Journal of Education and Learning; Vol. 3, No. 4; 2014
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

A Proposal on the Validation Model of Equivalence between PBLT and CBLT

Huilin Chen[1]

[1] School of Education, Shanghai International Studies University, Shanghai China

Correspondence: Huilin Chen, School of Education, Shanghai International Studies University, 550 Dalian Rd(w), Shanghai, 200083, China. E-mail: chenhuilin@shisu.edu.cn

Received: October 5, 2014 Accepted: October 25, 2014 Online Published: November 10, 2014

doi:10.5539/jel.v3n4p17 URL: http://dx.doi.org/10.5539/jel.v3n4p17

Abstract

The validity of the computer-based language test is possibly affected by three factors: computer familiarity, audio-visual cognitive competence, and other discrepancies in construct. Therefore, validating the equivalence between the paper-and-pencil language test and the computer-based language test is a key step in the procedure of designing a computer-based language test. By taking the test on Essentials of English-Speaking Countries as the case study, this paper elucidates the three-step model of validating the equivalence of the two types of test: investigating computer familiarity, assessing the impact of audio-visual cognitive competence, and examining other discrepancies in construct. The model proposed by this paper can offer some methodological insights on the way to establishing the validation model of the equivalence between the paper-and-pencil language test and the computer-based language test.

Keywords: paper-and-pencil language test, computer-based language test, test equivalence, validation model

1. Introduction

According to Bachman, "Validation has been, and continues to be, a recurring theme of the annual Language Testing Research Colloquium … validation has become the de facto paradigm for language testing research and development".(2000, p. 22) Since the 1980s, the newly developed Computer Based Language Testing (CBLT) has been gradually replacing the conventional Paper-and-Pencil Based Language Testing (PBLT). A key problem encountered in computerized testing is whether an item bank designed for the conventional PBLT can be adopted as the item bank for CBLT with the test validity basically unchanged.

To solve that problem, a lot of studies were conducted to investigate the possibility of validating the equivalence between PBLT and CBLT. Henning (1991) elucidated the validation challenges encountered in constructing the item bank for CBLT by focusing on test methods and procedures. Wainer (2000), however, only focused on method effects. Moreover, Li (2006) held that research on equivalence between PBLT and CBLT should "explore the equivalence or comparability of the results of the same test tasks under different modes of test presentation (in print or on screen) and different response modes (paper and pencil or mouse and keyboard)". The previous studies mentioned above have different research focuses probably because they might have adopted different definitions of test validity. The prevailing definition of test validity was put forward by American Psychological Association which defined test validity as "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (1999, p. 9). According to that authoritative definition, this study holds that the research on validating the equivalence between PBLT and CBLT should take both test construct and method effects into consideration and we should not only see the challenges but also formulate a practical procedure to validate the equivalence between PBLT and CBLT.

2. Framework and Procedure

Based on the aforementioned studies, this study holds that the difference in validity between PBLT and CBLT may be brought about by computer familiarity, audio-visual cognitive competence, and test construct. Therefore, in this study we consider a three-step procedure to validate the equivalence between PBLT and CBLT. Step one aims to estimate the impact of examinees' computer competence on their test results with reference to their levels of computer familiarity. Step two aims to estimate the impact of examinees' audio-visual competence by comparing the test results between PBLT and multimedia CBLT. Step three aims to detect other construct

discrepancies by comparing the construct validity of PBLT and CBLT. The three steps are sequentially interlocked and thus should be conducted in a sequential order.

2.1 Impact of Computer Familiarity

An important aspect which needs to be taken into consideration for validating CBLT is the computer familiarity of the examinees. Since the CBLT tests the language ability of examinees, validation of a CBLT depends on whether language ability is tested. However, there may exist discrepancies in computer familiarity among CBLT examinees. As for the examinees with lower levels of computer familiarity, computer familiarity can be regarded as an ability involved in CBLT and affect the validity and impartiality of CBLT. Computer familiarity, mainly reflected in the familiarity of operating computer hardware, such as using mouse and keyboard, was discovered to have significant impact on test performance (Ward, Hooper, & Hannafin, 1989; Llabre, Clements, Fitzhugh, Lancelotta, Mazzagatti, & Quinones, 1987). Recent studies on computer familiarity (Russell & Haney, 1997; Russell, 1999; Russell & Plati, 2001, 2002) mainly focused on the familiarity of inputting written forms into computers and discovered that some examinees were more accustomed to writing on computers and their performance was even better in the CBLT writing section than in the PBLT writing section.

CBLT designers maytake either of the following two actions to reduce the impact of computer familiarity on the validity of CBLT. The first action is to add a computer familiarization training procedure to the CBLT administration program. Studies (Gressard & Loyd 1987; Raub, 1981) showed that the difference of computer familiarity among examinees could be reduced so that the anxiety of the computer-anxious examinees could be reduced and thus their test performance could be enhanced. The second action is to make the response mode of CBLT as simple as possible. Using mouse, for example, may be a simpler response mode than using keyboard since examinees are generally more familiar with using mouse. Therefore, using mouse as the response mode may greatly reduce the impact of computer familiarity on the validity of CBLT.

Whether a computer familiarization training procedure is added or the response mode is simplified, computer familiarity questionnaires should always be required to be completed after CBLT pretesting. Based on the results from computer familiarity questionnaires, the impact of computer familiarity on pretesting results can be estimated. If computer familiarity does not have significant impact on pretesting results of a CBLT, computer competence is not thought to be a factor which can affect the real CBLT performance. That is to say, computer competence is not an element of the CBLT validity, which is a prerequisite of validating the equivalence between PBLT and CBLT. Researchers have produced different questionnaires to collect information on computer familiarity. Among them, the most influential are the Computer Experience Questionnaire designed by Lee (1986), the Computer Attitude Scale designed by Loyd and Gressard (1984), and the Computer Familiarity Questionnaire designed by Daniel Eignor, Carol Taylor, Irwin Kirsch, and Joan Jamieson (1998).

2.2 Impact of Audio-Visual Cognitive Competence

Since item banking procedures for CBLT are usually carried out by adopting PBLT pretesting(Henning, 1986; Hicks, 1986; Larson, 1987), the item characteristics (e.g. difficulty, discrimination, reliability, validity, etc.) obtained through such pretesting are only meaningful under PBLT situations. It would be inappropriate to use the item characteristics obtained from PBLT situations to construct a CBLT item bank. Since computer is the medium of CBLT which can be presented interactively through many channels, the interaction and multimedia effects may also affect test performance and can be regarded as a potential element of CBLT validity. The study conducted by Mazzeo and Harvey (1988) found out that the number of screens, screen size, font and letter size, image resolution, and other complex ways of presentation could have significant impact on CBLT performance. McKee and Levinson (1990) pointed out that those computer-related modes of item presentation probably had greatly changed the nature of items and that a PBLT and the relevant CBLT probably were not testing the same thing. CBLTs probably not only test the language ability but also are involved with audio-visual cognitive competence. Therefore, CBLT validity is always inconsistent with relevant PBLT validity due to different modes of item presentation.

In order to find out whether multimedia item presentation under CBLT has impact on test performance, comparisons can be made among the multimedia CBLT, text-mode CBLT, and text-mode PBLT. If the test performance under multimedia CBLT situations is significantly different from that under text-mode CBLT and text-mode PBLT situations while there is no significant difference between the latter two, it can be concluded that multimedia item presentation under CBLT may have an impact on test performance and that audio-visual cognitive competence may be an element of CBLT validity.

2.3 Test Construct

Test construct refers to all the abilities or skills involved in a certain test and the relationship among them. If a test can effectively measure a certain construct, that test can be thought to have construct validity concerned with that construct. Construct validity is defined as the degree to which test scores can explain theoretical construct or quality, or the appropriateness to which certain construct or quality can explain test scores. Since construct validity can estimate test validity objectively, it is one of the most powerful methods to estimate test validity. The test construct can be interpreted from data because test construct can be manifested in forms of observable characteristics of responses.

Like other types of tests, CBLT also has its construct. If a CBLT is derived from the relevant PBLT, the CBLT should also have the same construct as the relevant PBLT has and therefore they both should have the same construct validity. By analyzing the relevance between a CBLT and its relevant PBLT and the relationship between their internal structures, the equivalence between the two versions of the test can be definitely guaranteed.

Multitrait-multimethod analysis (MTMM) was adopted in this study to validate the construct equivalence between CBLT and PBLT because the validation of construct equivalence involved in this study mainly focuses on the relationships among traits, among methods, and between traits and methods. MTMM, first proposed by Campbell and Fiske (1959), aims to estimate convergent validity, discriminate validity, and mode effect.

3. A Case Study

The case study adopted the achievement test of the course "Essentials of English-Speaking Countries" to explore whether the three-step model elaborated above is feasible to validate the equivalence between PBLT and CBLT. Since the course "Essentials of English-Speaking Countries" was both taught and tested in English, the test adopted in the case study can be regarded as a kind of English language test. The subjects of the case study include 296 English major students who had just received the course lectures. Those subjects were randomly divided into three groups: Group A (96), Group B (100), and Group C (100). The test item bank included 300 multiple choice questions which could be divided into 60 knowledge points (5 items for one knowledge point) spreading in 35 chapters. The case study was conducted in a situation which was similar to pretesting. The research instruments adopted are SPSS, LISREL, and the computerized test administration package Fast Test Pro (Weiss, 2008).

3.1 Investigating Computer Familiarity

In this case study, we investigated the computer familiarity of all 296 subjects with the adapted version of Computer Familiarity Questionnaire (Eignor, Taylor, Kirsch, & Jamieson, 1998). Since the adapted questionnaire contains 11 questions with 4 choices for each, the scores from the adapted questionnaire range from 11 to 44, the higher the score the higher the computer familiarity. The questionnaire response distribution of the 296 subjects is shown in Table 1.

Table 1. Response distribution in Computer Familiarity Questionnaire

	Once a week or more often (score = 4)	Between once a week and once a month (score = 3)	Less than once a month (score = 2)	Never (score = 1)
How often is a computer available to you at school?	268	27	1	0
How comfortable are you with using a computer?	228	61	6	1
How comfortable are you with using a "mouse"?	255	40	1	0
How comfortable are you with using a computer to write a paper?	181	74	33	8
How comfortable would you be taking a test on a computer?	188	70	27	11

How would you rate your ability to use a computer?	33	114	141	8
How often do you use a computer?	289	6	1	0
How often do you use a "mouse" with a computer?	289	6	1	0
How often do you use word processing in English?	234	52	7	3
How often do you use spreadsheets?	101	114	67	14
How often do you use graphics?	222	52	21	1

According to the responses from the questionnaire, the average score of computer familiarity for the 296 subjects was 39.4, which demonstrated that the sample students generally had high levels of computer familiarity. It can also be seen from Table 1 that the scores for the two questions concerning the use of "mouse" were much higher than those for other questions. However, the scores for the questions concerning "writing on computer", "using spreadsheets", and "computer ability" were comparatively lower than those for other questions. The findings show that avoiding using complicated computer skills and encouraging the use of "mouse" probably are the ways to enhance computer familiarity.

Although the computer familiarity of the 296 subjects was at high levels, it is also necessary to find out whether their computer familiarity has impact on their performance in CBLTs. All the 296 subjects received a CBLT containing 60 text-mode items, one selected from each knowledge point. The impact of their computer familiarity on their test results was estimated through ANOVA in which the scores of each question in Computer Familiarity Questionnaire and the total scores of the questionnaire were taken as factors while the results of the 60-item CBLT were taken as the dependent variable. Table 2 shows the estimate results.

Table 2. Impact of computer familiarity on CBLT performance

Dependent variable: CBLT results					
Questions	Type III Sum of Squares	df	Mean Square	*F*	*p*
Computer use at school	264.263	2	132.131	3.098	.047
Comfort in using computer	267.138	3	89.046	2.081	.103
Comfort in using mouse	209.177	2	104.588	2.442	.089
Comfort in writing on computer	285.806	3	95.269	2.230	.085
Comfort in test on computer	272.572	3	90.857	2.125	.097
Computer ability	353.889	3	117.963	2.777	.042
Computer use frequency	266.946	2	133.473	3.131	.045
Mouse use frequency	266.946	2	133.473	3.131	.045
Use of word processing	206.976	3	68.992	1.605	.188
Use of spreadsheets	498.430	3	166.143	3.957	.009
Use of graphics	241.771	3	80.590	1.880	.133
Total score of computer familiarity	950.963	20	47.548	1.107	.341

From Table 2, it can be seen that only the scores of "Use of spreadsheets" had strong impact on the CBLT results while the scores of other questions had weak or no impact on the CBLT results. The total scores of Computer Familiarity Questionnaire also had no impact on the CBLT results (p = .341> .05), which demonstrated that computer familiarity generally has no significant impact on CBLT performance. That is to say, computer familiarity of the subjects does not influence the test validity, other things being equal. If the computer familiarity of sample examinees influences CBLT validity, greater importance should be attached to computer familiarization training procedure and response mode simplification.

3.2 Assessing the Impact of Audio-visual Cognitive Competence

In order to evaluate the impact of audio-visual cognitive competence on test validity, comparisons among different modes of item presentation were carried out. First of all, based on the 60-item text-mode CBLT results obtained in 3.1, the 96 subjects from Group A were further classified into three subgroups (A1, A2, and A3) whose test means were not significantly different (p = .908> .05 based on ANOVA). Then, besides the original text-mode PBLT form, the remaining 240 items in the item bank were all adapted into two other forms: text-mode CBLT items and multimedia CBLT items. The text-mode CBLT items were presented on the computer screen in two windows which contained item stems and alternatives respectively. The multimedia CBLT items had an additional window which contained pictures, sound clips, and video clips. Next, the 240 items were administered to Subgroup A1, Subgroup A2, and Subgroup A3 in forms of text-mode PBLT, text-mode CBLT, and multimedia CBLT respectively. It was found that the test results of the three subgroups showed significant difference (p = .006< .05 based on ANOVA). The ANOVA multiple comparisons are shown in Table 3.

Table 3. ANOVA multiple comparisons among modes of item presentation

			Mean difference	p
Bonferroni	text-mode PBLT	text-mode CBLT	-3.25000	1.000
		multimedia CBLT	-17.93750	.008
	text-mode CBLT	text-mode PBLT	3.25000	1.000
		multimedia CBLT	-14.68750	.041
	multimedia CBLT	text-mode PBLT	17.93750	.008
		text-mode CBLT	14.68750	.041

From Table 3, it can be seen that the test results of text-mode PBLT were not significantly different from those of text-mode CBLT (p = 1.000> .05) while the test results of both text-mode PBLT and text-mode CBLT were significantly different from those of multimedia CBLT (p = .008< .05; p = .041< .05). From the above multiple comparisons, we can conclude that the fonts and windows in CBLTs do not have significant impact on test validity; however, multimedia presentations like pictures, sound clips, and video clips have significant impact on test validity probably because the ability of recognizing multimedia materials is part of the test construct. If the validation of equivalence failed in this step, there would be no need to go further to the third step. In order to achieve validation of equivalence in this step, the CBLT should be designed to include as few multimedia materials as possible.

3.3 Examining Construct Discrepancy

If computer familiarity is not a factor to influence performance in CBLT and both PBLT and CBLT are presented identically in text-mode, the next step to validate the equivalence between PBLT and CBLT is to detect whether there exist subtle construct discrepancies between them.

First of all, based on the 60-item text-mode CBLT results obtained in 3.1, Group B and Group C were found to have no significant difference in test means (p = .723> .05 based on t-test), which demonstrated that Group B and Group C did not have significant difference in mastering the course knowledge. Then, the remaining 240 items in the item bank were delivered to Group B in text-mode PBLT form while the same items were delivered to Group C in text-mode CBLT form. Next, MTMM analysis was carried out to detect whether there existed construct discrepancies by comparing trait and method effects. In MTMM analysis, the 5 latent trait factors were geography, history, politics, economy, and culture, while the 2 method factors were PBLT and CBLT. The loadings between the test results of each chapter and the interactions of the 5 trait factors and the 2 method factors are shown in Table 4.

Table 4. Loadings between chapter results and factor interactions

Chapter	Geography		History		Politics		Economy		Culture	
	PBLT	CBLT	PBLT	CBLT	PBLT	CBLT	PBLT	CBLT	PBLT	CBLT
1	.71	.72								
2	- .09	- .09								
3	.71	.73								
4	.54	.54								
5	.71	.72								
6	.24	.24								
7	.00	.00								
8	.24	.24								
9	.31	.32								
10	.17	.18								
11			.03	.03						
12			.77	.01						
13			.01	.01						
14			.01	.01						
15					1.12	1.14				
16					1.14	1.16				
17					.71	.71				
18					.62	.62				
19							.96	.97		
20							.86	.86		
21							.65	.66		
22							.71	.71		
23									.90	.92
24									.52	.53
25									.54	.55
26									.79	.79
27									.71	.71
28									.89	.90
29									.64	.65
30									.26	.26
31									.42	.42
32									.62	.63
33									- .32	- .33
34									.27	.27
35									.03	.03

In Table 4, it can be found that test results concerning the same chapter have almost the same factor loadings regardless of whether the test is PBLT or CBLT (except for the test results concerning chapter 12). The results of MTMM analysis shows that the text-mode PBLT and the text-mode CBLT have almost the same construct, which can also be seen from the loadings between the two types of factors in Table 5.

Table 5. Loadings between two types of factors

	Geography	History	Politics	Economy	Culture
PBLT	.18	7.22	.16	.19	.22
CBLT	.14	6.11	.14	.15	.19

In Table 5, it can be seen that each trait factor has almost the same loadings with either of the two method factors, which further confirms that method factors (PBLT and CBLT) do not have impact on the construct validity of the test involved in the case study.

Only at this stage can the PBLT and the CBLT, which are identically presented in text-mode, be regarded as equivalent tests for sample examinees. If the validation of equivalence was achieved in the first two steps but failed in this step, the PBLT and CBLT still cannot be regarded as two equivalent tests and more studies should be carried out to find out the causes of their subtle differences in test construct.

4. Conclusion

The three validation steps mentioned above checked the three factors probably affecting the equivalence between PBLT and CBLT step by step. In accordance with the three validation steps, the three factors to be checked are computer competence, audio-visual cognitive competence, and test construct. The study carried out a case study to empirically demonstrate the reasonableness, the necessity, and the feasibility of the three-step model for validating the equivalence between PBLT and CBLT. In the case study, computer familiarity was found to have no significant impact on the CBLT results, the impact of audio-visual cognitive competence was avoided by only adopting text-mode item presentation, and the construct validity of the two test versions was almost equal. Therefore, it can be concluded that the equivalence between the PBLT version and the CBLT version of the test on Essentials of English-Speaking Countries was validated.

Although this study can be a reference for validating the equivalence between PBLT and CBLT, this study also has its limitations. First, the size of sample and the item pool is comparatively small. Second, the test content involved in the case study is only concerned with limited aspects of language testing domain. Third, elaborations on how to achieve test equivalence were quite limited. Future studies are needed to be carried out by putting the validation model into the practice of test equating between PBLT and CBLT and designing large scale CBLTs.

Acknowledgments

This paper is a partial fulfillment of the project “Application of cognitive diagnosis to language testing” (KX171266) sponsored by Shanghai International Studies University, Shanghai China.

References

American Psychological Association (APA). (1999). *Standards for educational and psychological testing*. Washington DC: American Education Research Association.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*, 1-42. http://dx.doi.org/10.1177/026553220001700101

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix.*Psychological Bulletin*, *56*, 81-105. http://dx.doi.org/10.1037/h0046016

Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). *Development of a scale for assessing the level of computer familiarity of TOEFL examinees*. Princeton, NJ: Educational Testing Service.

Gressard, C. P., & Loyd, B. H. (1987). An investigation of the effects of math anxiety and sex on computer attitude. *School Science and Mathematics*, *87*, 125-135. http://dx.doi.org/10.1111/j.1949-8594.1987.tb11684.x

Henning, G. (1986). Item banking via DBASE II: The UCLA ESL proficiency examination experience. In C. W. Stansfield (Ed.), *Technology and language testing* (pp. 69-77). Washington, DC: Teachers of English to Speakers of Other Languages.

Henning, G. (1991). Validating an item bank in a computer-assisted or computer-adaptive test. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 209-222). New York: Newbury House.

Hicks, M. (1986). Computerized multilevel ESL testing, a rapid screening methodology. In C. W. Stansfield (Ed.), *Technology and language testing* (pp. 79-90). Washington, DC: Teachers of English to Speakers of Other Languages.

Larson, J. W. (1987). Computerized adaptive language testing: A Spanish placement exam. In K. M. Bailey, T. L. Dale, & R. T. Clifford (Eds.), *Language testing research: Selected papers from the 1986 colloquium* (pp. 1-10). Monterey, CA: Defense language Institute.

Lee, J. (1986). The effects of past computer experience on computerized aptitude test performance. *Educational and Psychological Measurement*, *46*, 727-733. http://dx.doi.org/10.1177/0013164486463030

Li, Q. (2006). Equivalence studies of paper-and-pencil based language testing and computer based language testing: A survey. *Foreign Language World*, *114*, 73-78.

Llabre, M. M., Clements, N. E., Fitzhugh, K. B., & Lancelotta, G. (1987). The effect of computer-administered testing on test anxiety and performance.*Journal of Educational Computing Research*, *3*, 429-433. http://dx.doi.org/10.2190/GMX4-AV22-FD4R-A06P

Loyd, B., & Gressard, C. (1984). Reliability and factorial validity of computer attitude scales. *Educational and Psychological Measurement*, *44*, 501-505. http://dx.doi.org/10.1177/0013164484442033

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature*. Princeton, NJ: Educational Testing Service.

McKee, L. M., & Levinson, E. M. (1990). A review of the computerized version of the Self-Directed Search. *Career Development Quarterly*, *38*, 325-333. http://dx.doi.org/10.1002/j.2161-0045.1990.tb00222.x

Raub, A. C. (1981). *Correlates of computer anxiety in college students*. Unpublished doctoral dissertation, University of Pennsylvania.

Russell, M. (1999).Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, *7*. Retrieved from http://epaa.asu.edu/epaa/v7n20.html

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, *5*. Retrieved from http://epaa.asu.edu/epaa/v5n3.html/

Russell, M., & Plati, T. (2001). Mode of administration effects on MCAS composition performance for grades eight and ten. *Teachers College Record*, Retrieved from http://www.tcrecord.org/Content.asp?ContentID=10709

Russell, M., & Plati, T. (2002). Does it matter with what I write? Comparing performance on paper, computer and portable writing devices. *Current Issues in Education*, *5*. Retrieved from http://cie.ed.asu.edu/volume5/number4/

Wainer, H. et al (Eds.). (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ward, T. J., Hooper, S. R., & Hannafin, K. M. (1989). The Effect of Computerized Tests on the Performance and Attitudes of College Students. *Journal of Educational Computing Research*, *5*, 327-333. http://dx.doi.org/10.2190/4U1D-VQRM-J70D-JEQF

Weiss, D. J. (2008). *Manual for the FastTEST professional testing system, version 2*. St. Paul, MN: Assessment Systems Corporation.