Englisia MAY 2015
Vol. 2, No. 2, 86-94

VALIDITY IN COMPUTER-BASED TESTING: a literature review of comparability issues and examinee perspectives

Ika Kana Trisnawati
Universitas Muhammadiyah Aceh, Indonesia
ika.arraniry@gmail.com

Abstract

These past years have seen the growing popularity of the Computer-Based Tests (CBTs) in various disciplines, for various purposes, although the Paper-and Pencil Based Tests (P&Ps) are still in use. However, many question whether the use of CBTs outperforms the effectiveness of the P&Ps or if the CBTs can become a valid measuring tool compared to the PBTs. This paper tries to present the comparison on both the CBTs and the P&Ps and their respective examinee perspectives in order to figure out if doubts should arise to the emergence of the CBTs over the classic P&Ps. Findings show that the CBTs are advantageous in that they are both efficient (reducing testing time) and effective (maintaining the test reliability) over the P&P versions. Nevertheless, the CBTs still need to have their variables well-designed (e.g., study design, computer algorithm) in order for the scores to be comparable to those in the P&P tests since the score equivalence is one of the validity evidences needed in a CBT.

Keywords: *Computer-Based Tests; Paper-and-Pencil Based Tests; comparability; examinee perspectives; validity; reliability*

Introduction

The use of computers has significantly increased over the past decade in testing and assessment applications (Olsen, 2000; Gallagher, Bennett, Cahalan, & Rock, 2002; Russell, Goldberg, & O'Connor, 2003). One reason for the ever rising use is that many testing developers believe that computerized testing will be able to provide potential benefits (e.g., efficiency in testing administration) (Gallagher, Ben-

nett, Cahalan, & Rock, 2002). Further, examinees need only to answer shorter tests, as in adaptive testing, compare to traditional tests (paper-and-pencil based testing) to get their achievement measured (Bennett & Rock, 1995; Zenisky & Sireci, 2002).

Olsen (2000) defined such Computer-Based Tests as "tests or assessments that are administered by computer either stand-alone or networked configuration or by other technology devices linked to the Internet or the World Wide Web." However, despite the potentials gained from computerized tests, many studies keep trying to look at the validity of these computerized tests by conducting comparability studies between paper-and-pencil based testing (P&P) and computer-based testing (CBT). Additionally, studies have also been conducted to see the examinees' perceptions on computerized tests. Hence the objective of this paper is to review the validity studies in testing and assessments related to computer-based testing, particularly in the comparability of P&P tests and CBTs and the examinee surveys.

Validity Issues

As stated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), the definition of *validity* refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests." (p. 9). Due to the rapid movement from the P&P test administrations to the CBT system, there are some major concerns in terms of appropriateness of CBTs scores in relation to previous P&P test scores: equivalence across formats and criterion-related validity (Neuman & Baydoun, 1998)

In order for scores in different items or testing materials, different testing procedures, or test forms administered in different test formats to be used interchangeably, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) suggested that evidence of *score equivalence* should be provided. Green, Bock, Humphreys, Linn, and Reckase (1984) also stated that CBT and PPT are equally valid only if they have been demonstrated to yield equivalent measures.

In addition to the issues mentioned above, the administration factors are also considered affecting the examinees' performance throughout the test, such as during the transfer of problems from the screen to scratchwork space, lack of scratchwork

space, and inability to review and/or to skip test items (Russell, Goldberg, & O'Connor, 2003). Therefore, the effect of test modes to the examinees should also be investigated thoroughly.

Summary and Results

Mead & Drasgow (1993) conducted a meta-analysis of all research from 1980s to 1990s by comparing the computerized and paper-and-pencil versions of 123 timed power tests and 36 speeded tests. The tests were intended to measure young adults' and adults' cognitive ability. After correcting for measurement error on 159 cross-mode correlations, they found that the estimated cross-mode correlations were .97 for timed power tests and .72 for speeded tests. For speeded tests, Mead and Drasgow (1993) believed that the test mode affected examinees performance due to the longer time they read text from the screen. While the computer delivery algorithm, either linear or adaptive computer tests, did not result in any differences between CBT and P&P scores.

Wang & Kolen (2001) addressed comparability analyses between ACT Mathematics Assessment computerized adaptive version and paper-and-pencil version through simulation procedures. Raw scores from both versions were converted into the ACT Assessment scale score (range from 1 to 36) to better understand the score interpretation. The findings indicated that cumulative scores distribution in computerized adaptive tests are quite similar to one another, but they differed from the paper-and-pencil tests scale scores. Wang & Kolen (2001) assumed that the differences in scoring methods might influence the major difference in scale score distribution for both test versions.

Gallagher, Bennett, Cahalan, & Rock (2002) examined a computerized Mathematical Expression (ME) using ANOVA to detect construct-irrelevant variance. The test required examinee to enter mathematical expressions into the computer. The study took 178 participants from colleges and universities in the United States. Participants were asked to take parallel computer-based and paper-based tests consisting of ME items, plus a test of their skills in editing and entering data using the computer interface, and also complete questionnaires regarding their personal

background, computer familiarity, perceptions on the math task, and plans for graduate study.

Gallagher et al. (2002) found there was no statistical evidence to claim that individual differences in facility with the computer-based form affected performance on computerized mathematics tasks in a quantitatively skilled, computer-familiar population, and that mean performance, reliability, and relations with other variables were closely similar for both paper-and-pencil and computerized test modes. However, some examinees reported mechanical difficulties in responding on the computer screen and indicated a preference for the paper-and-pencil test. In addition, the authors noted that the findings could not be generalized to other population considering that the sample in this study had higher quantitative skill.

Pomplun, Frey, & Becker (2002) studied the score equivalence from two computerized and two paper-and-pencil versions of the Nelson-Danny Reading Test. The test provided three types of scores: vocabulary, comprehension, and total score. The results showed that both forms of computerized version had higher vocabulary scores than the paper-and-pencil version, and one form also had higher comprehension and total score in the computerized version. Pomplun et al. (2002) believed that such differences might be due to the response speed associated with the use of mouse when recording the responses compared to when examinees had to write their responses in the paper-and-pencil answer sheets. Yet, scale scores for the computerized versions had similar predictive power for course placement as paper-and-pencil did.

To investigate examinees efforts on computerized test, Wise and Kong (2005) analyzed the computer-based version of Information Literacy Test (ILT)—a low-stake assessment—on 506 freshmen at a southeastern university by employing a Response Time Effort (RTE), a new procedure to assess examinee test-taking effort. The study used the Reported Effort subscale of the Student Opinion Scale to measure examinees self-reported effort on the ILT. In addition, the Modified Caution Index developed by Harnisch and Linn (1981) was used to measure a person fit—identifying aberrantly responding examinees. Wise and Kong (2001) found that due to low-stake assessment administration, most examinees tended to response questions too

quickly (rapid-guessing behavior) for items that had the accuracy not exceeded the chance levels.

Recent comparability study conducted by Wang, Jiao, Young, Brooks, & Olson (2008) was to synthesize the administration mode effects of computer-based tests and paper-and-pencil tests on K–12 student reading assessments by applying a meta-analysis on studies conducted from 1980 to 2005. Findings indicated that the administration mode had no statistically significant effect on K–12 student reading achievement scores. There were four variables: study design, sample size, computer delivery algorithm, and computer practice, that made statistically significant contributions to predicting effect size. However, such variables as grade level, type of test, and computer delivery method did not affect the differences in reading scores between test modes.

Table 1 briefly describes six articles being reviewed in this paper that addressed comparability between CBTs and P&P tests, and examinees' perceptions on computerized tests.

Table 1. Summary

Author (Publication Year)	Validity Issues	
	Paper-and Pencil Equivalence	**Examinee Surveys**
Mead & Drasgow (1993)	Yes (power tests) No (speeded tests)	- Timing problem when reading on the screen
Wang & Kolen (2001)	No (score distribution)	N/A
Pomplun et al. (2002)	No	N/A
Gallagher et al. (2002)	Yes	- Encounter difficulty on CBT & P&P preference
Wise & Kong (2005)	N/A	- Highly motivated on high-stake assessment - Low motivation on low-stake assessment
Wang et al. (2008)	Yes (test administration, type of test, grade level, computer delivery method) No (study design, sample size, computer algorithm, computer practice)	N/A

Discussions

The advantages and disadvantages of CBTs

This paper has limited number of studies being reviewed; however, there are several points that can be identified as the effects of the use of CBTs. One clear ad-

vantage of the CBT administration, particularly in adaptive version (CAT), is to reduce testing time while maintaining the test reliability over P&P versions (Wang & Kolen, 2001).

However, suppose the scores in CAT are to be used interchangeably with the P&P test scores, then increasing the reliability of CAT over P&P test will lead to some inequities. Wang & Kolen (2001) illustrated that if a score of 18 on the ACT Assessment scale score is used as a cut score for collegiate sports eligibility by the NCAA, there were 71% of the examinees who had P&P scores of 18 or above while there were 68% of the examinees who got CAT scores of 18 or above. If the CAT scores were used due to the CAT reliability over the P&P test, then it would be only 3% fewer examinees in the CAT compared to the examinees who took the P&P tests that were eligible for collegiate sports.

Wise and Kingsbury (2000) pointed out several important issues regarding the examinee perspectives, as briefly mentioned previously, on computerized adaptive tests were related to the opportunity to *review items* as the P&P tests have, due to the fact that when examinees were able to go back to previous items and change their answers, the examinees' performances mostly increased. The examinees are also believed to have more confidence if they know that they have more control during the test (as shown in these studies: i.e., Glass & Singer, 1972; Blechmann & Dannemiller, 1976; Perlmuter & Monty, 1977).

Another concern is to develop a reasonable *time limit*, that is when two different ability groups (e.g., higher vs. lower) have to take equivalent number of items—higher group takes 20 hard items and lower group takes 20 easy items—then, the problem will be how to establish time limit for each test, because we cannot predict whether harder items will make higher group respond in longer time and easier items will apparently help lower group respond quickly, or vice versa. This problem will definitely affect each group performance, and complicate their score interpretation (Wise & Kingsbury, 2000).

The other significant issue is related to *equity*. For the test to promote fairness and comparability of scores among different types of examinees, equity among examinees should be established. Wise and Kingsbury (2000) mentioned a study by

Sutton (1997) that showed poor and minority students have had less access to computer at home and at school. Therefore, it can be concluded that less access led to less experience in computer practices, and any time limits given may also contribute to differences for examinees within such background.

Concerns on the use of CBTs in language assessment

A special attention should be addressed on the use of CBTs in assessing language ability considering that CBTs are growing more in language testing, especially in second language assessment (Ockey, 2009). As briefly discussed in aforementioned studies by Pomplun, Frey, & Becker (2002) and Wang, Jiao, Young, Brooks, & Olson (2008), the results interestingly show different issues. Scores on reading tests in Wang, Jiao, Young, Brooks, & Olson (2008) did not have significant differences although the test modes used were of paper and computer. However, in Pomplun, Frey, & Becker (2002), the reading comprehension scores from one form of the computer version outperformed those in the paper version, and more specifically in the vocabulary sections. What caused these results might have related to the response speed of the test takers of the CBTs, who found it quicker to move the mouse around to answer.

Nevertheless, a benefit of the CBT over P&P is that the ability of the CBT to predict the effect size, as in the case of Wang, Jiao, Young, Brooks, & Olson (2008). From these findings, concerns on using CBTs to assess language ability are more to whether the computerized testing is able to allow the test takers to get more familiar to the test medium and properly assess their language ability and whether it is capable to efficiently calculate and score tests with performance-or-productive based skills such as speech and writing since the computer still lacks of resources that are needed to assess those skills effectively (Ockey, 2009; Parhizgar, 2012).

Conclusions

Implications for future research

Findings from the CBT and P&P comparability studies were not consistent, however, those studies examining a variety of CBT and P&P versions have implied that scores in computerized tests can be equivalent to the scores in P&P tests so long

as the CBTs have its variables well-designed (e.g., study design, computer algorithm), otherwise the scores in CBTs will not comparable to scores in P&P tests, as the score equivalence is one of the validity evidences needed in a CBT. Therefore, future research should pay more attention on such psychometric issues in CBT, and that test developers should examine their own CBTs for having the intended results that fully interpret examinees 'real' ability.

References

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

Bennett, R. E., & Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered formulating hypothesis test. *Journal of Educational Measurement*, *32(1),* 19–36.

Gallagher, A., Bennett, R. A., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: detecting construct-irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment*, *8*(1), 27–41.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21(4), 347-360.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychological Bulletin*, 114(3), 449-458.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: when are they equivalent? *Applied Psychological Measurement,* 22(1), 71-83.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal*, 93, 836-847.

Olsen, J. B. (2000). Guidelines for computer-based testing. Retrieved May 17, 2008 from http://www.isoc.org/oti/articles/0500/olsen.html.

Parhizgar, S. (2012). Testing and technology: past, present and future. *Theory and Practice in Language Studies,* 2(*1*), 174-178.

Pomplun, M., Frey, S., & Becker, D. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62(2)*, 337-35.

Russell, M., Goldberg, A., & O'Connor, K. (2003). *Computer based-testing and validity: a look back and into the future*. Retrieved May 2, 2008 from http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1003&context=intasc.

Wang, T, & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: issues, criteria and an example. *Journal of Educational Measurement*, 38 (1), 19-49.

Wang, S., Jiao, H., Young, M. J., Brooks, T, & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K 12 reading assessments: a meta-analysis of testing mode effects. *Educational and Psychological Measurement, 28*(1), 5-24.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21,*135-155.

Wise, S. L., & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15(4),* 337-362.