The Validity and Reliability of Online Testing for the Assessment of Spatial Ability

Jeffrey Buckley, Dr Niall Seery and Dr Donal Canty
Department of Design and Manufacturing Technology
University of Limerick

Abstract

The assessment of spatial skills offers significant insight into cognitive capacities associated with disciplines such as graphics, engineering and design. The operationalization of this assessment is typically seen in the format of paper and pencil based tests. However, aligning with pertinent technological advances, a paradigm shift can be seen in the exploration of computer based online assessment. While research has identified a number of limitations to this approach, the use of computer based assessment merits recognition, especially as technology becomes increasingly integrated into modern society. This paper investigates the validity and reliability of online testing in the assessment of spatial skills. A study cohort (n=162) of 1st year post-primary pupils piloted a test center which consisted of digital versions of three spatial ability tests. Performance scores were compared with a national sample from a similar demographic who utilized paper based versions of the tests. Results indicate no statistically significant difference between modalities and suggest the applicability of expedited tests for larger cohorts while the full tests appear more suitable for individual results.

Introduction

Increases in the capacity of modern technology have resulted in a paradigm shift towards the exploration of online assessment as an alternative to traditional paper based assessments. The use of online ICT infrastructures for assessment affords opportunities such as collaborative peer assessment, a more diverse range of potential test items, and customization of feedback mechanisms. However, transitioning from or between paper based and online mediated assessment also presents a variety of considerations and limitations, particularly when the intent of the assessment architecture is to elicit levels of cognitive factors independent of semantic knowledge.

Various factors are suggested to influence performance on computer based assessment. For example, McDonald (2002) argues that individual differences such as computer anxiety and experience can have a negative effect however he also acknowledges that such factors are not static, suggesting their potential alleviation in conjunction with increases of computer integration into society. Considering this, when comparing performances between paper and pencil based assessments with computer based assessments it is advocated that mean scores and variances of tests taken should be identical across modalities (Wilson, Genco, & Yager, 1985).

Specifically in relation to the online assessment of spatial factors, the nature of the specific factor must be taken into account. Larson (1996) describes a widely acknowledged continuum for the positioning of spatial factors. One extreme is characterized by factors pertinent to cognitive speed such as the *speeded rotation* factor while the other is embodied by factors associated with cognitive power such as *spatial relations* and *visualization*. Mead and Drasgow (1993) identified differences between traditional and computer mediated assessment when tests were based on speed of response while Veurink and Hamlin (2015) found differences across both methods for the Purdue Spatial Visualization Test: Visualization of Rotations (PSVT:R) (Guay, 1977), a measure of spatial relations at the power end of the continuum.

While these findings offer significant insight into the use of computer mediated and online assessment, it is important that educational systems align with societal progression. While the use of computers has previously been a novel experience and may continue to be so at present, McDonald's (2002) acknowledgement of the potentially dynamic nature of pertinent individual differences merits recognition as computers transition from being novel to a standard convention. Aligning with this idea, this paper presents the initial developments of an online test center for spatial ability which is ultimately envisioned to have the capacity to capture a person's *spatial profile* (Buckley & Seery, 2016) through the provision of valid and reliable tests for multiple spatial factors.

Method

This study aimed to instigate the development of an online test center for spatial skills and was conducted in conjunction with the SPACE project, a national project examining spatial skills across all stages of the Irish educational system. The initial phase of the project aimed to generate a national spatial profile of pupils entering post-primary education through the administration of expedited versions of the PSVT:R (Guay, 1977), the Mental Cutting Test (MCT) (CEEB, 1939), and the Space Relations subtest of the Differential Aptitude Test (DAT:SR) (Bennett, Seashore, & Wesman, 1973). The online test center was trialed in sample of participating schools (Table 1) and contained full versions of the PSVT:R and the MCT with the expedited version of the DAT:SR. The tests were administered either by a member of the research team or by a participating teacher who received pertinent training prior to the study with the order of administration being randomized so as to avoid inducing an order bias. This protocol was also followed in the wider national study. When designing the tests, consideration was given to emulating the paper-based versions by not restricting the question order and allowing answers to be reviewed while still within the time limit. This was to cater for the results of Shermis and Lombard (1998) who found that such restrictions can induce anxiety.

Table 1. Demographic information for study cohort

School	Location	Size	Cohort Age		Cohort Gender		n
			Mean	Std. Deviation	Male	Female	
SCH1	Urban	≈550	13.40	0.88	46	15	61
SCH2	Urban	≈500	13.07	0.47	22	4	26
SCH3	Urban	≈700	12.83	0.39	19	4	23
SCH4	Rural	≈180	13.05	0.38	11	11	22
SCH5	Urban	≈850	13.26	0.73	14	5	19
SCH6	Urban	≈650	12.91	0.54	6	3	11

Findings

Considering the national study utilized expedited versions of the full tests, the intent of this analysis was to examine potential differences between the scores on the full tests and the participants' scores on ten specific items which align with the expedited versions. Descriptive statistics for these measures are presented in Table 2.

Table 2. Descriptive statistics for psychometric tests

	n	Min	Max	Mean	Std. Deviation	Skewness		Kurtosis	
						Statistic	Std. Error	Statistic	Std. Error
PSVT:R	116	0.00	76.67	31.29	17.32	.694	.225	-.099	.446
10 Item PSVT:R	116	0.00	100.00	32.93	20.60	.599	.225	.043	.446
MCT	55	8.00	44.00	23.35	8.97	.335	.322	-.426	.634
10 Item MCT	55	0.00	80.00	26.00	15.23	1.436	.322	2.841	.634
DAT:SR	73	0.00	70.00	37.12	19.04	-.145	.281	-.775	.555

To further this analysis, paired samples t-tests were conducted between performance on the full tests and performance on the ten extracted items. No statistical significance was found between the PSVT:R (M = 31.29, SD = 17.32) and the 10 item PSVT:R (M = 32.93, SD = 20.60), t (115) = -1.587, p = .117, or between the MCT (M = 23.35, SD = 8.97) and the 10 item MCT (M = 26.00, SD = 15.23), t (54) = -1.914, p = .061. High and statistically significant correlations were also found between the full PSVT:R and the 10 extracted items (r = .840, p < .001) and between the full MCT and the 10 extracted items (r = .756, p < .001).

While these results advocate for the substitution of the full tests with the expedited versions, it is worth noting the variances in maximum performances across these measures as well as the high skewness and kurtosis values for the extracted MCT items. In addition to this, the variances between individual participants scores on both versions of each test were also examined and showed a maximum variance for the PSVT:R tests as 33.33% with a maximum of 40% for the MCT tests. Cronbach's Alpha values were also found for each of the measures and were; PSVT:R

(α = .802), 10 item PSVT:R (α = .554), MCT (α = .147), 10 item MCT (α = .236), and DAT:SR (α = .452), which identify the full PSVT:R as the only sufficiently reliable measure with this cohort (Kline, 2000).

In addition to this, to examine the use of online spatial skills assessment in comparison with paper based assessment, the performance of the participants in this study cohort was compared with the performance of a sample of participants in the SPACE study (Seery, Buckley, Bowe, & Carthy, 2016) through a series on independent samples t-tests on the expedited versions on the tests. The results (Table 3) do not indicate a statistically significant difference between the cohorts suggesting no statistical difference between test modalities as all participants are part of a larger national sample.

Table 3. Analysis of performance differences between the study cohort and the national sample (Seery et al., 2016)

Measure	Current Study Cohort		National Sample (n=451)		*t*	*df*	*p*
	Mean	Std. Deviation	Mean	Std. Deviation			
DAT:SR	37.12	19.04	34.66	22.23	-.896	522	.371
PSVT:R	32.93	20.60	35.0	22.26	1.125	565	.261
MCT	26.00	15.23	26.34	16.03	.150	504	.881

The final analysis examined participants who made multiple attempts on the tests. Participants had the capacity to do this for a short period of time while the testing was in progress. The number of times each test was retaken is shown in Table 4.

Table 4. Analysis of multiple test attempts by participants

	No. of participants with multiple attempts	Mean no. of attempts
PSVT:R	6	3.17
MCT	3	3.67
DAT:SR	21	4.67

Discussion and Conclusion

The results from this analysis suggest that the utilization of expedited versions of the PSVT:R and the MCT may be appropriate when the intent is to determine results for large cohorts of participants however they are not as accurate for individual results as variances are observable of up to 40%. However, the use of the full tests is still advocated to support comparisons with other datasets and for an increased accuracy in results. The reliability of the tests was found to be relatively low in this study and also low in the initial results of the national study (expedited DAT:SR, α = .629; expedited PSVT:R, α = .613; expedited MCT, α = .350) (Seery et al., 2016). It

is posited that this is due to the pertinent spatial skills being at a malleable stage at this age as significant gender differences were also not observed.

While Veurink and Hamlin (2015) have shown that differences can occur between online and paper based assessments with the PSVT:R, such differences were not evident in this study. This may be due to differences in age, educational experiences or technological familiarity between the cohorts. Veurink and Hamlin (2015) surmise that sketching during the paper based version may have contributed to the differences they found. It is not known whether sketching was an approach adopted by participants in the national sample however if participants are sketching to solve the problems they may be circumventing the spatial reasoning capacities the tests are designed to espouse, in which case the use of online assessments may be more valid.

Finally, it is interesting to note the frequencies of attempts made by participants on each of the tests. A substantially larger number of re-attempts were made to the DAT:SR than to either of the PSVT:R or the MCT. As the mean performance on the DAT:SR was marginally higher this may be due to a lower difficulty level however it also only consisted of 10 items while the PSVT:R and MCT had 30 and 25 respectively. It is posited that an element of competition emerged either intrinsically or amongst the participants and that the lower number of items fostered this competition by facilitating faster performance feedback. Stemming from this, further research should be designed with the intent of examining the motivational aspects of online spatial ability tests and developmental activities as this may offer insight for the potential creation of an online spatial ability testing and development platform.

References

Bennett, G., Seashore, H., & Wesman, A. (1973). *Differential Aptitude Tests, Forms S and T*. New York: The Psychological Corporation.

Buckley, J., & Seery, N. (2016). Framing Spatial Cognition: Establishing a Research Agenda. In L. Sun, H. Steinhauer, & D. Lane (Eds.), *ASEE Engineering Design Graphics Division 70th Mid-Year Conference* (pp. 118–122). Daytona Beach, Florida: EDGD.

CEEB. (1939). *Special Aptitude Test in Spatial Relations*. New York: College Entrance Examination Board.

Guay, R. (1977). *Purdue Spatial Visualization Test: Rotations*. West Lafayette, Indianna: Purdue Research Foundation.

Kline, P. (2000). *Handbook of Psychological Testing*. England: Routledge.

Larson, G. (1996). Mental Rotation of Static and Dynamic Figures. *Perception & Psychophysics*, *58*(1), 153–159.

McDonald, A. (2002). The Impact of Individual Differences on the Equivalence of Computer-Based and Paper-and-Pencil Educational Assessments. *Computers and Education*, *39*(3),

299–312.

Mead, A., & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin*, *114*(3), 449–458.

Seery, N., Buckley, J., Bowe, B., & Carthy, D. (2016). Spatial Ability in Education: A National Study. In P. Tiernan (Ed.), *33rd International Manufacturing Conference*. Limerick, Ireland: University of Limerick.

Shermis, M., & Lombard, D. (1998). Effects of Computer-Based Test Administrations on Test Anxiety and Performance. *Computers in Human Behavior*, *14*(1), 111–123.

Veurink, N., & Hamlin, A. J. (2015). Comparison of On-line versus Paper Spatial Testing Methods. In *122nd ASEE Annual Conference & Exposition* (p. 26.381.1-26.381.11). Seattle: American Society for Engineering Education.

Wilson, R., Genco, K., & Yager, G. (1985). Assessing the Equivalence of Paper-and-Pencil vs. Computerized Tests: Demonstration of a Promising Methodology. *Computers in Human Behavior*, *1*(3), 265–275.