rotate and visualize cross-sections of 3-D objects. Second, spatial training narrowed gender differences in spatial skills at the immediate posttest perhaps because of ceiling effects. Third, the training group consistently outperformed the control group on examination scores for introductory Newtonian physics ($d = .38$) but not for other STEM courses; these physics improvements were evident for 2.5 months after training. Fourth, after 8–10 months, training differences did not exist for spatial skills, STEM course grades, physics self-efficacy, or declared majors. We now discuss the limitations and implications for each of these research findings.

6.1. Spatial training improves spatial skills

First, how do the substantial ceiling effects found on the spatial pretest and immediate posttest affect conclusions about immediate spatial training improvements? These ceiling effects *strengthen* such conclusions because ceiling effects tend to mask authentic group differences. These ceiling effects could also help explain why other research has found training-related improvements on the Paper Folding test (Wright et al., 2008) and the Lappan test (Sorby, 2009), but we did not. For those Paper Folding and Lappan results, an alternative hypothesis is: students with more average spatial skills may have benefited more from the same amount of spatial experience. However, in some opposition to this hypothesis, participants reported in the online surveys that the sketching exercises were challenging and appropriate to their learning needs. Students reported to the first author that the pre- and post-tests were comparatively easy; systematic ceiling effects confirmed those anecdotal reports. These results highlight a critical distinction between *generating* versus *selecting* correct mental transformations. Sketching requires students to generate mental transformations, rather than rely upon the generous scaffolding provided by multiple choice answers (Strasser et al., 2010). More challenging spatial tests, especially those involving sketching, may have revealed greater training differences.

6.2. Spatial training appears to narrow gender differences in spatial skills

Second, how do the ceiling effects affect conclusions about training narrowing gender differences in spatial skills immediately after instruction? These ceiling effects *weaken* claims about narrowed gender differences for the same reason: ceiling effects tend to mask authentic group differences. Larger gender differences could have existed in training group with tests more sensitive to high levels of spatial performance. The finding of narrowed gender differences is especially problematic since we found large spatial gender differences eight months after training and spatial training generally does not narrow gender differences (Uttal et al., in press).

6.3. Spatial training improves learning in Newtonian physics but not in other STEM courses

Third, why did spatial training improve course performance in introductory Newtonian physics, but not for other courses taken concurrently with spatial training (such as engineering)? Engineering course grades were assigned to student teams (rather than individual students) and were largely based on group projects and presentations. This course format suggests that the importance of individuals' spatial skills are not reflected on the final grade, consistent with the complete lack of correlation ($r \sim 0$) of initial spatial skills with engineering grades. However, students *most* frequently mentioned in online surveys that the spatial training was applicable to their current engineering course, especially for the training module on orthographic projections. Hence, the final grade likely did not reflect the positive contributions of spatial training in the parts of the engineering course that covered spatial topics. Finer grained measures of engineering success would be necessary to investigate such hypotheses about engineering outcomes; we did not have access to such metrics (e.g., scores on constructing engineering design sketches). Studies with more average student populations suggest a promising outlook for improving engineering learning with spatial instruction (Alias et al., 2003; Hsi et al., 1997; Sorby, 2009). Finally, compared to physics, the smaller correlation of initial spatial skills with grades in chemistry and mathematics courses suggests that those specific courses simply did not have as many spatial topics relevant to this study's training. Students' survey responses confirmed this hypothesis since chemistry and mathematics courses were mentioned much less frequently than their physics or engineering courses.

6.4. Spatial training effects did not last over eight to ten months

Fourth, why did we find no evidence of lasting spatial improvements when other researchers have? For example, past longitudinal research with about 10 h of training (Feng et al., 2007; Hedley, 2008; Pallrand & Seeber, 1984; Terlecki et al., 2008) has found little decrement in spatial skills after three to five months; these results suggest that the improvements were stable and likely to last for eight months as well. Furthermore, Sorby (2009) found long-term improvements in STEM course grades and engineering retention rates, although those results could be because of self-selection effects. Our extraordinarily talented STEM population may help explain this divergence from past research. As previously mentioned, our sample had extremely high initial spatial skills and students with more average spatial skills may have benefited more from the same amount of spatial experience. As some researchers have suggested (Newcombe, 2007; Uttal et al., in press), such results may represent a "spatial threshold" above which additional spatial skills may have little benefit to STEM success. However, training improvements in physics course performance ($d = .38$) were evident for two months after the last training session; this provides evidence against this threshold hypothesis given our sample's extremely high pretest spatial performance. In longitudinal data, mental rotation performance correlated with physics self-efficacy and some physics learning outcomes after controlling for general academic aptitude, also providing evidence against this threshold hypothesis.

6.5. General limitations: Control group and small sample size

The control group did not complete 12 h of extra practice in some other cognitive non-spatial domain such as writing or vocabulary.

The design of the control group, therefore, confounds the specific effects of spatial training with the general effects of additional cognitive practice and money received in the study (all students received equivalent monetary compensation for completing the spatial tests, but students in the training condition received $90 more to compensate for additional time spent in the training sessions). We respond to this control group limitation with two pieces of empirical evidence. First, past research with these specific training materials indicates that improvements in spatial skills are similar for a "no treatment" group to a group that completed only the software spatial training exercises; differences in spatial improvement only emerged when comparing these groups to groups that completed the *workbook* sketching exercises (Sorby, 2009). Therefore, improvements in spatial skills beyond a "no treatment" condition seem to arise from completing workbook exercises, rather than general effects such as participants' effort involved in the study. Second, if our training condition produced only general effects, we would expect similar improvements on all course grades. Instead, improvements were only found in the subject (Newtonian physics) that (1) correlated most strongly with initial levels of spatial skills, (2) was mentioned frequently on survey responses, and (3) had consistent quantitative improvements across different examinations throughout the semester. For these reasons, we acknowledge that the design of the control group can be improved but does not invalidate the positive results of this study, especially with regards to improved physics course performance.

Another general limitation regards this study's small sample size. The null results for non-physics courses and the null results for the 8–10 month longitudinal data could perhaps be explained by a lack of statistical power. However, we note that the effect size magnitudes for these null results were typically small ($d < .20$) and varied in direction (e.g., the control group sometimes nonsignificantly outperformed the training group), suggesting a lack of statistical trends for these null results. Power analyses indicated that our statistical tests were sufficiently powered to detect at least the large average effect size of $d = .67$ found in previous longitudinal training studies (Feng et al., 2007; Hedley, 2008; Pallrand & Seeber, 1984; Terlecki et al., 2008).

7. Implications and future research

This study finds that sketching-based spatial training challenges even gifted STEM undergraduates and can improve some long-term learning outcomes like grades in calculus-based Newtonian physics. Training only students who fall below a certain spatial performance guideline (e.g., Sorby, 2009) may exclude a wide range of students who could benefit from such instruction. In comparison to the years of systematic development of mathematical, verbal, and writing skills throughout K-12 instruction, 12 h of spatial training in college seems minuscule. Nevertheless, this comparatively short intervention led to demonstrable educational benefits for physics and for some spatial skills. We now outline directions for future research.

A central goal of future research should be to understand the *theoretical mechanisms* for how spatial training can improve spatial skills and learning in STEM fields. As Uttal et al. (in press) noted, specifying improvement mechanisms is centrally important for the design of educational interventions. Although research has specified some basic cognitive pathways to improved spatial skills such as changes in spatial strategies (Glück, Jirasko, Machat, & Rollett, 2001), there remains a dearth of evidence for explaining how spatial training can improve STEM course performance like in physics (NRC, 2006). Kozhevnikov and colleagues (Kozhevnikov, Hegarty, & Mayer, 2002; Kozhevnikov et al., 2007; Kozhevnikov & Thornton, 2006) have provided insightful data regarding the importance of spatial thinking in qualitative, conceptual physics problem solving. With quantitative, protocol analysis, and eye fixation data, Kozhevnikov et al. (2007) argued that, "multidimensional physics problems and spatial visualization tasks require the problem solver to simultaneously process multiple pieces of spatial information that tax the supplies of visual/spatial working memory resources" (p. 576).

Our results, however, suggested that this working memory hypothesis cannot explain the improvements that we found in physics, given the lack of correlation of spatial working memory with physics self-efficacy and all physics learning outcomes. Instead, our results suggested that improvements in specific skills like mental rotation may help explain improvements in physics at least for gifted STEM undergraduates. More in-depth qualitative data are needed to undercover the role of spatial thinking in specific physics contexts, especially comparing quantitative physics problem-solving to qualitative physics understanding.

Above all, we urge future research to investigate how spatial activities such as sketching can enrich existing STEM courses and directly improve STEM student success. Similar to other researchers (National Research Council, 2006; Newcombe, 2010; Piburn et al., 2005), we argue that spatial enriching activities should be *integrated with specific STEM educational topics* such as in engineering design activities (e.g., Youssef & Berry, in press) or analyzing topographic maps in geology courses (Piburn et al., 2005). Such an approach could help learners systematically improve their spatial skills over an extended period of time and also help concretize learning of spatial skills in specific domains. Furthermore, educators and policy makers are unlikely to accept spatial training as an "add-on" course to an already over-packed curriculum—a point that the National Research Council (2006) repeatedly made in its report *Learning to Think Spatially*. The training in this study was comparatively context-free and discipline-general. In online surveys, many students reported difficulty in connecting the training to their current STEM courses. If spatial training topics were instead directly embedded into physics curriculum, we may have found larger and longer-lasting improvements compared the one-third of a letter grade improvement in introductory Newtonian physics. However, this remains a hypothesis for future research.

Future investigations can further delineate the role of spatial skills training in STEM fields, and undercover alternative strategies to improve STEM success for *all* learners including highly gifted STEM undergraduates. Results from this study need replication but present a promising outlook for such future investigations given the improvements that we found in physics learning up to two months after training (although these effects did not last for eight to ten months after training). The National Research Council (2006) called for a national reform to infuse spatial thinking into standards-based curriculum especially in STEM courses. Connecting research on spatial training to long-term STEM learning outcomes can help make that ambitious and critically important goal a reality.

Acknowledgements

This research was funded by the Shanahan Student-Directed Research Funds at Harvey Mudd College. We also thank Peter N. Saeta for his helpful advice and mentoring, Wendy Menefee-Libey for her support from Harvey Mudd's Academic Excellence program, Marcia C. Linn, Jonathan Wai and Joseph J. Williams for comments on this paper, and Alyssa Dray, Brendan Folie, and Neal Pisenti for their time in teaching the spatial training program. Portions of this research will appear in the 2011 Cognitive Science Society conference proceedings.

References

Adams, W. K., Perkins, K. K., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado learning attitudes about science survey. *Physical Review Special Topics - Physics Education Research*, *2*, 1–14.

Alias, M., Black, T. R., & Gray, D. E. (2003). The relationship between spatial visualization ability and problem solving in structural design. *World Transaction on Engineering and Technology Education*, *2*(2), 273–276.

Baenninger, M., & Newcombe, N. S. (1995). Environmental input to the development of sex-related differences in spatial and mathematical ability. *Learning and Individual Differences*, *7*, 363–379.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York, NY: Freeman.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Blasko, D. G., & Holliday-Darr, K. A. (2010). Longitudinal analysis of spatial skills training in engineering graphics. *Proceedings of the 65th Midyear Meeting of the Engineering Design Graphics Division* (pp. 138–151).

Çalişkan, S., Selçuk, G. S., & Erol, M. (2007). Development of physics self-efficacy scale. *Proceedings of the 2007 American Institute of Physics Conference*, *899*. (pp. 483–484).

Casey, M. B., Nuttall, R. L., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology*, *31*, 697–705.

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, *135*, 218–261.

CEEB (1939). *Special Aptitude Test in Spatial Relations, developed by the College Entrance Examination Board, USA.*

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review*, *17*(2), 193–199.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

Cooke-Simpson, A., & Voyer, D. (2007). Confidence and gender differences on the Mental Rotations test. *Learning and Individual Differences*, *17*, 181–186.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, *18*(10), 850–855.

Ferrini-Mundy, J. (1987). Spatial training for calculus students: Sex differences in achievement and in visualization ability. *Journal for Research in Mathematics Education*, *13*, 126–140.

Glück, J., Jirasko, M., Machat, R., & Rollett, B. (2001). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences*, *13*, 1–22.

Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). NY: Psychology Press.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*, 1–51.

Hedley, M. (2008). *The use of geospatial technologies to increase students' spatial abilities and knowledge of certain atmospheric science content* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3313371)

Hegarty, M. (2010). Components of spatial intelligence. In B. H. Ross (Ed.), *The psychology of learning and motivation*, *Vol. 52*. (pp. 265–297)San Diego: Academic Press.

Hegarty, M. (2010b, December). *The role of spatial thinking in undergraduate science education*. Paper presented at the meeting of the National Research Council Committee on the Status, Contributions, and Future Directions of Discipline-Based Education Research, Irvine, CA.

Hegarty, M., Keehner, M., Khooshabeh, P., & Montello, D. R. (2009). How spatial ability enhances, and is enhanced by, dental education. *Learning and Individual Differences*, *19*, 61–70.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141–151.

Hsi, S., Linn, M. C., & Bell, J. E. (1997). The role of spatial reasoning in engineering and the design of spatial instruction. *Journal of Engineering Education*, *86*(2), 151–158.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology. General*, *133*(2), 189–217.

Keppel, G. (1991). *Design and analysis.* Englewood Cliffs, NJ: Prentice Hall.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(4), 989–998.

Kozhevnikov, M., Hegarty, M., & Mayer, R. (2002). Visual/spatial abilities in problem solving in physics. In M. Anderson, B. Meyer, & P. Olivier (Eds.), *Diagrammatic representation and reasoning* (pp. 155–173). London: Springer-Verlag.

Kozhevnikov, M., Motes, M. A., & Hegarty, M. (2007). Spatial visualization in physics problem solving. *Cognitive Science*, *31*, 549–579.

Kozhevnikov, M., & Thornton, R. (2006). Real-time data display, spatial visualization ability, and learning force and motion concepts. *Journal of Science Education and Technology*, *15*, 113–134.

Linn, M. C., Chang, H. -Y., Chiu, J. L., Zhang, H., & McElhaney, K. (2010). Can desirable difficulties overcome deceptive clarity in scientific visualizations? In A. Benjamin (Ed.), *Successful remembering and successful forgetting: a Festschrift in honor of Robert A. Bjork* (pp. 239–262). New York, NY: Routledge.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, *56*(6), 1479–1498.

Lippa, R. A., Collaer, M. C., & Peters, M. (2010). Sex differences in mental rotation and line angle judgements are positively correlated with gender equality and economic development across 53 nations. *Archives of Sexual Behavior*, *39*(4), 990–997.

Lohman, D. F. (1988). Spatial abilities as traits, processes, and knowledge. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence*, *Vol. 4.* (pp. 181–248)Hillsdale, NJ: Erlbaum.

Lohman, D. F., & Nichols, P. D. (1990). Training spatial abilities: Effects of practice on rotation and synthesis tasks. *Learning and Individual Differences*, *2*(1), 67–93.

Moore, D. S., & Johnson, S. P. (2008). Mental rotation in human infants: A sex difference. *Psychological Science*, *19*, 1063–1066.

National Research Council (2006). *Learning to think spatially.* Washington D. C.: The National Academies Press.

National Science Board (2010). *Preparing the next generation of STEM innovators: Identifying and developing our nation's human capital.* Arlington, VA: National Science Foundation.

Newcombe, N. S. (2007). Taking science seriously: Straight thinking about sex differences. In S. Ceci, & W. Williams (Eds.), *Why aren't more women in science? Top gender researchers debate the evidence* (pp. 69–77). Washington, DC: APA Books.

Newcombe, N. S. (2010). Picture this: Increasing math and science learning by improving spatial thinking. *American Educator*, *8*, 29–43.

Newcombe, N. S., & Shipley, T. F. (in press). Thinking about spatial thinking: New typology, new assessments. In J. S. Gero (ed.), *Studying visual and spatial reasoning for design creativity*. New York, NY: Springer.

Pallrand, G. J., & Seeber, F. (1984). Spatial ability and achievement in introductory physics. *Journal of Research in Science Teaching*, *21*(5), 507–516.

Peters, M., Laeng, B., Latham, K., Johnson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse Mental Rotations Test: Different versions and factors that affect performance. *Brain and Cognition*, *28*, 39–58.

Peters, M., Manning, J. T., & Reimers, S. (2007). The effects of sex, sexual orientation, and digit ratio (2D:4D) on mental rotation performance. *Archives of Sexual Behavior*, *36*, 251–260.

Piburn, M. D., Reynolds, S. J., McAuliffe, C., Leedy, D. E., Birk, J. P., & Johnson, J. K. (2005). The role of visualization in learning from computer-based images. *International Journal of Science Education*, *27*(5), 513–520.

Quinn, P. C., & Liben, L. S. (2008). A sex difference in mental rotation in young infants. *Psychological Science*, *19*, 1067–1070.

Reis, S. M., & Renzulli, J. S. (2010). Is there still a need for gifted education? An examination of current research. *Learning and Individual Differences*, *20*, 308–317.

Richland, L. E., Linn, M. C., & Bjork, R. A. (2007). Cognition and instruction: Bridging laboratory and classroom settings. In F. Durso, R. Nickerson, S. Dumais, S. Lewandowsky, & T. Perfect (Eds.), *Handbook of applied cognition* (2nd ed.). New York, NY: Wiley.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology. General*, *125*(1), 4–27.

Sims, V. K., & Mayer, R. E. (2002). Domain specificity of spatial expertise: The case of video game players. *Applied Cognitive Psychology*, *16*(1), 97–115.

Small, M. Y., & Morton, M. E. (1983). Research in college science teaching: Spatial visualization training improves performance in organic chemistry. *Journal of College Science Teaching*, *13*(1), 41–43.

Smith, I. M. (1964). *Spatial ability: Its educational and social significance.* London: University of London Press.

Sorby, S. A. (2009). Educational research in developing 3-D spatial skills for engineering students. *International Journal of Science Education*, *31*, 459–480.

Sorby, S. A., & Wysocki, A. F. (2003). *Introduction to 3-D spatial visualization: An active approach.* Clifton Park, NY: Thomson-Delmar Learning.

Stransky, D., Wilcox, L. M., & Dubrowski, A. (2010). Mental rotation: Cross-task training and generalization. *Journal of Experimental Psychology. Applied*, *16*(4), 349–360.

Strasser, I., Koller, I., Strauß, S., Csisinko, M., Kaufman, H., & Glück, J. (2010). Use of strategy in a 3-dimensional spatial ability test. *Journal of Individual Differences*, *31*, 74–77.

Terlecki, M. S., Newcombe, N. S., & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied Cognitive Psychology*, *22*(7), 996–1013.

The College Board (2010). SAT percentile ranks for males, females and total group: Mathematics. Retrieved April 22, 2011, from. http://professionals.collegeboard.com/profdownload/sat-mathematics-percentile-ranks-2010.pdf

Tuckey, H., Selvaratnam, M., & Bradley, J. (1991). Identification and rectification of student difficulties concerning three-dimensional structures, rotation, and reflection. *Journal of Chemical Education*, *68*(6), 460–464.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A., Warren, C., et al. (in press). Training spatial skills: What works, for whom, and for how long? *Psychological Bulletin*.

Vandenberg, S., & Kuse, A. R. (1978). Mental rotations: A group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*, 599–604.

Veurink, N. L., Hamlin, A. J., Kampe, J. C. M., Sorby, S. A., Blasko, D. G., Holliday-Darr, K. A., et al. (2009). Enhancing visualization skills—Improving options and success (EnViSIONS) of engineering and technology students. *Engineering Design Graphics Journal*, *73*(2), 1–17.

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil test of mental rotation: A meta-analysis. *Psychonomic Bulletin & Review*, *18*(2), 267–277.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250–270.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*, 817–835.

Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., & Kosslyn, S. M. (2008). Training generalized spatial skills. *Psychonomic Bulletin & Review*, *15*, 763–771, doi: 10.1007/s10798-010-9151-3.

Youssef, B. B., & Berry, B. (in press). Learning to think spatially in an undergraduate interdisciplinary computational design context: a case study. *International Journal of Technology and Design Education*.