Learning and Individual Differences 45 (2016) 176–184

Contents lists available at ScienceDirect

Learning and Individual Differences

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

Sex difference in spatial ability for college students and exploration of measurement invariance☆

Xiaoying Xu [a], Eun Sook Kim [b], Jennifer E. Lewis [c,*]

[a] Department of Chemistry, University of South Florida, United States
[b] Department of Educational and Psychological Studies, University of South Florida, United States
[c] Department of Chemistry and Center for the Improvement of Teaching and Learning in Undergraduate STEM Education, University of South Florida, United States

ARTICLE INFO

Article history:
Received 31 July 2014
Received in revised form 2 July 2015
Accepted 26 November 2015

Keywords:
Spatial ability
Measurement invariance
Confirmatory factor analysis
Sex gap

ABSTRACT

This study investigates psychometric evidence for a spatial ability instrument with a focus on measurement invariance and sex difference in this domain. The Purdue Visualization of Rotations (ROT) test was administered to science major students enrolled in a college chemistry course to measure mental rotation ability. Confirmatory factor analysis (CFA) was used to evaluate the fit of three alternative models. A bi-factor model is a better fit for the data than other models (1-factor, 4-factor 2nd-order), which indicates that the probability of answering a test item correctly is affected not only by student ability but also by an item-writing pattern. Results from multiple-group CFA support the measurement invariance of the bi-factor model between sexes and the outperformance of males over females with medium to large effect sizes. In conclusion, the ROT is appropriate to measure student mental rotation ability in college classrooms and to detect a difference between sexes.

Spatial ability has been proposed as important for higher education including astronomy, geology, chemistry and engineering (Black, 2005; Liben, Kastens, & Christensen, 2011; Lord, 1985; Pribyl & Bodner, 1987; Wai, Lubinski, & Benbow, 2009). For example, geologists need to think about the configuration and transformation of rock layers, and to read a plethora of graphs and maps. Students without sufficient spatial skills need aids or extra effort to understand space-related tasks. Currently, there are ongoing efforts to train college students' spatial skills to improve performance in the Science, Technology, Engineering, and Mathematics (STEM) areas at the college level. Miller and Halpern (2013) found the training of spatial ability improved student introductory physics scores in a one-year study for college students. Sorby and Baartmans (1996) developed a ten-week course to train spatial thinking skills in the early college engineering curriculum. In the development of spatial ability training programs, researchers need robust measurement tools that align with instructional goals in order to better understand the effectiveness of those programs.

Spatial ability has been defined as the ability to "mentally manipulate, rotate, twist, or invert a pictorially presented stimulus object" (McGee, 1979). In studying spatial ability, researchers use terms that could mean either overall spatial ability or one aspect of spatial ability, such as spatial visualization (Salthouse, Babcock, Skovronek, Mitchell, & Palmon, 1990; Sorby, 1999), mental rotation (Shepard & Metzler, 1971; Stieff, 2007), or spatial orientation (Merchant, Goetz, & Keeney-Kennicutt, 2012; Tartre, 1990). Therefore, there is a need to distinguish between specific spatial ability skills, so that researchers can describe more clearly what is being studied. Lohman (1979) proposed a framework with major and minor components such as spatial relations, spatial orientation, visualizations and transformations (Harle & Towns, 2011; Lohman, 1979). In their review of the literature, Linn and Petersen (1985) instead suggested three categories: spatial perception, mental rotation and spatial visualization. The spatial relations component in Lohman's framework and the mental rotation category proposed by Linn and Petersen both concern tasks requiring mentally rotating an object in space. When such a task involves the object as a whole without any change in shape, there is really no distinction between the two (Sorby, 1999). The present study can therefore be said to relate to spatial relations or mental rotations. We will use the latter term.

Because the mental rotation component of spatial ability is one of the cognitive areas with the most persistent sex gap favoring males (Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995; Weckbacher & Okamoto, 2014), it has been posited as a potential reason for women's underperformance and underrepresentation in STEM areas (Ceci, Williams, & Barnett, 2009; Hill, Corbett, & St. Rose, 2010). Recently, meta-analytic techniques have revealed significant national variability in performance differences between female and male middle school

☆ This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
* Corresponding author at: Department of Chemistry, 4202 E. Fowler Avenue CHE205, Tampa, Florida 33620, United States.
E-mail address: jennifer@usf.edu (J.E. Lewis).

http://dx.doi.org/10.1016/j.lindif.2015.11.015
1041-6080/

students in math (Else-Quest, Hyde, & Linn, 2010) and science (Nosek et al., 2009), implicating societal factors rather than innate ability. Females are still underperforming in college level science in the United States, even with interventions intended to address cultural factors (Miyake et al., 2010) and are underrepresented in STEM careers worldwide (Blickenstaff, 2005), so there remains a need for researchers to understand whether the observed differences in spatial ability between the sexes are significant with respect to this problem. Unfortunately, studies on possible test bias with respect to sex have been rarely reported for spatial ability instruments, with some notable exceptions (e.g., Arendasy & Sommer, 2010; Bors & Vigneau, 2011), although some interesting recent work with respect to mental processing speed has ruled out this factor as a contributor to the persistent sex difference in performance on a timed mental rotation task (Debelak, Gittler, & Arendasy, 2014). Investigating possible test bias is critical for determining whether an observed sex difference stems from test bias (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999). The Purdue Visualization of Rotations Test (ROT) is chosen in this study because it is frequently used to examine the relationship of spatial ability with achievement, affective variables, and/or sex differences in the context of STEM courses (Bodner & Guay, 1997; Brownlow, McPheron, & Acks, 2003; Carter, LaRussa, & Bodner, 1987; Morgil, Yavuz, Oskay, & Arda, 2005; Southam & Lewis, 2013), there is no existing study of possible test bias for this instrument, and the instrument provides a way to measure mental rotation ability under normal classroom conditions. While computerized mental rotation tests have been developed (Arendasy & Sommer, 2010), the need for paper and pencil tests, such as ROT, that can be administered under normal classroom conditions remains.

ROT is chosen also because it has accumulated some psychometric evidence of validity and reliability. Educators need measurement tools that are both practical and robust in order to facilitate better understanding and training of spatial ability in undergraduate courses. According to the *Standards for Educational and Psychological Testing*, one source of validity based on response processes is whether test takers apply the intended strategy or use unintended approaches to solve the items (American Educational Research Association et al., 1999). The administration conditions for ROT are designed to minimize the use of alternate analytical strategies. The test takers are required to finish all 20 items in 10 min, and they are not allowed to make any mark on the booklet; this reduces the possibility of using analytical reasoning strategies. A meta-analysis of studies using ROT and similar instruments has revealed these tests usually favor males and has suggested that the time limit plays a role (Maeda & Yoon, 2013). Second, confirmatory factor analysis has been used to support a one-factor structure (Yoon, 2011). Additionally, internal consistency reliability has been reported for the ROT at multiple sites with Cronbach's alpha values around .8 (Battista, Wheatley, & Talsma, 1982; Sorby & Baartmans, 1996; Yoon, 2011). However, there is no report on alternate possible models or measurement invariance. In addition, little is known about the temporal stability of test scores of ROT.

One goal of this study is to gather more evidence as illustrated in the *Standards* and previous studies for validity and reliability of ROT scores so that instructors can use the ROT as an appropriate measure in college classrooms (American Educational Research Association et al., 1999; Arjoon, Xu, & Lewis, 2013; Cook, Zendejas, Hamstra, Hatala, & Brydges, 2014; Frederickson & Petrides, 2008; Freund & Holling, 2011; Markon, Chmielewski, & Miller, 2011). The other is to establish measurement invariance for the group mean comparison across sexes for a student sample enrolled in a college chemistry course. When a test is utilized to compare groups, it is important to establish measurement invariance, i.e., the test should measure the latent variable in the same way across groups (American Educational Research Association et al., 1999; Arjoon, Xu, & Lewis, 2013; Kim & Yoon, 2011). Without this information, it is premature to draw a conclusion regarding whether ROT's tendency to favor males is related to a true performance difference or simply to test bias. Specifically, three research questions guide this study:

(1) How stable is the ROT score when students retake the test?
(2) How similarly do ROT items function between the sexes?
(3) What is the sex difference, if any, as measured by the ROT test?

1. Research methods

1.1. Instruments

In this study, ROT (Bodner & Guay, 1997) was used to measure the mental rotations component of spatial ability. This 20-item paper-and-pencil test asks students to discern how an example object has been rotated, and to choose from five drawings how a different object will look when the same rotation is applied. The ROT is derived from the Purdue Spatial Visualization Tests: Visualization of Rotations, which was based on the Shepard-Metzler test as adapted by Vandenberg (Shepard & Metzler, 1971; Vandenberg & Kuse, 1978). An example of a ROT test item is illustrated in Fig. 1. The correct answer is A.

1.2. Participants and data collection

ROT was given to students enrolled in a college chemistry course for science majors during the second week of four semesters (Fall 2008, Spring 2009, Fall 2009, and Spring 2010) at a large Southeastern public research university in the United States. The administrators followed the test procedure described in the literature (Bodner & Guay, 1997). There are 4168 ROT records in total. The ROT data were examined for each semester separately to establish psychometric evidence concerning whether the test functioned similarly across time and sex. The data were combined only to track students who retook the course and the ROT. Demographic information such as sex, major, and SAT math scores was obtained via the registrar. The top five majors in which students ultimately graduated ($n = 986$) were biomedical sciences (36%), biology (19%), health sciences (11%), psychology (5%) and public health (4%).

In order to give a general idea of how the demographic distribution varies across semesters, Table 1 presents the number and SAT Math scores for both females and males of each semester. Sexes distribute similarly across semesters, with more females than males. Males tended to score slightly higher than females on SAT Math, with effect sizes quantified as medium according to Cohen's d effect size guidelines (Cohen, 1988). Because this difference in math ability may confound the spatial ability comparison, the SAT math score will be included in the analysis.

1.3. Statistical data analysis

1.3.1. Descriptive statistics

Descriptive analysis was performed including mean, standard deviation, skewness and kurtosis for the ROT total score. A reliability estimate was calculated by Cronbach's alpha to reflect how items function consistently to assess the intended construct. Cronbach's alpha of greater than 0.7 is considered as above the satisfactory level for research purposes (Murphy & Davidshofer, 2005).

1.3.2. Test–retest relationship for temporal stability

The 309 students who retook the General Chemistry I course and thus had more than one ROT score were used to track score change and correlation. Note that a typical way to measure test–retest reliability is to give a test to a representative sample within an appropriate time interval such as two weeks (Meyer, 2010). The test–retest relationship here is a convenient but not an ideal way to investigate temporal