Educ Psychol Rev (2013) 25:69–94
DOI 10.1007/s10648-012-9215-x

REVIEW ARTICLE

A Meta-Analysis on Gender Differences in Mental Rotation Ability Measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R)

Yukiko Maeda · So Yoon Yoon

Published online: 9 December 2012

Abstract This meta-analysis was conducted to estimate the magnitude of gender difference in three-dimensional (3-D) mental rotation ability and to investigate how factors related to test administration conditions play a role in varying gender difference effect sizes and threatening validity. Individuals' 3-D mental rotation ability was measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R). We integrated 70 effect sizes of gender differences in mental rotation ability measured by the PSVT:R which were obtained from 40 primary studies. The results indicated that male participants outperformed females on the test (Hedges' $g=0.57$). The I^2 statistic indicated 41.7 % of variation in effect sizes reflects real heterogeneity. The moderator analysis indicated that male superiority on spatial ability tasks measured by the PSVT:R is related to the implementation of time limits. The gender difference became larger when stringent time limits (equal or less than 30 s per item) were implemented.

Keywords Meta-analysis · Gender difference · Spatial ability · Mental rotation · The PSVT:R

Introduction

The Relation of Mental Rotation Ability to Academic Learning

Spatial ability is generically defined as "the ability to generate, retain, retrieve and transform well-structured visual images" (Lohman 1996, p. 112). However, the notion of spatial ability varies, so does the operationalization of the ability for measurement across studies, which

Y. Maeda (✉)
Department of Educational Studies, Beering Hall of Liberal Arts and Education, Purdue University,
Room 5152 100N. University Street, West Lafayette, IN 47907-2067, USA
e-mail: ymaeda@purdue.edu

S. Y. Yoon
Institute for P-12 Engineering Research and Learning (INSPIRE), School of Engineering Education,
Purdue University, 315 N. Grant Street, West Lafayette, IN 47907-2023, USA
e-mail: yoon18@purdue.edu

may contribute to the difficulty in providing a unified definition (Höffler 2010; Lohman 1996). Yet, researchers agree that the ability constitutes human intelligence and is an essential prerequisite for learning (Höffler 2010). Studies dealing with spatial ability have accumulated evidence to support a positive link between the ability and performance in various academic fields, particularly in the fields of science, technology, engineering and mathematics (STEM).

For example, it is consistently reported that spatial ability relates to mathematical conceptualization (Smith 1964), problem-solving skills (Moses 1977), creative and higher order thinking skills in science and mathematics (Shepard 1978; Smith 2009), and design and graphical representation skills in engineering and technology (Strong and Smith 2001/2002). Furthermore, a series of studies led by the Lubinski and Benbow research group provided evidence for predictive validity of spatial ability to make inferences about academic success in STEM areas (Humphreys et al. 1993; Shea et al. 2001) and indicated a unique contribution of spatial ability to predict success in STEM beyond SAT mathematics and verbal scores (Lubinski 2010; Wai et al. 2009). Nonetheless, spatial ability, in spite of its critical role in STEM education, has received less attention in educational settings compared to other cognitive abilities, such as verbal and mathematical abilities (Gohm et al. 1998).

The construct of spatial ability can be further divided into several sub-factors depending on the emphasis given to a specific aspect of the mental process on visualizing images (Carroll 1993; Lohman 1996). Although there is no consensus on dimensions in spatial ability, researchers seem to agree that spatial ability consists of at least two correlated, but theoretically separable, core dimensions: spatial visualization (SV), which usually involves the spatial transformation of objects, and spatial relation/orientation, which does not necessarily involve spatial transformation (Just and Carpenter 1985). Mental rotation ability requires a cognitive process to mentally transform or rotate two- or three-dimensional (3-D) objects to whatever direction is indicated through SV (Carroll 1993). Compared to other sub-factors in spatial ability, mental rotation tends to show larger individual and/or gender differences. In addition, mental rotation ability and its relationship to academic performance are of particular interest to STEM researchers, as learning in STEM disciplines often involves tasks, such as technical drawing, geometrical problem-solving, or molecular structure representations of chemical materials, which require a strong visuoperceptual sense to grasp the dynamics and structures of objects in 3-D space and the spatial transformation of objects. Accordingly, some researchers have attempted to take students' performances on a mental rotation test into account when designing instructions (e.g., Stieff 2007) and providing appropriate interventions for improving academic performance in a subject area (e.g., Sorby 2009). As the significant role of mental rotation in academic success is recognized, the use of spatial or mental rotation tests is expected to increase in educational research and practice.

Gender Differences in Mental Rotation Ability

In the meantime, understanding gender differences in spatial ability has been an issue posited by psychologists and educational researchers since the early history of spatial ability research (e.g., Carroll 1993; Eliot 1987; Harris 1978). In particular, among research on different sub-factors of spatial ability, research focusing on mental rotation ability provides consistent evidence for gender differences that favor males (e.g., Linn and Petersen 1985; Voyer et al. 1995). While the trend seems evident, the magnitude of gender difference varies across studies. For example, Linn and Petersen (1985) conducted the first meta-analysis with 29 effect sizes retrieved from 9 studies, and found significant heterogeneity of the effect sizes in mental rotation. Voyer et al. (1995) expanded the study by Linn and Petersen with 50 effect sizes retrieved from 44 studies reported between1974 and 1993. In the meta-analysis, Voyer et al. (1995) could further identify the variation in the effect sizes by the types of the tests used to measure spatial ability.

However, the reasons for the observed gender differences in mental rotation ability and the variation of their magnitudes are not yet fully understood (e.g., Masters 1998; Moè 2009). Our literature review suggests at least five distinguishable factors that may explain the cause and the extent of these gender differences. We labeled them as biological, strategic, experiential, affective, and test administration factors in this study and briefly summarized them below. Biological differences between females and males may contribute to gender differences in mental rotation ability. For example, researchers (Jordan et al. 2002; Koscik et al. 2009) attributed the observed gender differences to functional and morphological differences in the brains of males and females. Other researchers indicated that hormonal differences (Hausmann et al. 2000) or genetic differences (Boles 1980; Harris 1978; Thomas and Kail 1991) are related to the differences in performance on mental rotation tasks. The results of the meta-analysis by Linn and Petersen (1985) indicated that gender differences are related to the unique processes of biological development for each sex, and Geiser et al. (2008) supported their findings.

Others further suggested that biological sex differences lead to gender differences in cognitive processing and adopting a certain problem-solving strategy for metal rotation tasks (Geiser et al. 2006; Guay 1978; Jordan et al. 2002; Thomsen et al. 2000). For example, males and females adopt different mental strategies to evaluate response options in each item (Hirnstein et al. 2009). Also, different strategies may be used by each gender when guessing plausible correct responses (Voyer and Saunders 2004). As a result, males may perform better than females on mental rotation tasks.

Alternatively, the magnitude of gender differences in mental rotation ability may be attributed to the amount of previous spatial experiences an individual has had. In other words, mental rotation ability can be fostered through activities involving spatial tasks (Feng et al. 2007; Ginn and Pickens 2005); researchers found that some activities related to spatial tasks may be more prevalent among boys than girls, such as playing computer and/or video games (Cherney 2008; Feng et al. 2007; Quaiser-Pohl et al. 2006), playing certain types of sports, (Ginn and Pickens 2005), or playing with building toys, such as Lego® or blocks (Deno 1995). These activities provide boys with spatial experiences more often than girls, which may contribute to boys' superiority on mental rotation tasks.

Recent studies reported that an individual's affective state can influence his or her performance on spatial tasks. For example, an individual's performance differs according to his or her level of confidence in solving of spatial tasks (Cooke-Simpson and Voyer 2007). Other studies showed that when an individual's psychological state is shaped by common gender beliefs (Cooke-Simpson and Voyer 2007; Moè 2009; Moè and Pazzaglia 2006) and/or gender stereotypes (Ortner and Sieverding 2008; Sharps et al. 1994), the individual's performance on spatial tasks is affected. For example, gender differences in performance on mental rotation tasks increased when students were primed to believe that males have an advantage over females when it comes to successfully completing mental rotation tasks (Ortner and Sieverding 2008).

The last factor we identified is related to test administration conditions. The characteristics of an instrument used to measure mental rotation ability and the conditions under which the instrument is administered can affect an individual's performance (Anastasi and Urbina 1997; McCallin 2006). Several studies indicate that certain administration conditions may contribute to males' superior performance on mental rotation tests. For example, Goldstein et al. (1990) found that gender differences disappeared in the average scores on the Mental Rotation Test (MRT) (Vandenberg and Kuse 1978) when no time limit was applied to solve the items. However, they also found that a particular way of assigning scores increased the gender gap in favor of males.[1] These findings are not conclusive, since other researchers reported contradicting results (e.g., Masters 1998; Titze et al. 2008). The test delivery mode may also contribute to differences in spatial performance by gender. Monahan et al. (2008) found that superior performance of males to that of females on a mental rotation test decreased from a large to a medium effect size (from Cohen's $d=1.05$ to 0.59) when the same spatial test was administered in a computer-based format instead of a paper-and-pencil format when all other testing conditions were identical.

It is recognized that the speed for solving spatial problems differs by gender, and that this difference may produce gender differences in mental rotation scores when a time limit is implemented (Peters 2005; Voyer and Saunders 2004). For example, with the revised MRT test (Peters et al. 1995), Peters (2005) found that more women could not reach the end of a test (with a 6-min time limit to solve two sets of 12 items) than men, because women tended to require more time to complete mental rotation tasks than men. However, no change in the magnitude of gender differences was observed, even when participants were given twice as long as the original time limit to solve the exact same set of items, and the scores of both men and women increased. Peters speculated that the increase in scores that occurred for both men and women was caused by having additional time, which equally helped both genders to solve items correctly. Therefore, he conjectured that men would perform better than women regardless of time restriction, and men's superiority would be more significant as the time given to solve each item is shortened. Interestingly, Voyer et al. (2004) found that adapting a time limit may also affect females' response behaviors; females are more prone to guessing answers if they have a longer time to respond.

Impacts of Administration Conditions on Interpreting the Gender Differences in Mental Rotation Ability

As we reviewed, the relevant literature generally supports males' outperformance on mental rotation tests with few exceptions (Brownlow et al. 2003; Yue et al. 2001, June); several factors may explain how and to what extent the gender differences exist. However, while all factors except for those related to test administration conditions can contribute to actual gender differences in mental rotation ability, the gender differences caused by the administration conditions of a test largely resulted from variant methodological choices used to measure the ability. The variation in test scores resulting from the influence of test administration conditions can be considered as construct-irrelevant variance (CIV) because it does not reflect the variation due to actual individual differences of ability or knowledge; rather, it reflects the variation due to measurement errors (Haladyna and Downing 2004; McCallin 2006; Netemeyer et al. 2003). In general, CIV arises from the influence of psychological or situational factors, such as different test administration conditions that affect performance of a specific group of people or individuals (Haladyna and Downing 2004; Messick 1984). Haladyna and Downing (2004) stated that altering test administration conditions from the

[1] The MRT consists of 20 items and each multiple choice item of the MRT had two correct answers when it was developed. A participant gets one point when he or she identifies the two correct answers. A participant gets no point when he or she identifies only one or none of the two correct answers. Thus, the maximum score of the MRT is 20 points. An alternative scoring method gives a point for each correct answer and accordingly the maximum score is 40 points.

standardized conditions, such as implementing time limits and/or computer-based testing (CBT), is a potential source for systematic errors associated with CIV. Huff and Sireci (2001) also explained how the implementation of CBT may introduce CIV due to individual differences in familiarity with computer platform and interface, computer proficiency, and ability to work quickly while taking a test. As a systematic measurement error, CIV introduces bias in an estimate of a respondent's true ability score. Although all of the identified five factors that may cause gender differences in spatial ability deserve attention in order to explain the variation in effect sizes of the gender differences, this study focused on the impact of the factors related to test administrations because the variation due to these factors is potentially considered to be a measurement error, and thus it will threaten accurate interpretation of the effect sizes for gender differences in mental rotation ability.

Purpose of the Study

The main goal of the current study was to investigate whether some of variations in the observed effect sizes, representing gender differences in mental rotation ability, resulted from procedural artifacts. In other words, using a meta-analytic method, we first quantitatively synthesized the relevant studies on gender differences in 3-D mental rotation ability measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R) in order to characterize the typical magnitude of the differences and their variations across studies. Second, we focused on the factors related to test administration as moderators of interest to estimate the artificial variance in gender difference effect sizes due to methodological choices.

Specific research questions addressed in this meta-analysis addresses are: (a) how large is the typical effect size that indicates gender differences in mental rotation ability measured by variants of the PSVT:R?; (b) to what extent do effect sizes vary across studies?; and (c) how do effect sizes differ according to the modes of administration, length of the test, time limits, figure modifications, as well as study characteristics (such as participants' education levels and academic majors) and the type of publication? Since the differences caused by the administration conditions can be considered measurement errors, and the errors will hinder the accurate interpretation of gender differences of mental rotation, identifying the amount of variation and the characteristics that caused the variation should help us to better understand gender differences under a specific definition of spatial ability.

Methods

Population of Studies

The population of the current meta-analysis is the studies that provide the information of gender differences in mental rotation ability, which is operationally defined as the ability that is measured by the PSVT:R (Guay 1976). Narrowing the definition of mental rotation ability by focusing on the PSVT:R might limit the generalizability of our findings. However, this decision was made because of the following reasons. First, we concur with the claim of Caplan et al. (1985) and Voyer et al. (1995) that meta-analysts face difficulties when synthesizing and/or comparing effect sizes of spatial abilities across studies, because studies rarely provide clear definitions of spatial ability and the rationale for the use of the selected test.

Second, as Morris and DeShon (2002) discussed, the conversion of the effects obtained from different studies on a common metric, such as Cohen's *d* for synthesis, would make sense only when the definition of the theoretical construct of spatial ability represented by the scores on a specific test is equivalent. Third, the assumption about the same underlying construct across scores on different mental rotation tests is less likely to be hold, because employed tasks in each test may have different complexities in spatial problem solving, which may evoke the differences in the cognitive process or the strategic approach to solve the tasks (Voyer et al. 1995). Finally, and most importantly, focusing on the particular test made it possible for us to pay close attention to test administration issues to explain the heterogeneity of the effect sizes, which is the main objective of this study. The practical implication of our findings will be straightforward for the users of the PSVT:R.

Among various mental rotation tests, we chose the PSVT:R for our study because of its inconsistent administration and prevalent use in educational research. The original PSVT, which consists of three subtests entitled Developments, Rotations, and Views, respectively, was developed by Guay (1976). The test is comprised of 36 items, 12 from each subtest. The PSVT:R is an extended version of the subtest, Rotations, used to measure 3-D mental rotation ability of individuals age 13 or older in 20 min (Guay 1980). The PSVT:R has 30 items consisting of 13 symmetrical and 17 nonsymmetrical figures of 3-D objects, which are drawn in a two-dimensional (2-D) isometric format. A sample item of the PSVT:R is shown in Fig. 1. All figures contain cubes or cylinders with varied truncated slots. In each item, the respondents' task is to mentally rotate a figure in the same direction visually indicated in the instructions and identify the most appropriate choice among the five options.

The PSVT:R[2] has been frequently used in educational research, particularly in STEM disciplines, for over three decades (Contero et al. 2005; Field 2007), while the MRT has been popularly used in the fields of psychology and social studies. The prevalent use of the PSVT:R, especially in STEM research, may be supported by the following characteristics of the test.

1. Several studies have reported relatively strong reliability and/or validity evidence to support the use of the PSVT:R as summarized in Table 1.
2. The ability to solve the tasks utilized in the PSVT:R is related to the ability to solve tasks often found in the STEM disciplines (Bodner and Guay 1997; Yue 2004).
3. The PSVT:R contains items that are difficult enough to distinguish STEM students by their level of mental rotation ability (Black 2005; Yue 2006, October).
4. The PSVT:R has been frequently cited as the strongest measure of spatial visualization ability of mental rotation that most incorporates the holistic or gestalt spatial thinking process and least incorporates the analytic or analogical spatial thinking process (Black 2005; Branoff 1998; Guay et al. 1978).
5. The PSVT:R is available for researchers free of charge and is easy to score because of its multiple choice format.

Nevertheless, the test was not included in previous landmark meta-analyses conducted by Linn and Petersen (1985) and Voyer et al. (1995).

In addition, the PSVT:R is often modified for practical reasons without any additional evidence to support the equivalence of its property with that of the original PSVT:R. Examples of modifications made while administrating the PSVT:R include the change of test delivery mode from paper-and-pencil to a computer-based format (Branoff 1998, 2000; Yue 2002), the addition of visual aids on items, such as adding 3-D axes into figures

[2] The instrument is also cataloged in the *Educational Testing Service Test Collection* (Educational Testing Service 2009) and *An International Directory of Spatial Tests* (Eliot and Smith, 1983).

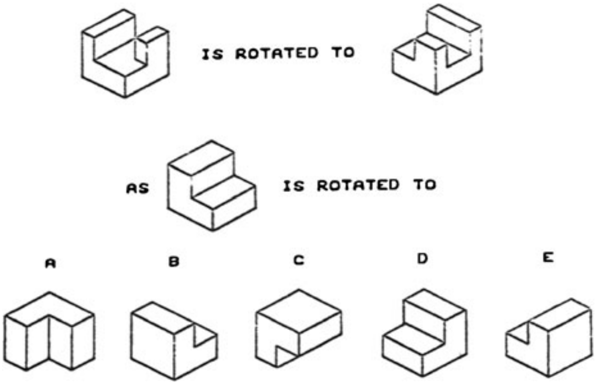


Fig. 1 A sample item of the PSVT:R (Guay 1976)

(Hagevik 2003; Smith 2009), the change of figures to trimetric views (Branoff 2000), the omission of some items to make the test length shorter (Battista 1990; Parolini 1994), and the implementation of a time limit different from 20 min, which was suggested by Guay (1980) (Battista 1990; Hassan and Abed 1999). These changes affect the accurate estimation of an individual's mental rotation score (Anastasi and Urbina 1997) and therefore may result in bias in estimating the effect size of gender difference. Inconsistent administration of the instrument among researchers may be due to the fact that neither a technical manual of the PSVT:R is available, nor is the PSVT:R standardized. For those reasons, we considered the PSVT:R to be an appropriate example to demonstrate the variation in gender differences due to administration differences, as well as to provide new evidence of gender differences in mental rotation ability in a frequently used test in educational research.

Sampling of Primary Studies

We attempted to retrieve all studies conducted between 1976, when Guay first developed the PSVT, and 2011. Scholarly work fitting this description was identified through electronic databases (e.g., Education Resources Information Center (ERIC), PsycINFO, and ProQuest Dissertation and Thesis Database), search engines (e.g., Google Scholar, and Yahoo!). Online search keywords were combinations of the following: PSVT:R, ROT, PSVT, spatial ability, visualization, spatial test, gender, sex, Purdue, Guay, Bodner, rotation. We also examined references of identified articles through the electronic search. A total of 181 studies, including 63 journal articles, 31 dissertations, 13 master's theses, 173 conference proceedings, and one technical report, were identified as potential studies to be included in the meta-analysis. A study was included in the current meta-analysis when the study met the following inclusion criteria: (a) the PSVT:R or any variants of the instrument including the Purdue Visualization of Rotations Test (ROT)[3] and a subsection of the PSVT (i.e., PSV:R)[3] were used, (b) Cohen's *d* effect size or sufficient statistical information (e.g., means and standard deviations of the PSVT:R scores, *t* statistic, *F* statistic with 1 numerator *df* for gender difference, correlation coefficient between gender and the PSVT:R scores, or other statistics) to calculate Cohen's *d* effect sizes for gender differences were available, and (c) sufficient details of critical features of study design including participants characteristics

[3] The Purdue Visualization of Rotations Test (ROT) is a shortened version of the PSVT:R with 20 items, and the PSV:R is a 12-item shortened version of the PSVT:R

Table 1 Reliability and validity evidence of the PSVT:R

Author	Sample	Reliability	Validity
Alkhateeb (2004)	180 education major undergraduate students enrolled in mathematics courses	r_α=0.81	
Battista et al. (1982)	82 pre-service elementary teachers enrolled in a geometry course	$r_{KR\text{-}20}$=0.80	
Branoff (2000)[a]	139 undergraduate students enrolled in graphic communication courses		r_P=0.67 with the MRT
Guay and McDaniel (1978)	101 undergraduate students	r_{SH}=0.90	r_P=0.61 with the Shepard-Metzler Rotations Tests
		$r_{KR\text{-}20}$=0.86	r_P=0.25 with the RMPFBT
Maeda and Yoon (2011)[a]	585 first year engineering students	r_α=0.84	One factor structure was confirmed through exploratory and confirmatory factor analysis
Sorby and Baartmans (1996)	492 first year engineering students	$r_{KR\text{-}20}$=0.82	
Sorby and Baartmans (2000)	27 first year engineering students	$r_{KR\text{-}20}$ of pre-test=0.82	
		$r_{KR\text{-}20}$ of post-test=0.71	
Sorby (2000)	Approximately 80 undergraduate students		r_P=0.53 with the MCT
			r_P=0.19 with the DAT:SR (not significant)
Yoon (2011)[b]	1022 undergraduate students	r_α=0.86	One factor structure was confirmed through confirmatory factor analysis

r_α Cronbach's alpha reliability coefficient; r_{SH} Split-half reliability coefficient; $r_{KR\text{-}20}$ Kuder–Richardson reliability coefficient; r_P Pearson product–moment correlation coefficient; *RMPFBT* Revised Minnesota Paper Form Board Test; *MCT* the Mental Cutting Test (MCT) (College Entrance Examination Board 1939); *DAT:SR* Differential Aptitude Tests: Spatial Relations (Bennett et al. 1973)

[a] This study used a computer-based form of the PSVT:R

[b] The Revised PSVT:R was used

were provided. If the same study was reported through two different venues (e.g., reported in a journal article and in a dissertation), we selected the document that provided the most detailed information about the study to include in the meta-analysis. Otherwise, a study that did not meet at least one of these criteria was excluded from the meta-analysis. As a result, 40 of the 181 studies met all criteria and were included in the sample for the meta-analysis.[4]

Coding of Studies

Note that the effect of administration conditions on the magnitude of gender gap is our primary interest. Therefore, we developed a coding scheme to code key characteristics of

[4] References with * in the reference list indicate the studies used for the current meta-analysis.

primary studies, including the type of the PSVT instrument (PSVT:R, ROT, or PSV:R), the number of questions in the instrument, the time restriction for administration, the administration mode (i.e., computer-based or paper-and-pencil), and the inclusion of additional visual aids in figures as well as study characteristics (such as types of documentation). We also coded participants' education levels and academic majors of participants in primary studies. These features served as possible moderators to explain the variation in effect sizes across primary studies (Lipsey 2009; Wilson 2009).

The primary effect size chosen for this meta-analysis is the standardized mean difference, Hedges' *g* (Hedges 1981) because Cohen's *d* statistic tends to overestimate the population effect sizes with a small sample size (Borenstein 2009). We also computed Cohen's *d* (Cohen 1977) as it is typically used in research. However, the correlation between these two indices is 0.9997, and the average discrepancy in the magnitude between two indices is 0.008 in this study. Therefore, we mainly report the results with Hedges'*g*, but also reported Cohen's *d*, if necessary.

Once the validity of the coding scheme was established with multiple iterations of trial coding and revisions of coding scheme, both authors coded the same five articles together, and then checked the consistency of the coding to examine the reliability of coding. If an inconsistency occurred, it was resolved through discussion, so coding criteria and processes were clarified. After interrater reliability was established, each author coded half of the located primary studies, and whenever difficulty in coding occurred, both authors crosschecked the coding over the primary studies. At regular research meetings, these ambiguities and issues were discussed and resolved through the consensus between two authors.

Table 2 summarizes the coded key variables and 70 effect sizes obtained from the 40 studies included in this meta-analysis. Note that a single study can provide more than one effect size for gender differences. Inclusion of multiple effect sizes obtained from the same sample will cause dependency among data. The problems associated with dependent effect sizes has been recognized (e.g., Strube 1987; Wood 2008) and several methods to handle the problem were proposed, such as selecting one effect size from each uniquely identified sample, averaging the multiple effect sizes within the study, or applying statistical adjustment to combine multiple effect sizes within a study. Applying statistical adjustment was not feasible because the required additional statistics for the adjustment were not available in primary studies. Therefore, we decided to select one effect size if the study reported multiple effect sizes from the identical sample. More specifically, if a study utilized a pre-post or repeated measures design to investigate an intervention effect on the PSVT:R scores, we selected the pre-test or initial test scores to calculate an effect size. For example, Eraso (2007) reported gender differences on spatial ability measured by the PSVT:R for both pre- and post-tests, resulting in two effect sizes. We selected the effect size computed with pre-test scores for analyses with the independent effect sizes. We also retrieved multiple effect sizes of gender differences from more than one independent group, as shown in Table 2. For example, Provo (1996) collected data from four separate groups of first year veterinary medicine students in her study. The effect size obtained from each group was treated independently and included in the meta-analysis.

Data Analysis

Because the 70 effect sizes retrieved from the 40 studies are different in terms of the characteristics of samples and utilized research design and procedures, the random-effects

Table 2 Primary Studies and Variables Employed to Calculate Effect Sizes

Study	ID	Type	Test	N_Q	T_t	T_i	Mode	Visual aid	N_M	N_F	Education Level	Major	d	SE_d	g	SE_g
Battista (1980)	1	J	PSVT:R	30	–	–	PBT	–	25	32	UG	Mixed	0.41	0.27	0.40	0.27
Battista (1990)	2	J	PSVT:R	20	8	24	PBT	–	75	53	HS	–	0.95	0.19	0.94	0.19
Black (2005)	3	J	PSVT:R	30	20	40	PBT	–	33	64	UG	Non-STEM	0.38	0.22	0.38	0.21
Bock (2005)[a]	4	T	PSVT:R	30	–	–	CBT	–	39	39	UG	STEM	1.12	0.24	1.11	0.24
Branoff (1998)	5	D	PSVT:R	30	–	–	CBT	–	64	17	UG	Mixed	0.65	0.28	0.64	0.28
Branoff (1999)[b]	6	J	PSVT:R	30	–	–	CBT	–	95	31	UG	Mixed	0.61	0.21	0.61	0.21
	7	J	PSVT:R	30	–	–	CBT	3-axes	93	30	UG	Mixed	0.18	0.21	0.18	0.21
Branoff (2000)	8	J	PSVT:R	30	–	–	CBT	–	118	21	UG	Mixed	0.27	0.24	0.27	0.24
	9	U	PSVT:R	30	–	–	CBT	Trimetric	117	21	UG	Mixed	0.46	0.24	0.45	0.24
Branoff and Connolly[c] (1999)	10	C	PSVT:R	30	–	–	CBT	–	–	–	UG	Mixed	0.68	–	0.66	–
	11	C	PSVT:R	30	–	–	CBT	3-axes	–	–	UG	Mixed	0.59	–	0.57	–
Brownlow et al. (2003)	12	J	ROT	20	–	–	PBT	–	16	32	UG	Non-STEM	−0.08	0.31	−0.08	0.30
	13	J	ROT	20	–	–	PBT	–	20	28	UG	Non-STEM	1.39	0.33	1.37	0.32
	14	J	ROT	20	–	–	PBT	–	15	18	UG	Non-STEM	0.91	0.37	0.88	0.36
Brownlow et al. (2011)	15	J	PSVT:R	10	–	–	PBT	–	16	16	UG	Mixed	0.51	0.36	0.50	0.35
	16	J	PSVT:R	10	–	–	PBT	–	16	16	UG	Mixed	0.29	0.36	0.28	0.35
	17	J	PSVT:R	10	–	–	PBT	–	16	16	UG	Mixed	0.60	0.36	0.58	0.35
Brus et al. (2004)	18	C	PSVT:R	30	–	–	PBT	–	184	61	UG	STEM	0.54	0.15	0.54	0.15
Chae et al. (2008)[d]	19	C	ROT	20	10	30	PBT	–	42	13	ES,MS,HS	–	0.20	0.32	0.20	0.31
	20	C	ROT	20	10	30	PBT	–	21	30	ES,MS,HS	–	0.97	0.30	0.95	0.30
Connolly et al. (2009)[b]	21	C	PSVT:R	28	–	–	PBT	–	57	12	UG	Mixed	0.20	0.32	0.20	0.31
Dean (2009)	22	J	ROT	20	10	30	PBT	–	67	9	UG	Mixed	0.45	0.36	0.44	0.35
	23	J	ROT	20	10	30	PBT	–	13	25	UG	Mixed	1.00	0.36	0.98	0.35
Eraso (2007)	24	D	PSV:R	12	10	50	PBT	–	8	13	HS	–	0.52	0.46	0.50	0.44
	25	D	PSV:R	12	10	50	PBT	–	10	12	HS	–	0.65	0.44	0.63	0.42
	26	D	PSV:R	12	10	50	PBT	–	9	12	HS	–	0.03	0.44	0.02	0.42
	27	D	PSV:R	12	10	50	PBT	–	17	14	HS	–	0.62	0.37	0.61	0.36

Table 2 (continued)

Study	ID	Type	Test	N_Q	T_t	T_i	Mode	Visual aid	N_M	N_F	Education Level	Major	d	SE_d	g	SE_g
	28	D	PSV:R	12	10	50	PBT	–	13	15	HS	–	0.42	0.38	0.40	0.37
	29	D	PSV:R	12	10	50	PBT	–	26	9	HS	–	0.18	0.39	0.18	0.38
Ernst and Clark (2009)[b]	30	J	PSVT:R	30	–	–	PBT	–	362	210	MS+HS	–	0.35	0.09	0.35	0.09
Grabow (2003)	31	T	ROT	20	10	30	PBT	–	58	76	HS	–	0.47	0.18	0.47	0.18
Guay (1978)	32	C	PSV:R	12	–	–	PBT	–	82	135	UG	Mixed	0.93	0.15	0.93	0.15
Guay and McDaniel (1978)	33	R	PSVT:R	30	–	–	PBT	–	50	51	UG	Mixed	1.04	0.21	1.03	0.21
Hagevik (2003)	34	D	PSVT:R	30	–	–	PBT	3-axes	97	67	MS	–	1.10	0.17	1.09	0.17
Hake (2002)	35	C	PSVT:R	30	–	–	PBT	–	110	93	UG	Non-STEM	0.82	0.15	0.82	0.15
Harris et al. (2009)	36	J	ROT	20	–	–	PBT	–	10	10	UG	STEM	0.04	0.45	0.04	0.43
Hassan and Abed (1999)	37	J	PSVT:R	20	8	24	PBT	–	12	28	BA	–	1.33	0.38	1.30	0.37
Koch (2006)	38	D	PSVT:R	30	–	–	PBT	–	34	13	UG	STEM	0.56	0.33	0.55	0.33
Kovac and Rensselaer (1989)	39	J	PSV:R	12	–	–	PBT	–	28	30	MS	–	0.03	0.26	0.03	0.26
Lindsay (2001)	40	D	ROT	20	10	30	PBT	–	45	48	UG	STEM	0.62	0.21	0.61	0.21
	41	D	ROT	20	10	30	PBT	–	41	59	UG	STEM	0.76	0.21	0.75	0.21
Maeda and Yoon (2011)	42	C	PSVT:R	30	25	50	CBT	–	480	105	UG	STEM	0.24	0.11	0.24	0.11
Parolini (1994)[e]	43	T	PSVT:R	15	–	–	PBT	–	418	117	UG	STEM	0.66	0.11	0.66	0.11
Poulin et al. (2004)	44	J	ROT	20	10	30	PBT	–	86	132	UG	Mixed	0.74	0.14	0.74	0.14
Provo (1996)	45	T	ROT	20	–	–	PBT	–	10	12	UG	Non-STEM	1.12	0.46	1.07	0.44
	46	T	ROT	20	–	–	PBT	–	5	17	UG	Non-STEM	0.59	0.52	0.57	0.50
	47	T	ROT	20	–	–	PBT	–	5	17	UG	Non-STEM	0.30	0.51	0.29	0.49
	48	T	ROT	20	–	–	PBT	–	16	40	UG	Non-STEM	0.31	0.30	0.31	0.29
Santone (2009)	49	D	PSV:R	12	10	50	Mixed	–	377	66	UG	Mixed	0.51	0.13	0.50	0.13
Schoenfeld-Tacher (2000)	50	D	ROT	20	10	30	PBT	–	202	255	UG	Mixed	0.62	0.10	0.62	0.10
Smith (2009)[b]	51	D	PSVT:R	20	20	60	PBT	3-axes	138	16	UG	STEM	0.37	0.26	0.37	0.26
Sorby et al. (2006a)	52	C	PSVT:R	10	–	–	PBT	–	22	18	HS	–	0.82	0.33	0.80	0.32
Sorby et al. (2006b)	53	J	PSVT:R	10	–	–	PBT	–	63	66	MS, HS	–	0.48	0.18	0.48	0.18
Stumpf and Eliot (1995)[f]	54	J	PSVT:R	30	–	–	PBT	–	95	57	MS, HS	–	0.56	0.17	0.56	0.17

Table 2 (continued)

Study	ID	Type	Test	N_Q	T_t	T_i	Mode	Visual aid	N_M	N_F	Education Level	Major	d	SE_d	g	SE_g
Titus and Horsman (2009)[b]	55	J	PSVT:R	10	3	18	PBT	–	21	36	UG	Mixed	0.15	0.27	0.15	0.27
	56	J	PSVT:R	10	3	18	PBT	–	34	39	UG	STEM	0.79	0.24	0.78	0.24
	57	J	PSVT:R	10	3	18	PBT	–	13	13	UG	STEM	0.90	0.41	0.87	0.40
Yoon (2011)	58	D	PSVT:R	30	–	–	PBT	–	92	63	UG	Non-STEM	0.23	0.16	0.23	0.16
	59	D	PSVT:R	30	–	–	PBT	–	39	190	UG	Non-STEM	0.79	0.18	0.79	0.18
	60	D	PSVT:R	30	–	–	PBT	–	38	17	UG	STEM	0.55	0.30	0.54	0.29
	61	D	PSVT:R	30	–	–	PBT	–	83	179	UG	Non-STEM	0.40	0.13	0.40	0.13
	62	D	PSVT:R	30	–	–	PBT	–	26	17	UG	Non-STEM	0.37	0.31	0.36	0.31
	63	D	PSVT:R	30	–	–	PBT	–	46	32	UG	STEM	0.94	0.24	0.94	0.24
	64	D	PSVT:R	30	–	–	PBT	–	77	21	UG	STEM	1.00	0.26	0.99	0.25
	65	D	PSVT:R	30	–	–	PBT	–	6	9	UG	Mixed	0.74	0.54	0.69	0.51
Yue (2002)	66	C	PSVT:R	30	–	–	CBT	–	6	6	HS	–	0.48	0.59	0.44	0.54
	67	C	PSVT:R	30	–	–	CBT	–	71	14	UG	STEM	0.18	0.29	0.18	0.29
	68	C	PSVT:R	30	–	–	CBT	–	36	11	UG	STEM	0.81	0.35	0.79	0.35
Yue and Chen (2001)	69	C	PSVT:R	30	–	–	CBT	–	16	2	UG	STEM	0.01	0.75	0.01	0.71
	70	C	PSVT:R	30	–	–	CBT	–	19	5	UG	STEM	−0.53	0.51	−0.51	0.49

The positive effect size indicates males' higher performance on the test

ID effect size identification number; *Type* publication type (*C* Conference proceeding, *D* Dissertation, *J* Journal, *T* Thesis, *R* Technical Report);N_Q Number of questions; T_t time restriction in minutes; T_i time per item in seconds; *Mode* Administration mode; *PBT* Paper-and-pencil test; *CBT* a test was delivered by using a computer; N_M Number of males; N_F Number of females; *MS* middle school students; *HS* high school students; *UG* Undergraduate students; *BA* College graduates; *Mixed* mixed majors; *STEM* majors in Science, Technology, Engineering, and Mathematics; *Vet* Veterinary; *d* Cohen's measure of effect size; *g* Hedges' measure of effect size; SE_d standard error of *d*; SE_g standard error of *g*

[a] Half of participants had questions randomly with a reverse order

[b] Data were collected from the authors through e-mail communications

[c] Number of participants by gender were not provided, but effect sizes were calculated from F-values with df=35 and 31, respectively

[d] Participants were enrolled in a gifted program during summer

[e] A z value was used as a t value to calculate effect size

[f] Participants were mathematically talented students

model was selected as an analytical framework for the current meta-analysis. More specifically, the meta-analysis was executed in the following steps. First, histograms with independent effect sizes ($n=70$) were created. Second, the standard error was calculated and the 95 % confidence interval was established around the observed effect size (cf., Borenstein 2009). Third, the weighted average of the effect sizes was computed. The weight assigned to each effect size was an inverse of its variance (i.e., the sum of within- and between-study variance). Fourth, the chi-square test of homogeneity of effect sizes (Q test, Hedges and Olkin 1985) was conducted to see if there was a significant variation across effect sizes and to assure the choice of the random-effects model was appropriate. If there is significant variation in effect sizes, Q statistic is significant at degrees of freedoms of $k-1$, where k is the number of tested effect sizes. The I^2 statistics, proposed by Higgins et al. (2003), was also reported as an alternative index to quantify the amount of heterogeneity of effect sizes.

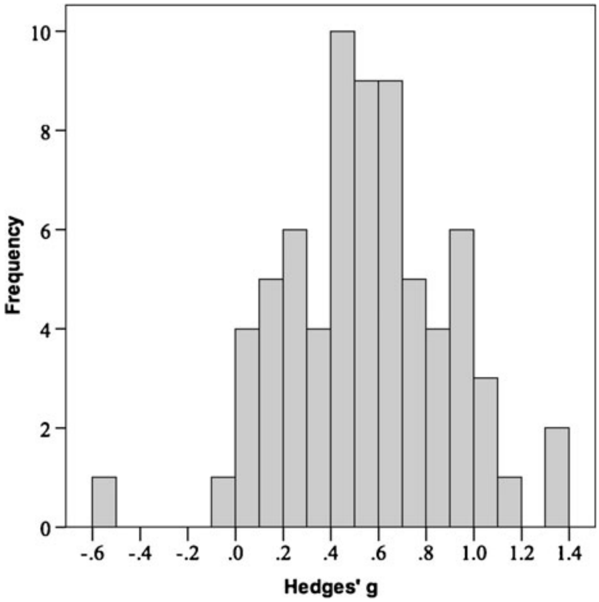
Finally, inferential analyses were conducted to identify moderators to explain the variation in effect sizes. We again used a Q test to compare the weighted average effect sizes for subgroups created by a moderator variable (Borenstein 2009). The Q, representing the weighted sum of squares for subgroups about the overall weighted average, follows a chi-square distribution with degrees of freedom of $p-1$, where p is the number of subgroups. We paid close attention to the administration conditions to explain the variation in moderator analyses, while some additional moderating effects were examined for an exploratory purpose. The moderator variables of interest include: (a) administration mode (e.g., paper-and-pencil or computer-based test), (b) length of test (number of questions), (c) time limit, and (d) manipulation of figures. Additional variables that we explored for the potential moderating effect include education levels, academic majors of primary samples, and type of documentation of studies. As expected from the findings by Voyer et al. (1995), the result of the Q test described further in the result section supported the use of a random-effects model in the subsequent analyses. Thus, we estimated the magnitude of variation (τ^2) in population effect sizes under a random-effects model, and analyzed the data with potential moderators.

Results

Heterogeneity of Effect Sizes

The histogram of Hedges' g based on all 70 effect sizes is presented in Fig. 2. It depicted a possible outlier ($g=-0.51$ reported by Yue and Chen, 2001, June), while showing positively skewed distributions. However, we included the effect size for further analyses because our close examination of the article by Yue and Chen (2001, June) did not find any distinct characteristics of the study that may explain the deviated effect size. The forest plot of 70 independent effect sizes, along with their 95 % CIs, is depicted in Fig. 3. Due to missing information of variance estimates in a primary study, the 95 % CIs of two effect sizes from Branoff and Connolly (1999) were not available; therefore, these 95 % CIs were not represented in Fig. 3 and the two effect sizes were excluded in the following analyses. The plot indicates the variation of effect sizes across studies with relatively large CIs. The chi-square test of homogeneity of effect sizes rejected the null hypothesis ($Q(67)=114.90$, $p<0.01$), implying that effect sizes do vary across studies, and it is reasonable to assume that some of the observed variation is due to variation in parameter effect sizes. Under the random-effects model, the estimated overall average effect size for gender difference was 0.57, which indicates that, on average, the PSVT:R score of a male individual is about 0.57 standard deviation higher than that of a female. Results also indicated that there was

Fig. 2 Distribution of 70 Hedges' *g* effect sizes



significant variation in parameter effect sizes, τ^2=0.035, $\chi^2(67)$=115.01, p<0.01. The I^2 statistics was 41.7 %, which indicates that the obtained effect sizes show moderate inconsistency (Higgins et al. 2003). Thus, the results warranted the need for additional analyses to explain the variation in observed effect sizes with moderators.

Moderating Effects by Test Administration Conditions and Study Design

The weighted average effect sizes by key moderators were summarized in Table 3. Note that the effect size obtained from Santone (2009) was also excluded from proceeding analyses because the researcher administered the test to participants as either online or paper–pencil formats and the statistics were not reported by the type of test delivery method, which is one of the key moderators in our analysis. The descriptive statistics indicated a possible moderation effect by one of the examined test conditions, the implementation of time limits, but not other conditions. The average effect size of gender differences measured by computer-based testing $(\overline{g} = 0.43)$ was smaller than that measured by paper-and-pencil testing $(\overline{g} = 0.60)$. The variation in effect sizes was smaller when spatial ability was measured by computer-based testing than by paper-and-pencil. However, the Q test was not significant, $\chi^2(1) = 2.89, p = 0.09$.

Originally, the PSVT:R, ROT, and PSV:R contained 30 items, 20 items, and 12 items, respectively. In addition, some researchers customized the PSVT:R by eliminating some of the items in the instrument for their research purposes. For example, Hassan and Abed (1999) and Smith (2009) administered 20 out of 30 items in the PSVT:R, while Sorby et al. (2006a) and Sorby et al. (2006b) administered 10 out of 30 items in the PSVT:R to measure the same ability. We therefore computed the weighted average effect sizes by the test length (i.e., "short" tests with 20 or less items versus "long" tests with more than 20 items) to investigate whether the use of a smaller number of items to measure the ability might affect the magnitude of the gender difference effect size. The short tests yielded a slightly larger weighted average effect size $(\overline{g} = 0.61)$ than long tests $(\overline{g} = 0.54)$. However, the

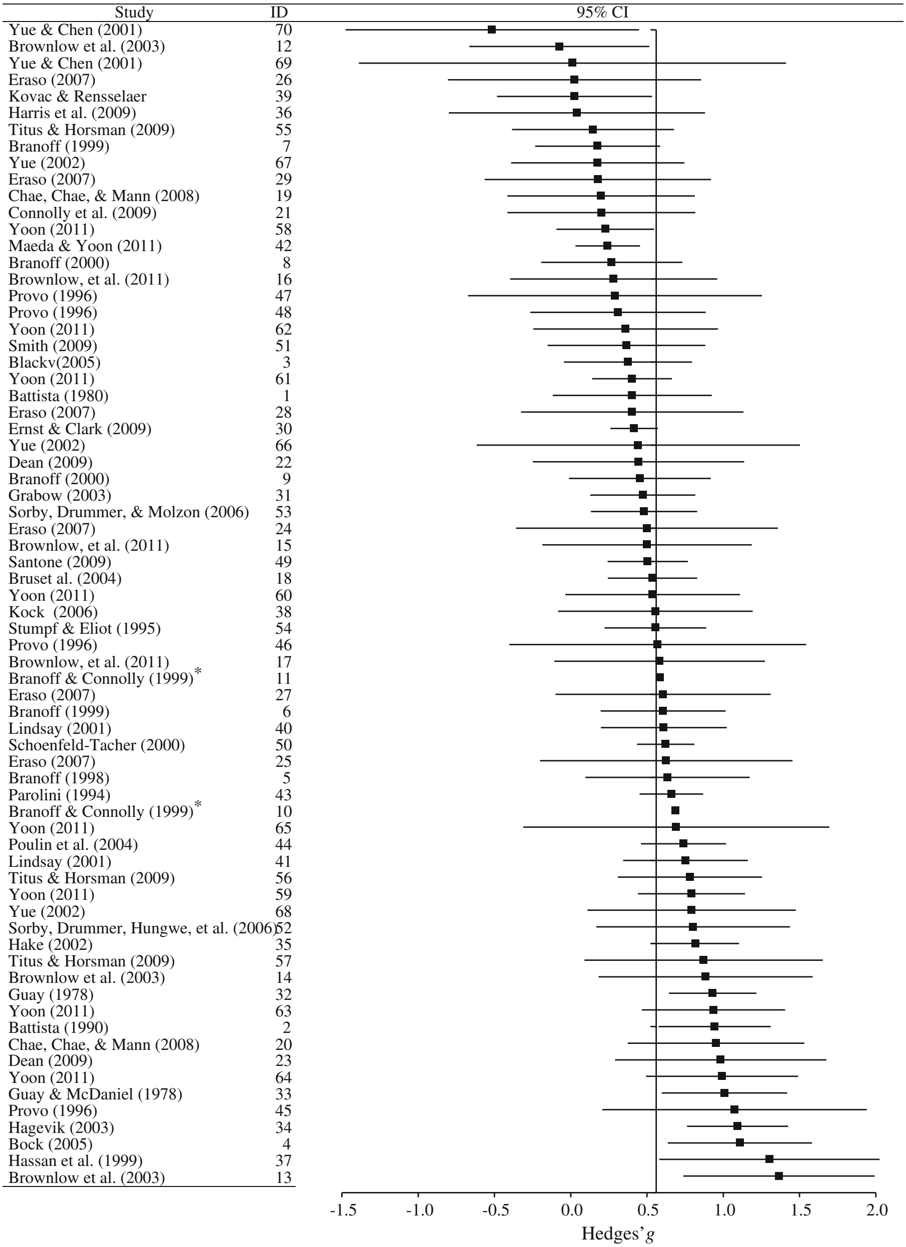


Fig. 3 Forest plot of 70 Hedges' *g* effect sizes with 95 % confidence intervals. Note. *Due to missing information of variance estimates in a primary study, the 95 % *CI*s of two effect sizes from Branoff and Connolly (1999) were not represented

difference was not statistically significant, implying that the observed variation may be attributed to sampling errors.

A close investigation of each study with a time limit for measuring the ability revealed that no systematic time restriction was implemented across these studies. For example, Eraso

Table 3 Subgroup Analyses under Random Effects by Possible Moderator Variables

Variable		*K*	Weighed	Hedges' *g*		*Q*	*df*	*p*
			N	*M*	SD			
Administration mode								
	PPT	55	629	0.602	0.341	2.885	1	0.089
	CBT	12	99	0.429	0.294			
Test length[a]								
	Short (≤20)	37	407	0.606	0.306	0.687	1	0.407
	Long (>20)	30	307	0.544	0.308			
Time limit								
	No time limit	44	406	0.570	0.330	13.553	2	0.001
	Long time limit(>30 s)	9	160	0.307	0.127			
	Short time limit (≤30 s)	14	250	0.676	0.235			
Figure modification								
	No change	63	678	0.572	0.304	0.020	1	0.888
	Changed figure	4	19	0.539	0.366			
Education level								
	Post-secondary	50	502	0.576	0.326	0.063	1	0.802
	Grade 5-12	17	162	0.554	0.302			
College major								
	Non-STEM	13	99	0.546	0.357	0.113	2	0.945
	STEM	18	167	0.588	0.321			
	Mixed	18	241	0.578	0.262			
Publication type								
	Non-journal	42	427	0.600	0.300	0.896	1	0.344
	Journal	25	254	0.525	0.325			
Total		67						

K number of effect size

[a] If the number of questions in a test is over 20, then the length of the test is long. Otherwise, the test length is short

(2007) and Santone (2009) set 10 min as the time limit to complete 12 items of the PSVT:R (50 s per question), while Dean (2009) and Lindsay (2001) set the same 10 min to complete 20 items of the ROT (30s per item). Therefore, the time limits seemed to function differently depending on the number of items to solve for a given time limit. For our moderating analysis on lime limit, we classified the effect sizes into three categories based on the length of time limit per item; no time limit, long time limit (i.e., longer than 30s per item) and short time limit (i.e., equal or shorter than 30s per item). Our observation indicates that larger gender differences seem to appear when a shorter amount of time was given to solve each item (e.g., 18, 24, or 30s per item) $(\overline{g} = 0.67)$ than when more time was given (e.g., 40, 50, or 60s per item) $(\overline{g} = 0.31)$. Under the condition where no time limit was implemented, the average effect size was 0.57. The results suggests the impact of time limits on the magnitude of gender difference, $\chi^2(2) = 13.553,\ p < 0.01$. We found no sizable difference in the weighted average effect size due to adding visual aids in figures or changing the presentation of figures in the instrument.

We further investigated the differences in the weighted average effect sizes by looking at variables related to the characteristics of samples and studies. The descriptive statistics show that the magnitude $(\overline{g} = 0.57)$ of the gender difference in post-secondary students was higher than that $(\overline{g} = 0.55)$ in students in grades 5 through 12, although the difference was marginal, $\chi^2(1) = 0.063, p = 0.80$. Unexpectedly, the weighted average effect size of non-STEM majors of college students $(\overline{g} = 0.55)$ was smaller accompanied by a larger standard deviation (SD= 0.36) than those of mixed groups of majors $(\overline{g} = 0.58)$ and STEM majors $(\overline{g} = 0.59)$. However, the differences were not statistically significant, $\chi^2(2) = 0.113, p = 0.95$. Finally, as generally expected, the gender difference effect sizes reported in published studies in journals tended to be smaller $(\overline{g} = 0.53)$ than those reported in other forms, such as dissertations, theses, and conference proceedings $(\overline{g} = 0.60)$. However, the difference was not statistically significant, $\chi^2(1) = 2.62, p > 0.05$. These results indicated that the gender differences in mental rotation ability measured by the PSVT:R and its variants seemed consistent across different educational levels and disciplines. In addition, comparable magnitudes of the gender differences were reported in both journal articles and other types of research reports.

Discussion and Conclusions

Gender Differences in the PSVT:R

Since Hedges and Nowell (1995) revealed gender differences in cognitive abilities using national probability samples, accurately documenting the directions and magnitudes of the differences in various abilities became an essential first step to elucidate the trend and the cause of these differences. Gender differences in mental rotation ability that favor males have also been frequently reported in empirical studies. However, the notion of mental rotation ability or its operationalization varies across studies when gender differences were discussed. In addition, the definitional and operational differences explained the variation in the effect sizes in previous meta-analyses (e.g., Linn and Petersen 1985; Voyer et al. 1995). While various mental rotation tests are available, the PSVT:R has been used primarily in research on educational settings, particularly in STEM disciplines. For example, the test has been predominantly used to measure students' spatial visualization ability of mental rotation in engineering education (Contero et al. 2005; Field 2007). This is because, compared to other popular tests for research in engineering education, the PSVT:R is unique in that the 3-D objects utilized in the test have inclined, oblique, and curved surfaces, which requires a higher level of spatial visualization ability compared to visualizing objects with simple cubic surfaces that are typically used in other popular tests (Yue 2004). Like other instruments for mental rotation ability, gender differences in the PSVT:R have been observed in various studies (see studies listed in Table 2).

However, no attempt has been made to integrate studies that used the PSVT:R and its variants to investigate the consistency or inconsistency of gender differences in the scores. More importantly, no study to date has systematically reviewed the influence of administration conditions on either mental rotation performance itself or the gender gap in the performance measured by the PSVT:R. The latter investigation is particularly important because the error variance or CIV resulted from the procedural differences is likely to hamper the accurate interpretation of the effect size. Therefore, by employing a meta-analytic technique, we aimed to quantify the average magnitude of gender differences in mental rotation ability and to explore the amount of variation in these effect sizes accounted for by different test administrative conditions.

Table 4 highlights and summarizes the differences between the previous meta-analytic studies on gender differences in mental rotation ability by Linn and Petersen (1985), Voyer

Table 4 Summary of Three Meta-Analytic Studies on Gender Differences

	Linn and Petersen (1985)	Voyer et al. (1995)	Current meta-analysis
Inclusion or exclusion criteria	•Published studies	•Published studies	•Published and unpublished studies
	•Conference papers	•Studies used in Maccoby and Jacklin (1974) and Linn and Petersen (1985)	•1976–2011
	•1974–1982	•1974–1993	
Technique	Random-effects model	Random-effects model	Random-effects model
Moderators	•Age of participants	•Age of participants	•Administration conditions (test length, time limit, figural edits, administration mode).
	•Types of spatial ability	•Types of spatial ability	
		•Test to measure the SA	
		•For MRT, scoring methods	•Education level, College Major, Publication type
Number of Effect Sizes	172 effect sizes	286 effect sizes	70 effect sizes
	▪Mental rotation (MR)	▪Mental rotation (MR)	Mental Rotation (MR) measured by PSVT:R
	▪Spatial perception (SP)	▪Spatial perception (SP)	
	▪Spatial visualization (SV)	▪Spatial visualization (SV)	
Findings	•MR: *d*=0.73, *k* =29, measured mostly by PMA space or MRT, which shows higher effect size	Heterogeneous effect sizes with overall *d*=0.37	Heterogeneous effect sizes with overall MR: *g*=0.57 (*SD*=0.19)
	•SP: *d*=0.44	•MR: *d*=0.56	41 % of the observed variance reflects real difference in effect sizes.
	•SV: *d*=0.13	•SP: *d*=0.44	
		•SV: *d*=0.19	
		Scoring methods of MRT could explain the heterogeneity of variance	
Conclusions	•Different types of spatial ability test (MRT vs. PMA) will produce the different level of gender difference, possibly due to the differences in cognitive and strategic processing to solve the question	•Effect sizes vary depending on the test used to measure spatial ability.	•Effect sizes vary across studies
	•Categorization of spatial ability is a key moderator (Significant for mental rotation and spatial perception)		•Mean effect sizes vary for time limits limit
			Effect sizes were consistent regardless of education level and reporting format

K number of effect size; *d* Cohen's measure of effect size; *g* Hedges' measure of effect size

et al. (1995), and this meta-analysis study in terms of research design and results. We found the weighted average effect size to be 0.57, favoring males. According to Cohen (1977), this is a moderate to large effect size indicating a meaningful gender difference in mental rotation ability. This result is consistent with the previous meta-analyses for gender differences in mental rotation ability (Linn and Petersen 1985; Voyer et al. 1995) in terms of the direction of gender difference. However, the magnitude of the effect size is smaller than those reported in these studies. For example, Linn and Petersen reported the average effect size of 0.73, with 95 % CIs ranging from 0.50 to 0.96, that favors males at age levels from around 10 to 60 years old. Voyer et al. (1995) replicated Linn and Petersen's meta-analysis by including additional recent studies and studies that utilized other spatial ability instruments and reported a weighted mean effect size of 0.67 from the studies utilizing the MRT. Nevertheless, as claimed by these studies (Linn and Petersen 1985; Voyer et al. 1995) and the results of this study, gender differences on mental rotation ability indicate the higher performance of males on mental rotation tasks and a relatively large difference in the magnitude of gender differences compared to other subtypes of spatial ability.

What is worth emphasizing is that there exists a relatively large variation in the effect sizes of gender difference across studies. The estimated variance in parameter effect sizes is 0.035 and the standard deviation is 0.19. This variation occurs not simply due to sampling errors, but also due to several systematic characteristics of the primary studies in which average mental rotation scores were reported by gender. In other words, while we expect to see higher performance of males on the PSVT:R, the magnitude of gender differences may vary depending on the study design implemented by a primary study.

For this meta-analysis, we specifically focused on factors related to test administration as moderators to explain the variation because, if test administration plays a significant role in explaining the variation, some of the observed gender differences could be attributed to procedural differences resulting from inconsistent implementation of the PSVT:R. The average effect sizes by administration conditions reported in Table 3 suggested the possible impact of testing conditions on explaining the variation in effect sizes. Among the conditions we investigated, the implementation of a time limit explained the variation in effect sizes. Our result indicates that the performance gap on the PSVT:R between male and female respondents tend to be small when they have at least more than 30 s to solve an item with time limits. Interestingly, the gender difference under the condition is smaller compared to the condition with no time limit. Consistent with the finding by Peters (2005), the gender difference was largest when a shorter time limit was set per item.

At least two possible explanations for the impact of a time limit on effect sizes are available. First, the average effect sizes might be different across the conditions (i.e., short, long, or no time limit) due to a methodological artifact. As described earlier, the variation in test administration conditions will produce CIV and the variance will affect the accuracy of scores and score interpretations by including elements that are irrelevant to the construct that the scores intended to represent (McCallin 2006). Setting a time limit could be a source for construct irrelevant variance (McCallin 2006).

When the PSVT:R was originally developed, Guay (1980) recommended using a 20-min time limit to solve 30 items (40s per item) on the test to minimize the influence of using analytic strategy to solve items. Although we do not know whether the recommended time limit is appropriate (because of lack of information provided by Guay), this recommendation has received little attention in subsequent studies using the PSVT:R. More specifically, most studies utilizing the PSVT:R rarely mentioned how they set up the time limit (if they had one), or they did not have any time limit condition.

Note that it is possible that setting a time limit may affect the performance of both groups by introducing CIV equally or unequally. For example, increasing the time to take the PSVT:R may benefit both males and females (Peters 2005) and as a result, the gender differences remain at the same magnitude, while individual scores for both groups are higher with the time limit than without any time limit. On the other hand, imposing a time limit might introduce anxiety for completing the test on time more so for females than males, which might produce a larger gender gap on performance. In fact, the result of this study supports the differential impact of setting a time limit on the test performance by gender.

Second, the study also revealed that a specific interval of time limits seems to be a critical factor that affects the magnitude of gender differences when time limits were imposed, which suggests another possible explanation for the link between the implementation of time limits and the magnitude of gender difference. More specifically, it is possible to assume that the speed of cognitive processing for problem solving ("speededness") differs by gender (faster for males than females) and the difference might affect the performance on the PSVT:R (Haladyna and Downing 2004; Lu and Sireci 2007). As a result, larger gender differences might be observed under a stringent time limit condition, while smaller gender differences might be observed under a lenient time condition. Because neither a technical manual of the PSVT:R nor a comprehensive study investigating psychometric properties of the test exist, it is not clear whether speededness is part of the construct measured by the PSVT:R or CIV. If speededness is part of the construct, a standardized time limit should be imposed when the PSVT:R is administered. Inappropriate implementation of a time limit poses a potential threat to both the reliability and validity of the scores (Lu and Sireci 2007; McCallin 2006).

However, we want to accentuate that we focused on identifying a procedural source of error variation in estimating the effect size, which would interfere with the accurate interpretation of the effect size. Our finding regarding the impact of a time limits does not suggest anything about how these time limits enhance or diminish the validity of the interpretation of the PSVT:R scores. In other words, our finding only suggests that an individual with a certain mental rotation ability would get different scores on the PSVT:R when different time limits were imposed, and the need for additional experimental studies to identify the condition where observed scores accurately reflect individual's true mental rotation ability. In particular, our literature review indicated that no investigations existed regarding how speededness relates to performance on the PSVT:R. Nor does a study exist that indicates what length of a time limit accrues the most construct irrelevant variance. Thus, further investigation on specifying an appropriate time limit to enhance the validity of inferences drawn from the PSVT:R scores is imperative in future experimental studies.

We also emphasize that the results regarding other test administrative conditions (i.e., administration mode, figure modification, and test length) should not be used to justify modifications of the test for the conditions to meet particular research purposes without additional validation. Since tests can function differently to estimate an individual's ability depending on the way in which they are administered, this would pose a threat to the validity of the score interpretation (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] 1999).

For example, the difference in administration mode did not impact the magnitude of gender difference on the PSVT:R at the group level. However, this does not mean that the PRVT:R scores at the individual level are unaffected by the administration modes. Two delivery modes of test administration, paper-and-pencil based and computer-based, may also produce CIV because the differences in test delivery modes may differently affect the test scores for both males and females. In the case of computer-based testing, respondents are assumed to have some familiarity with computer or technology in order to take the test. If computer-related issues, including less familiarity with computers, influence an individual's performance on the test, the scores do not reflect the true mental rotation ability of the individual and, as a result, accurate interpretation of the scores will be challenged. On the other hand, the use of the paper-and-pencil test may help respondents peruse the entire set of items and make them easier to return to previous items to change responses. Accordingly, the response strategy may also impact test performance.

Similarly, adding visual aids did not show a significant effect on the magnitude of the gender differences in this study. Again, it is possible that additional visual aids were equally helpful for both males and females to solve items and increase their test scores, and resulted in a non-significant effect on explaining the variation in the gender difference effect sizes. A similar interpretation can be applied to the other moderators investigated in the analysis. In sum, the non-significant effect of test administration conditions on explaining the effect size variation does not mean that participants will obtain accurate scores under various test administration conditions. It only suggests that males and females will not be affected *differently* by testing conditions to measure mental rotation ability. It is possible that all participants' test scores are similarly affected by a certain condition, and, as a result, the magnitude of the effect size remains unchanged. Therefore, it is necessary to investigate how different administration conditions affect individual test scores on the PSVT:R to fully understand the measurement properties of this instrument including score reliability and validity of inferences in future studies.

Future Directions of Research with the PSVT:R

Based on the findings of this study, it is evident that the observed variation in gender difference effect sizes is not simply due to sampling errors. There is a systematic variation in effect sizes. Although the implementation of time limits explain the variation, the large amount of variation in the gender difference effect sizes remains unexplained in this study. It is possible that other factors, such as biological, strategic, experiential, and affective (or an interaction of these factors), may contribute to the variations. Thus, a consideration of other factors is warranted for future research in order to fully understand the notion of gender differences in mental rotation ability measured by the PSVT:R.

We encountered difficulties in synthesizing the literature in mental rotation ability measured by a variant of the PSVT:R because of inconsistent reporting practices of statistical results and methodological issues in primary studies. Even though we identified more than 181empirical studies that reported some statistics of the PSVT:R scores, we could include only 40 studies in this meta-analysis because some studies did not report sufficient information to calculate Hedges' *g* effect sizes. Reporting sufficient information will help not only meta-analysts to conduct comprehensive synthesis but also any readers to understand the nature of the data and interpret the reported results accurately. Thus, we suggest that researchers follow the guidelines of reporting practice of empirical results by Wilkinson and Task Force on Statistical Inference (1999) and provide accurate and sufficient information on administration procedures and empirical results.

In addition, based on our meta-analysis, the inconsistent use of the PSVT:R in past studies was evident. As previously indicated, we urge researchers to conduct primary studies that demonstrate an appropriate instrumentation of the PSVT:R to maximize validity. This also suggests the need for establishing a standardized administration procedure of this instrument that empowers researchers to obtain scores that accurately reflect the level of mental rotation ability. This will increase the reliability of the scores and the validity of the inferences based on the scores.

In summary, this meta-analysis quantitatively synthesized the empirical research findings on gender differences observed in the PSVT:R, and found a moderating effect of time limits on gender differences in mental rotation ability, which could be considered as a source of measurement error. The findings of this study provided a benchmark regarding the expected effect sizes in mental rotation ability measured by the PSVT:R by gender, and the expected amount of the variation in the effect sizes across studies. More importantly, the study demonstrated that observed variation in effect sizes could be attributed to the unstandardized use of the PSVT:R when measuring mental rotation ability, which may threaten the validity of reported findings regarding gender difference.

References

References marked with an asterisk indicate studies included in the meta-analysis. The in-text citations to studies selected for meta-analysis are not preceded by asterisks.

Alkhateeb, H. M. (2004). Spatial visualization of undergraduate education majors classified by thinking styles. *Perceptual and Motor Skills, 98*, 865–868. doi:10.2466/PMS.98.3.865-868.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, NW: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

*Battista, M. T. (1980). Interrelationships between problem solving ability, right hemisphere processing facility and mathematics learning. *Focus on Learning Problems in Mathematics, 2*, 53–60.

*Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education, 21*, 47-60. doi:10.2307/749456.

Battista, M. T., Wheatley, G. H., & Talsma, G. (1982). The importance of spatial visualization and cognitive development for geometry learning in pre-service elementary teachers. *Journal for Research in Mathematics Education, 13*, 332–340. doi:10.2307/749007.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1973). *Differential aptitude tests*. New York, NY: The Psychological Corporation.

*Black, A. A. (2005). Spatial ability and earth science conceptual understanding. *Journal of Geoscience Education*, 53, 402-414. Retrieved from http://www.nagt.org/nagt/jge/abstracts/sep05.html#v53p402.

*Bock, A. M. (2005). *Gaze duration estimates and eye movements related to mental rotation tasks*. Unpublished master's thesis. The University of Iowa, Iowa.

Bodner, G. M., & Guay, R. B. (1997). The Purdue visualization of rotations test. *The Chemical Educator, 2*(4), 1–18. doi:10.1333/s00897970138a.

Boles, D. B. (1980). X-linkage of spatial ability: a critical review. *Child Development, 51*, 625–635.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). New York: Russell Sage.

*Branoff, T. J. (1998). The effects of adding coordinate axes to a mental rotations task in measuring spatial visualization ability: an information processing approach relating to teaching methods of undergraduate technical graphics education. *Dissertation Abstracts International*, 59(03), 709A. (UMI Number: 9825977).

*Branoff, T. J. (1999). Coordinate axes and mental rotation tasks: a Solomon four group design. *Engineering Design Graphics Journal, 63*(3), 5–14.

*Branoff, T. J. (2000). Spatial visualization measurement: a modification of the Purdue Spatial Visualization Test—Visualization of Rotations. *Engineering Design Graphics Journal, 64*(2), 14–22.

*Branoff, T. J., Connolly, P. E. (1999, June). The addition of coordinate axes to the Purdue Spatial Visualization Test—Visualization of Rotations: a study at two universities. *Proceedings of the American Society for Engineering Education* (*ASEE*) *Annual conference and expositions*, Charlotte, North Carolina.

*Brownlow, S., McPheron, T. K. Acks, C. N. (2003). Science background and spatial abilities in men and women. *Journal of Science Education and Technology, 12*, 371-380. 10.1023/B:JOST.0000006297.90536.7c.

*Brownlow, S., Janas, A. J., Blake, K. A., Rebadow, K. T., Mello, L. M. (2011). Getting by with a little help from my friends: mental rotation ability after tacit peer encouragement. *Psychology, 2*, 363-370. 10.4236/psych.2011.24057.

*Brus, C., Zhao, L., Jessop, J. (2004, June). Visual–spatial ability in first-year engineering students: a useful retention variable? *Proceedings of the American Society for Engineering Education* (*ASEE*) *Annual Conference and Expositions, Salt Lake City, Utah*.

Caplan, P. J., MacPherson, G. M., & Tobin, P. (1985). Do sex-related differences in spatial abilities exist? A multilevel critique with new data. *American Psychologist, 40*, 789–799. doi:10.1037/0003-066X.40.7.786.

Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor-analytic studies*. New York: Cambridge Univ. Press.

*Chae, Y., Chae, S., Mann, R. L. (2008, November). *Gifted Spatial Learners*: *Class selection and Gender.* Poster session presented at the Research Gala of Research and Evaluation Division at the 55th Annual Convention of the National Association for Gifted Children in Tampa, Florida.

Cherney, I. D. (2008). Mom, let me play more computer games: they improve my mental rotation ability. *Sex Roles, 59*, 776–786. doi:10.1007/s11199-008-9498-z.

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

College Entrance Examination Board. (1939). *CEEB Special Aptitude Test in spatial relations*. New York, NY: College Entrance Examination Board.

*Connolly, P., Harris, L. V. A., Sadowski, M. (2009, June). Measuring and enhancing spatial visualization in engineering technology students. *Proceedings of the American Society for Engineering Education* (*ASEE*) *Annual conference and expositions*, Austin, Texas.

Contero, M., Naya, F., Company, P., Saorin, J. L., & Conesa, J. (2005). Improving visualization skills in engineering education. *Computer Graphics in Education, 25*(5), 24–31.

Cooke-Simpson, A., & Voyer, D. (2007). Confidence and gender differences on the Mental Rotations Test. *Learning and Individual Differences, 17*, 181–186. doi:10.1016/j.lindif.2007.03.009.

*Dean, D. H. (2009). Spatial visualization and the gender gap in videogame interest among young adults. *Young Consumers*: *Insight and Ideas for Responsible Marketers, 10*, 225–237.

Deno, J. (1995). The relationship of previous experiences to spatial visualization ability. *Engineering Design Graphics Journal, 59*, 5–17. doi:10.1108/17473610910986035.

Educational Testing Service (2009). Educational Testing Service Test Collection. Retrieved from http://204.50.92.130/ETS_Test_Collection/Portal.aspx?lang=en-US.

Eliot, J. (1987). *Models of psychological space: psychometric, developmental, and experimental approaches*. New York: Springer.

*Eraso, M. (2007). *Connecting visual and analytic reasoning to improve students' spatial visualization abilities: A constructivist approach*. Unpublished doctoral dissertation, Florida International University, Florida.

*Ernst, J. V., Clark, A. C. (2009). Technology-based content through virtual and physical modeling: a national research study. *Journal of Technology Education, 20*(2), 23–36.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science, 18*, 850–855. doi:10.1111/j.1467-9280.2007.01990.x.

Field, B. W. (2007). Visualization, intuition, and mathematics metrics as predictors of undergraduate engineering design performance. *Journal of Mechanical Design, 129*, 735–743. doi:10.1115/1.2722790_.

Geiser, C., Lehmann, W., & Eid, M. (2006). Separating "rotators" from "non-rotators" in the Mental Rotations Test: a multigroup latent class analysis. *Multivariate Behavioral Research, 41*, 261–293. doi:10.1207/s15327906mbr4103_2.

Geiser, C., Lehmann, W., & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence, 36*, 556–563. doi:10.1016/j.intell.2007.12.003.

Ginn, S. R., & Pickens, S. J. (2005). Relationships between spatial activities and scores on the mental rotation test as a function of sex. *Perceptual and Motor Skills, 100*, 877–881. doi:10.2466/PMS.100.3.877-881.

Goldstein, D., Haldane, D., & Mitchell, C. (1990). Sex differences in visual-spatial ability: the role of performance factors. *Memory & Cognition, 18*, 546–550.

Gohm, C. L., Humphreys, L. G., & Yao, G. (1998). Underachievement among spatially gifted students. *American Educational Research Journal, 35*, 515–531.

*Grabow, R. (2003). *The relationship of visual spatial ability to performance in solving stoichiometric problems in a high school chemistry class*. Unpublished masters' thesis, California State University, California.

Guay, R. B. (1976). *Purdue spatial visualization test*. West Lafayette, IN: Purdue Research Foundation.

*Guay, R. B. (1978). *Factors affecting spatial test performance*: *Sex, handedness, birth order, and experience.* Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, CA (ERIC Document Reproduction Service No. ED167612).

Guay, R. B. (1980). *Spatial ability measurement*: *a critique and an alternative*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED189166).

*Guay, R. B., McDaniel, E. (1978). Correlates of performance on spatial aptitude tests. (A final report on Grant No. DAHC 19-77-G-0019) Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Guay, R. B., McDaniel, E., & Angelo, S. (1978). Analytic factors confounding spatial ability measurement. Paper presented at the annual convention of the American Psychological Association, Toronto, CA. In R.

B. Guay & E. McDaniel (Eds.), *Correlates of performance on spatial aptitude tests*. Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

*Hagevik, R. A., (2003). *The effects of online science instruction using geographic information systems to foster inquiry learning of teachers and middle school science students*. *Dissertation Abstracts International*, 64(10), 3635A. (UMI Number: 3107767).

*Hake, R. R. (2002). *Relationship of individual student normalized learning gains in Mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization*. Poster session presented at the Physics Education Research Conference (PERC), Boise, Idaho.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*, 17–27. doi:10.1111/j.1745-3992.2004.tb00149.x.

Harris, L. J. (1978). Sex differences in spatial ability: Possible environmental, genetic, and neurological factors. In M. Kinsbourne (Ed.), *Asymmetrical function of the brain* (pp. 405–522). New York: Cambridge University Press.

*Harris, M. A., Peck, R. P., Colton, S., Morris, J., Neto, E. C., Kallio, J. (2009). A combination of hand-held models and computer imaging programs helps students answer oral questions about molecular structure and function: A controlled investigation of student learning. *CBE Life Sciences Education, 8*, 29-43. 10.1187/cbe.08–07–0039.

*Hassan, M. M., Abed, A. S. (1999). Differences in spatial visualization as a function of scores on hemisphericity of mathematics teachers. *Perceptual and motor skills, 88*, 387-390. 10.2466/PMS.88.2.387-390.

Hausmann, M., Slabbekoorn, D., Van Goozen, S. H. M., Cohen-Kettenis, P. T., & Güntürkün, O. (2000). Sex hormones affect spatial abilities during the menstrual cycle. *Behavioral Neuroscience, 114*, 1245–1250. doi:10.1037/0735-7044.114.6.1245.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128. doi:10.3102/10769986006002107.

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

Hedges, L. V., & Olkin, I. (1985). *Statistical model of meta-analysis*. New York: Academic.

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal, 327*, 557–560. doi:10.1136/bmj.327.7414.557.

Hirnstein, M., Bayer, U., & Hausmann, M. (2009). Sex-specific response strategies in mental rotation. *Learning and Individual Differences, 19*, 225–228. doi:10.1016/j.lindif.2008.11.006.

Höffler, T. N. (2010). Spatial ability: Its influence on learning with visualizations—a meta-analytic review. *Educational Psychology Review, 245–269*. doi:10.1007/s10648-010-9126-7.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16–25. doi:10.1111/j.1745-3992.2001.tb00066.x.

Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership and the role of spatial visualization in becoming an engineer, physical scientist, or artist. *Journal of Applied Psychology, 78*, 250–261.

Jordan, K., Wüstenberg, T., Heinze, H.-J., Peters, M., & Jäncke, L. (2002). Women and men exhibit different cortical activation patterns during mental rotation tasks. *Neuropsychologia, 40*, 2397–2408. doi:10.1016/S0028-3932(02), 00076-3.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: accounts of mental rotation and individual differences in spatial ability. *Psychological Review, 92*, 137–172.

*Koch, D. S. (2006). *The effects of solid modeling and visualization on technical problem solving*. Unpublished doctoral dissertation, The Virginia Polytechnic Institute and State University, Virginia.

Koscik, T., O'Leary, D., Moser, D. J., Andreasen, N. C., & Nopoulos, P. (2009). Sex differences in parietal lobe morphology: relationship to mental rotation performance. *Brain and Cognition, 69*, 451–459. doi:10.1016/j.bandc.2008.09.004.

*Kovac, R. J. Rensselaer, B. (1989). The validation of selected spatial ability tests via correlational assessment and analysis of user-processing strategy. *Educational Research Quarterly, 13*(2), 26–35.

*Lindsay, H. A. (2001). *Factors related to achievement in sophomore organic chemistry at the University of Arkansas*, Unpublished Doctoral dissertation, University of Arkansas.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Development, 56*, 1479–1498. doi:10.2307/1130467.

Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 147–158). New York: Sage.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*(4), 29–37. doi:10.1111/j.1745-3992.2007.00106.x.

Lohman, D. F. (1996). Spatial ability and G. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: their nature and measurement* (pp. 97–116). Hillsdale, NJ: Erlbaum.

Lubinski, D. (2010). Spatial ability and STEM: a sleeping giant for talent identification and development. *Personality and Individual Differences, 49*, 344–351. doi:10.1016/j.paid.2010.03.022.

*Maeda, Y., Yoon, S. Y. (2011). Scaling the Revised PSVT-R: Characteristics of the first year engineering students' spatial ability. *Proceedings of the American Society for Engineering Education* (*ASEE*) *Annual Conference and Exposition, 2011-2582, Vancouver, BC, Canada*.

Masters, M. S. (1998). The gender difference on the Mental Rotations Test is not due to performance factors. *Memory & Cognition, 26*, 444–448.

McCallin, R. C. (2006). Test administration. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 625–652). Mahwah, NJ: Erlbaum.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215–237.

Moè, A. (2009). Are males always better than females in mental rotation? Exploring a gender belief explanation. *Learning and Individual Differences, 19*, 21–27. doi:10.1016/j.lindif.2008.02.002.

Moè, A., & Pazzaglia, F. (2006). Following the instructions: effects of gender beliefs in mental rotation. *Learning and Individual Differences, 16*, 369–377. doi:10.1016/j.lindif.2007.01.002.

Monahan, J. S., Harke, M. A., & Shelley, J. R. (2008). Computerizing the Mental Rotations Test: are gender differences maintained? *Behavior Research Methods, 40*, 422–427. doi:10.3758/BRM.40.2.422.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105–125. doi:10.1037/1082-989X.7.1.105.

Moses, B. E. (1977). *The nature of spatial ability and its relationship to mathematical problem-solving*. Indiana University, IN: Unpublished doctoral dissertation.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures*. Thousand Oaks, CA: Sage.

Ortner, T. M., & Sieverding, M. (2008). Where are the gender differences? Male priming boosts spatial skills in women. *Sex Roles, 59*, 274–281. doi:10.1007/s11199-008-9448-9.

*Parolini, L. L. (1994). *Gender differences on predictors of success on the Purdue Spatial Visualization Test*: *Rotations*. Unpublished master's thesis, Michigan Technological University, Michigan.

Peters, M. (2005). Sex differences and the factor of time in solving Vandenberg and Kuse mental rotation problems. *Brain and Cognition, 57*, 176–184. doi:10.1016/j.bandc.2004.08.052.

Peters, M., Laeng, B., Lathan, K., Jackson, M., Zaiouna, 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.

*Poulin, M., O'Connell, R. L., Freeman, L. M. (2004). Picture recall skills correlate with 2D:4D ratio in women but not men. *Evolution and Human Behavior, 25*, 174-181. 10.1016/j.evolhumbehav.2004.03.004.

*Provo, J. A. (1996). *The effect of examination of a cross section on students' ability to visualize anatomy in three dimensions*. Unpublished master's thesis, Purdue University, Indiana.

Quaiser-Pohl, C., Geiser, C., & Lehmann, W. (2006). The relationship between computer-game preference, gender, and mental-rotation ability. *Personality and Individual Differences, 40*, 609–619. doi:10.1016/j.paid.2005.07.015.

*Santone, A. (2009). *Visuospatial characterization and analysis of spatial ability of video game players*. Unpublished doctoral dissertation, Purdue University, IN.

*Schoenfeld-Tacher, R. M. (2000). *Relation of student characteristics to learning of basic biochemistry concepts from a multimedia goal-based scenario*. Unpublished doctoral dissertation, University of Northern Colorado.

Sharps, M. J., Price, J. L., & Williams, J. K. (1994). Spatial cognition and gender: instructional and stimulus influence on mental image rotation performance. *Psychology of Women Quarterly, 18*, 413–425. doi:10.1111/j.1471-6402.1994.tb00464.x.

Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: a 20-year longitudinal study. *Journal of Educational Psychology, 93*, 604–614. doi:10.1037//0022-0663.93.3.604.

Shepard, R. N. (1978). Externalization of mental images and the act of creation. In B. S. Randhawa & W. E. Coffman (Eds.), *Visual learning, thinking, and communication* (pp. 133–190). New York: Academic.

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

*Smith, M. E. (2009). *The correlation between a pre-engineering student's spatial ability and achievement in an electronics fundamentals course*. Unpublished doctoral dissertation, Utah State University, UT.

Sorby, S. A. (2000). Spatial abilities and their relationship to effective learning of 3-D solid modeling software. *Engineering Design Graphics Journal, 64*(3), 30–35.

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., & Baartmans, B. J. (1996). A course for the development of 3-D spatial visualization skills. *Engineering Design Graphics Journal, 60*(1), 13–20.

Sorby, S. A., & Baartmans, B. J. (2000). The development and assessment of a course for enhancing the 3-D spatial visualization skills of first year engineering students. *Journal of Engineering Education, 89*, 301–307.

*Sorby, S. A., Drummer, T., Hungwe, K., Parolini, L., Molzon, R. (2006a, July). Preparing for engineering studies: Improving the 3-D spatial skills of K-12 students. *Proceedings of 9th International Conference on Engineering Education* (*ICEE*), San Juan, Puerto Rico.

*Sorby, S. A., Drummer, T., Molzon, R. (2006b). Experiences in using spatial skills testing instruments with younger audiences. *Journal for Geometry and Graphics, 10*, 227–235.

Stieff, M. (2007). Mental rotation and diagrammatic reasoning in science. *Learning and Instruction, 17*, 219–234. doi:10.1016/j.learninstruc.2007.01.012.

Strong, S., & Smith, R. (2001/2002). Spatial visualization: Fundamentals and trends in engineering graphics. *Journal of Industrial Technology, 18*, 1–6.

Strube, M. J. (1987). A general model for estimating and correcting the effects of non-independence in meta-analysis. *Multiple Linear Regression Viewpoints, 16*, 40–47.

*Stumpf, H., Eliot, J. (1995). Gender-related differences in spatial ability and the k factor of general spatial ability in a population of academically talented students. *Personality and Individual Differences, 19*, 33–45. 10.1016/0191-8869(95)00029-6.

*Titus, S., Horsman, E. (2009). Characterizing and improving spatial visualization skills. *Journal of Geoscience Education, 57*, 242–254.

Titze, C., Heil, M., & Janse, P. (2008). Gender differences in the Mental Rotations Test (MRT) are not due to task complexity. *Journal of Individual Differences, 29*, 130–133. doi:10.1027/1614-0001.29.3.130.

Thomas, H., & Kail, R. (1991). Sex differences in speed of mental rotation and the X-linked genetic hypothesis. *Intelligence, 15*, 17–32. doi:10.1016/0160-2896(91)90020-E.

Thomsen, T., Hugdahl, K., Ersland, L., Barndon, R., Lundervold, A., Smievoll, A. I., Roscher, B. E., & Sundberg, H. (2000). A functional magnetic resonance imaging (fMRI) study of sex differences in a mental rotation task. *Medical Science Monitor, 6*, 1186–1196.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations: a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47*, 599–604. doi:10.2466/PMS.47.6.599-604.

Voyer, D., Rodgers, M., & McCormick, P. A. (2004). Timing conditions and the magnitude of gender differences on the Mental Rotations Test. *Memory & Cognition, 32*, 72–82.

Voyer, D., & Saunders, K. A. (2004). Gender differences on the mental rotations test: a factor analysis. *Acta Psychologica, 117*, 74–94. doi:10.1016/j.actpsy.2004.05.003.

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. doi:10.1037/0033-2909.117.2.250.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: aligning over fifty years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*, 817–835. doi:10.1037/a0016127.

Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 159–176). New York: Sage.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist, 54*, 594–604. doi:10.1037/0003-066X.54.8.594.

Wood, J. A. (2008). Methodology for dealing with duplicate study effects in a meta-analysis. *Organizational Research Methods, 11*, 79–95. doi:10.1177/1094428106296638.

*Yoon, S. Y. (2011). *Psychometric properties of the Revised Purdue Spatial Visualization Tests*: *Visualization of Rotations* (*The Revised PSVT*:*R*) (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses. (Order Number: 3480934).

*Yue, J. (2002, June). Spatial visualization skills at various educational levels. *Proceedings of the American Society for Engineering Education* (*ASEE*) *Annual Conference and Expositions*, Montréal, Quebec, Canada.

Yue, J. (2004, June). Spatial visualization by orthogonal rotations. *Proceedings of the American Society for Engineering Education* (*ASEE*) *Annual Conference and Expositions*, Salt Lake City, Utah.

Yue, J. (2006, October). Spatial visualization by isometric drawing. Proceedings of the2006 IJME-INTERTECH Conference, Union, New Jersey.

*Yue, J. Chen, D. M. (2001, June). Does CAD improve spatial visualization ability? *Proceedings of the American Society for Engineering Education* (*ASEE*) *Annual Conference and Expositions, Albuquerque, New Mexico*.