Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 7, Number 23, November, 2001 ISSN=1531-7714

Effects of Removing the Time Limit on First and Second Language Intelligence Test Performance

Jennifer Mullane and Stuart J. McKelvie
Department of Psychology
Bishop's University

Abstract

Canadian post-secondary students with a moderate level of second language competence in English or French took the Wonderlic Personnel Test with the standard 12-min time limit or with no time limit. Participants who took the timed test in their second language scored lower than those who took it in their mother tongue, but the disadvantage was greater for limited French-proficient (LFP) students than for limited English-proficient (LEP) students. Scores increased with no time limit and the gain was greater on the French test for LFP students than for mother tongue students. On the English test, the gain was similar for LEP students and mother tongue students. It is concluded that the time accommodation can be applied to clients who are taking an intelligence test in their second language.

Maximum performance psychological tests enjoy widespread use in North America, particularly for educational and employment decisions. Ever since they were administered en masse to U.S. army recruits and immigrants in the early 1900s, the testing of minorities, particularly limited-English proficient (LEP) clients, has been controversial (Samelson, 1977). The issue is test fairness or "intercept bias": the possibility that the minority group scores lower than the majority on the test but not on the criterion (Anastasi & Urbina, 1997). LEP clients may be disadvantaged on a standardized intelligence test given in English, but they may be just as capable in school or on the job as native English speakers (Cummins, 1987).

In the U.S., 15 to 20% of students speak a language other than English at home (Geisenger & Carlson, 1992). In Canada, the corresponding percentage is 32 (23% speaking the second official language, French) (Statistics Canada, 1996). Although it may (Angus, 2000) or may not (Lam, 1993; Rivera, Vincent, Hafner & LaCelle-Peterson, 1997) be appropriate to administer achievement tests to LEP clients in the majority language, doing so with tests of general cognitive functioning (intelligence) is questionable. Test scores of LEP clients will vary with their English skills, and are likely to underestimate their ability to learn, which should remain relatively constant (Angus, 2000). Indeed, even bilingual children perform more poorly than monolinguals on standardized tests (Valdes & Figueroa, 1994).

To alleviate this problem, students could be assessed with a nonverbal test of intelligence, with a test in their own language, or with a modified test in the majority language. Although nonverbal tests have been defended (e.g., Bracken & McCallum, 2001), they may not measure the same cognitive functions as verbal intelligence tests (Angus, 2000; Anastasi & Urbina, 1997), and they are less successful than verbal tests at predicting educational performance (Angus, 2000; Geisenger & Carlson, 1992; Gregory, 2000). Such problems can be avoided by administering the test in the client's own language, but there can be problems in translation, with no guarantee that the second-language version measures the same construct as the original (Wonderlic, 1992). Moreover, few tests have been translated from English (Geisenger & Carlson, 1992).

The third solution is to modify the original test or test conditions to minimize the disadvantage to the LEP client. For example, grammar may be simplified, a glossary of terms may be provided, or time constraints may be relaxed (Azar, 1999; Rivera et al., 1997). In a survey of statewide assessment programs, it was found that 52% of states permitted amendments for LEP students, and the most popular one (81%) was giving extra time. As with test translation, there is no guarantee that such changes will preserve test validity, but an accommodation can be said to work if it provides a "differential advantage" to the LEP client (Azar, 1999). This means that the gain that accrues from the accommodation must be greater for LEP than for English mother tongue clients. It has been found that SAT scores increase for LEP clients who have extra time, but it is not clear if the gain is greater than for other regular students (Azar, 1999). More research is needed on the effect of test modifications (Rivera et al., 1997).

The purpose of the present study is to investigate the effect of removing the time limit for students taking a

standardized intelligence test in their second language. To evaluate whether the accommodation worked, the results were compared with those obtained by students who took the test in their mother tongue. In our study, participants completed the 50-item spiral-omnibus Wonderlic Personnel Test (WPT) with the standard time limit or with unlimited time. The WPT is widely employed in personnel selection, and can be used in educational settings (Wonderlic, 1992). Although Murphy (1984) doubts the manual's claim that it measures the "ability to learn," he states that it is a useful predictor of job performance. Furthermore, it can be classified as a test of general intelligence, because the items are based on the Otis Self-administering Tests of Mental Ability, and cover numerical reasoning, verbal reasoning, synonyms-antonyms, nonverbal reasoning, information, and attention to detail (McKelvie, 1992). Moreover, its loadings on Aptitude G of the General Aptitude Test Battery exceed .56, and its scores correlate moderately to very highly (up to .90; Dodrill, 1981; Dodrill & Warner, 1988) with the Wechsler Adult Intelligence Scale in general and psychiatric groups. Wonderlic (1992) also states that the test correlates reasonably (from .30 to .45) with academic achievement, so that it can be used as a selection and counseling tool in postsecondary education. Correlations at the university level are lower than those at high school (McKelvie, 1989, 1992), but this may be due to restriction of range.

Under standard instructions, the WPT is administered with a 12-min time limit, which makes it a highly-speeded test (Davou & McKelvie, 1984). Belcher (1992) states that this may be unfair for the increasing number of LEP clients. When answering an item, they may have to translate it into their mother tongue to fully understand its meaning. By spending more time than the native English speaker on item comprehension, they will not be able to complete as many items in the restricted time allowed. Although they may answer correctly on the items that they attempt, their obtained score will underestimate their true score. Indeed, the Wonderlic (1992) manual reports that Hispanic Americans scored about six points lower than white Americans on the English WPT. It is also notable that one study has shown that the time accommodation worked on the WPT for students with weak study habits (Davou & McKelvie, 1984). In the standard timed condition, scores were lower for students with weak than with strong study habits. However, the gain with unlimited time was greater for the weak than the strong studiers.

The WPT has also been translated into French for Canadian use, particularly in Québec, where the majority mother tongue is French. Here, English speakers are the minority. On the other hand, in the Québec academic setting where testing took place, the language of instruction is English and students are enrolled from all parts of Canada (as well as the U.S. and overseas). Here, French speakers are the minority. Thus, we had the opportunity to examine the effect of removing the time limit not only for LEP (French mother tongue) clients taking the test in English, but also for LFP (limited-French proficient, English mother tongue) clients taking the test in French. It was hypothesized that, although all participants would benefit from the relaxed time constraint, the gain would be greater for people taking the test in their second language. That is, the gain should be greater on the English WPT for French (LEP) compared to English mother tongue students, and on the French WPT for English (LFP) compared to French mother tongue students.

Method

Participants

The participants were 133 (89 women, 44 men; mean age 21.6 yr.) postsecondary college and university students who reported English or French as their mother tongue and at least moderate competence in the other language. After matching for gender, English and French first-language participants were assigned randomly to one of four experimental conditions resulting from two levels of two independent variables: timing (Limited Time, Unlimited Time) and test language (English, French).

Materials and Procedure

As described above, the Wonderlic Personnel Test (WPT) measures general intelligence. Its reliability has been estimated as .84 to .94 (test-retest, even for a 5-yr, period; Dodrill, 1983), .88 to .94 (split-half; Wonderlic, 1992), and .73 to .95 (alternate-form; Schoenfeldt, 1985).

Although participants reported that they had moderate competence in their second language, this was assessed more formally with two written tests of second-language proficiency: a 13-item self-report measure (Self Evaluation Questionnaire) in their mother tongue, and a 17-item objective test (Objective Cloze Test) in their second language. The Objective Cloze Test (OCT) was developed from an experimental version of the Test of English as a Foreign Language (TOEFL) (Hale, Stansfield, Rock, Hicks, Butler, & Oller, 1989). In the cloze procedure, words in a text are deleted and must be supplied by the examinee. In the OCT, they were given four choices from which to choose. Scores from multiple-choice cloze tests correlate highly with those from fill-in-the-blank versions (Chappelle & Abraham, 1990).

According to Al-Fallay (1997), the cloze procedure has concurrent validity as a predictor of other second-language achievement tests, but its face validity is questionable. We followed their recommendation to supplement the cloze test with another technique, and chose self-assessment, which has been useful (Ross, 1998). The Self Evaluation Questionnaire (SEQ) was adapted from the Bishop's University ESL (English as a Second Language) department's screening questionnaire. Items tapped various aspects of language ability and were answered on a 7-point Likert scale. Examples are: "When I speak French among a small group of people that I know well, I feel 1 (uneasy) to 9 (very much at ease)"; "I can understand newspaper articles without the use of a dictionary: 1 (not at all) to 7 (perfectly)". Both the OCT and SEQ were constructed in English and translated professionally into French.

Participants were tested individually or in small groups. After signing a consent form, they completed the SEQ then the OCT. The SEQ was given first to avoid contamination of self-reports by perceived performance on the objective test. Following these tests, the WPT was administered then participants were debriefed. The WPT was scored using the standard key for the English version. However, for the French version, two adjustments had to be made because of imprecise translation.

Results and Discussion

Second-Language Competence

The correlation between OCT and SEQ scores for all 133 participants was .529, $p < .01$. This indicates that the two tests were related, but tapped slightly different aspects of second-language competence, supporting the view that it should be assessed with more than one technique (Al-Fallay,1997).

In the present design, second-language competence should be moderate and similar in the different experimental conditions. If participants were perfectly bilingual (scoring close to maximum), second-language testing would not an issue, and if they were essentially monolingual (scoring close to 0), they would not be taking a test in another second language. To evaluate this, 2 X 2 X 2 (Timing X Mother Tongue X Test Language) factorial ANOVAs were conducted for each second-language test. There was a significant effect of mother tongue for both OCT scores, $F(1, 125) = 4.49$, $p < .05$, and for SEQ scores, $F(1, 125) = 11.57$, $p < .01$. Mother-tongue French speakers scored higher on the English OCT than mother-tongue English speakers scored on the French OCT (Ms = 12.3. 11.3). French speakers also rated themselves as more proficient in English than English speakers rated themselves in French (Ms = 62.5, 53.6). Although the tests were professionally translated from English to French, the OCT may not be equally difficult in each language. However, the fact that the results agreed with those of the SEQ indicates that our French speakers were more proficient in English than were our English speakers in French. At the same time, none of the mean scores was extreme. On the OCT, the maximum possible score was 17 (Ms were 12.3, 11.3) and on the SEQ it was 91 (Ms were 62.5, 53.6). Moreover, the lack of any significant interactions indicates that second-language competence was matched in the four timing/test language conditions.

Because the major comparisons of interest were the effects of time for people taking the test in their first or second language, the different levels of second-language competence were dealt with by including OCT and SEQ scores as covariates in the analysis of WPT scores.

Intelligence Test Performance

Initially, a 2 X 2 X 2 (Timing X Mother Tongue X Test Language) factorial ANOVA (with SEQ and OCT as covariates) was conducted on WPT scores. Table 1 shows the means in each of the eight conditions. Not surprisingly, there was a significant effect of timing, $F(1, 123) = 150.46$, $p < .001$, with higher scores in the unlimited time than in the limited time condition (Ms = 32.58, 22.17). Converted to Cohen's (1977) standardized effect size (d), this difference was 2.10 which clearly exceeds his guideline of 0.80 for a large effect.

Table 1: Mean Wonderlic Personnel Test Scores in Each Condition

	Limited Time			Unlimited Time		
Test Language	n	M	SD	n	M	SD
English						
English Mother Tongue	16	26.33	4.24	17	34.24	6.08
French Mother Tongue (LEP)	18	23.03	5.86	13	30.70	7.50
French						
English Mother Tongue (LFP)	17	16.46	4.62	16	31.94	5.10
French Mother Tongue	18	22.74	4.36	18	33.49	3.07

Note. Maximum score = 50. Means are adjusted for covariates. LEP = limited-English proficient, LFP = limited-French proficient.

There was also a significant effect of test language, $F(1, 123) = 9.32$, $p < .01$, with lower scores on the French WPT than on the English WPT (Ms = 26.09, 28.66). However, timing interacted significantly with test language, $F(1, 123) = 12.15$, $p < .01$. The gain from the limited to the unlimited time condition was greater on the French test (Ms = 19.42, 32.76; gain = 13.34) than on the English test (Ms = 24.92, 32.39; gain = 7.47). There was also a significant interaction between mother tongue and test language, $F(1, 123) = 13.92$, $p < .01$. For people whose mother tongue was English, scores were lower on the French than on the English test (Ms = 24.29, 30.03), but for those whose mother tongue was French, the scores were very similar (Ms = 27.88, 27.28).

These results indicate that the effects of timing and of mother tongue were greater on the French than on the English test. Because it was predicted that the effect of timing would be greater for second- than for first-language participants

on each test, planned 2 X 2 (Timing X Mother Tongue) ANOVAs (again with OCT and SEQ as covariates) were conducted separately for the English and French WPTs. For the French version of the test, all three effects were significant: timing, $F(1, 63) = 157.18$, $p < .01$, mother tongue, $F(1, 63) = 13.15$, $p < .01$, and their interaction, $F(1, 63) = 5.04$, $p < .05$. Scores were higher in the unlimited than in the limited time condition, and for French mother tongue than for English mother tongue (LFP) participants. However, and of particular interest here, the latter difference was only significant with limited time. Here, LFP students (English native speakers) performed more poorly than French native speakers, $t(63) = 4.24$, $p < .01$, but when the time constraint was removed, they did not, $t(63) = 1.07$, $p > .10$. Another way of looking at this is that both groups of participants benefited from extra time, but the gain was greater for the LFP students (15.5 points, $d = 3.54$) than for the French native speakers (10.8 points, $d = 2.44$). In other words, the time accommodation worked because it provided a differential advantage to the LFP clients (Azar, 1999).

For the English version of the test, there were significant effect of timing, $F(1, 58) = 34.96$, $p < .01$, and of mother tongue, $F(1, 58) = 6.31$, $p < .05$, but not their interaction, $F(1, 58) = 0.01$, $p > .90$. Scores were higher in the unlimited than in the limited time condition, and for English mother tongue than for French mother tongue (LEP) participants. Thus, although LEP students performed more poorly than English native speakers with the standard 12-min time limit, and although they benefited from unlimited time, the gain was not greater than that for English speakers. In fact, the effect size for time was $d = 1.46$ (7.7 points) for French speakers and $d = 1.50$ for English speakers (7.9 points). Here, the time accommodation did not work.

Why did the time accommodation work for LFP but not for LEP participants? The answer may be that, in the timed condition, the LFP disadvantage on the French test (Ms = 16.46, 22.74; difference = 6.28) was greater than the LEP disadvantage on the English test (Ms = 23.03, 26.33; difference = 3.30). The accommodation may only be effective if the disadvantage is great.

But why was the LFP disadvantage greater than the LEP disadvantage? The most obvious reason is that French second-language competence was less than English second-language competence as assessed on the OCT and SEQ. In fact, the mean levels of competence were not extreme, and the differences were controlled via covariance analysis[1]. Perhaps the answer is that the tests did not fully capture the fact that the participants whose second language was English were studying in an English-speaking institution, and who probably had more practice listening, reading and writing in their second language than did participants who second language was French. In fact, the SEQ only had one question about frequency of second-language use, and it only referred to speaking. To aid in the identification of clients likely to perform poorly on the timed WPT in their second language, the SEQ should be expanded to include information about reading and writing. It might also be noted that if the present study had also been conducted in a French-speaking institution, the results might have been reversed. That is, the gain from unlimited time might have been greater for LFP than for LEP students.

Although these results show that the time accommodation can work, they do not show whether the WPT scores obtained with unlimited time are as valid as those obtained with limited time. Wonderlic (1992) himself discusses this issue, stating that "while untimed scores are valid assessments of cognitive ability, they are not as accurate as the timed scores." Notably, a 25-item short form of the WPT given with unlimited time was as reliable as the full version (when corrected for length), and also had a similar criterion validity coefficient for predicting university grades (McKelvie, 1994). However, because his studies indicated that people taking the test under both conditions scored about six points higher with unlimited time, Wonderlic recommends that this score be used to estimate the timed score by subtracting six points from it.

French mother-tongue speakers (LEP participants) scored only slightly lower than English mother-tongue speakers on the English WPT with unlimited time. Because English mother-tongue speakers (LFP) did not score significantly lower than French mother-tongue speakers on the French WPT in this condition, extra time minimized or removed the second language disadvantage. Therefore, we suggest that people taking an intelligence test in their second language be permitted the accommodation of unlimited time. In the case of the WPT, their timed score can then be estimated by subtracting six points.

Conclusion

These results provide experimental evidence that the time accommodation can work for people whose second-language intelligence test limited time score is clearly lower than that of mother-tongue participants, and it does no injustice to those whose limited time score is only slightly lower. Therefore, we recommend removing time limits on standardized intelligence tests for clients taking them in their second language. The present measures of second-language competence should be expanded, and future research should obtain more information about the psychometric properties (norms, reliability, validity) of untimed intelligence tests.

-

References

Al-Fallay, I. (1997). Investigating the reliability and validity of the fixed ratio multiple-choice cloze test. *Human and Social Sciences, 24*, 507-526.

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

Angus, W. A. (2000). Using achievement tests, diagnostic (achievement) tests, and tests of intelligence with ESP populations. http://www.psychtest.com/ESLtest.htm.

Azar, B. (1999). Fairness a challenge when developing special needs tests. *APA Monitor Online, 30.* http://www.apa.org/monitor/dec99/in2.html.

Belcher, M. J. (1992). Review of the Wonderlic Personnel Test. In J. J. Kramer & J. C. Conoley (eds.), *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Bracken, B. & McCallum, R. S. (2001). Assessing intelligence in a population that speaks more than tow hundred languages: A nonverbal solution. In L. A. Suzuki and J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications,* 2nd ed. San Francisco, CA: Jossey-Bass, Inc.

Chapelle, C. A., & Abraham, R. G. (1990)/ Cloze method: What difference does it make? *Language Testing, 7,* 121-146.

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

Cummins, J. (1997). Psychoeducational assessment in multicultural school systems. *Canadian Journal for Exceptional Children, 3,* 115-117.

Davou, D., & McKelvie, S. J. (1984). Relationship between study habits and performance on an intelligence test with limited and unlimited time. *Psychological Reports, 54,* 367, 371.

Dodrill, C. B. (1981). An economical method for the evaluation of general intelligence in adults. *Journal of Consulting and Clinical Psychology, 49,* 668-673.

Dodrill, C. B. (1983). Long-term reliability of the Wonderlic Personnel Test. *Journal of Consulting and Clinical Psychology, 51,* 316-327.

Dodrill, C. B., & Warner, M. H. (1988). Further studies of the Wonderlic Personnel Test as a brief measure of intelligence. *Journal of Consulting and Clinical Psychology, 56,* 145-147.

Edinger, J. D., Shipley, R. H., Watkins, C. E., & Hammett, E. B. (1985). Validity of the Wonderlic Personnel Test as a brief IQ measure in psychiatric patients. *Journal of Consulting and Clinical Psychology, 53,* 937-939.

Geisenger, K. F., & Carlson, J. F. (1992). Assessing language-minority students. Practical Assessment, Research & Evaluation, 3(2). Available online: http://pareonline.net/getvn.asp?v=3&n=2.

Gregory, R. J. (2000). *Psychological testing: History, principles, and applications,* 3rd ed. Boston: Allyn and Bacon.

Hale, G., Stansfield, C., Rock, D., Hicks, M., Butler, F., & Oller, J. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing, 6,* 49-78.

Lam, T. C. M. (1993). Testability: A critical issue in testing language minority students with standardized achievement tests. *Measurement and Evaluation in Counseling and Development, 26,* 179-191.

McKelvie, S. J. (1989). The Wonderlic Personnel Test: Reliability and validity in an academic setting. *Psychological Reports, 65,* 161-162.

McKelvie, S. J. (1992). Does memory contaminate test-retest reliability? *The Journal of Psychology, 119*, 59-72.

McKelvie, S. J. (1994). Validity and reliability findings for an experimental short form of the Wonderlic Personnel Test in an academic setting. *Psychological Reports, 75,* 907-910.

Murphy, K. R. (1984). The Wonderlic Personnel Test. In D. J. Keyser and R. C. Sweetland (Eds.), *Test critiques* (volumes I-VI). Kansas City, MO: Test Corporation of America, pp. 769-775.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). Statewide assessment programs: Policies and practicies for the inclusion of limited English proficient students. *Practical Assessment, Research & Evaluation, 5(13).* Available online: http://pareonline.net/getvn.asp?v=3&n=2.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiental factors. *Language Testing, 15,* 1-20.

Samelson, F. (1977). World War I intelligence testing and the development of psychology. *Journal of the History of the Behavioral Sciences, 13,* 274-282.

Schoenfeldt, L. F. (1985). Review of the Wonderlic Personnel Test, In J. V. Mitchell Jr. (ed.). *The ninth mental measurements yearbook.* Vol 2. Lincoln, NE: University of Nebraska Press.

Statistics Canada (1996). 1996 Census figures. http://www.statcan.ca/english/Pgdb/People/Population/demo29d.htm).

Valdes, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias.* Stamford, CT: Ablex Publishing Corporation.

Wonderlic, E. F. (1992). *Wonderlic Personnel Test User's Manual.* Libertyville, IL: E. F. Wonderlic.

Footnote

1A second analysis of WPT scores was conducted in which the sample size was reduced to 109 by removing participants with higher English and lower French second-language OCT and SEQ scores. The results were the same, with the exception that LEP participants did not score significantly lower than English mother tongue participants on the English WPT. The key point is that, once again, the gain in scores from extra time was only of significant benefit to LFP participants on the French WPT.

Author Note

We thank Dr. Richard Kruk for advice on second language proficiency testing, Denise Bernier for pointing out translation ambiguities, and four reviewers for helpful critical comments. Send correspondence to Stuart J. McKelvie, Department of Psychology, Bishop's University, Lennoxville, Québec J1M 1Z7, Canada. Electronic address is smckelvi@ubishops.ca.

Descriptors: Bilingual Education; Test Format; Evaluation Methods; Intelligence Tests; Language Proficiency; Time Factors [Learning]

Citation: Mullane, Jennifer & Stuart J. McKelvie (2001). Effects of removing the time limit on first and second language intelligence test performance. *Practical Assessment, Research & Evaluation*, 7(23). Available online: http://PAREonline.net/getvn.asp?v=7&n=23.