Applied H.R.M. Research, 2003, Volume 7, Number 2, 41-48

Cognitive Ability Testing and Employment Selection: Does Test Content Relate to Adverse Impact?

Peter A. Hausdorf
University of Guelph

Manon Mireille LeBlanc
Queen's University

Anuradha Chawla
University of Guelph

Cognitive ability tests are commonly used in the employment selection process. Despite their criterion-related validity, they also demonstrate adverse impact (Campbell, 1996; Hartigan & Wigdor, 1989; Hunter & Hunter, 1984). This study assessed two cognitive ability tests (one with verbal, numeric and spatial subtests - the General Aptitude Test Battery and the other predominantly spatial - the Raven's Standard Progressive Matrices) as selection tools for police constable applicants with a specific focus on adverse impact. Both the Raven's Standard Progressive Matrices (SPM) and General Aptitude Test Battery (GATB) demonstrated adverse impact against minority candidates. Implications for research and practice on the use of cognitive ability testing in employment selection are discussed.

One of the most serious criticisms of general cognitive ability tests raised by psychometricians and human rights advocates is that, in spite of being able to predict job and training success (Campbell, 1996; Hartigan & Wigdor, 1989; Hunter & Hunter, 1984), these tests demonstrate adverse impact against minorities (e.g., Campbell, 1996; Hartigan & Wigdor, 1989). In other words, when cognitive ability tests are used in selection decisions as part of a top-down hiring process, majority-minority group differences in test scores invariably lead to lower job selection rates for minority groups because they score lower on average than majority group members (Hartigan & Wigdor, 1989; Jensen, 1980; Schmidt, 1988).

One such test, the General Aptitude Test Battery (GATB), has been widely used as a measure of cognitive ability in employment selection. Although it does predict job performance (Hartigan & Wigdor, 1989; Kirnan & Geisinger, 1990), it has demonstrated differential prediction and adverse impact against African Americans (a 1 SD mean difference) in the U.S. (Hartigan & Wigdor, 1989; Sackett & Wilk, 1994; Wigdor & Sackett, 1993).

To reduce adverse impact, alternative "culture-fair" tests of intelligence have been developed and are used for the assessment and selection of people across minority groups (Anastasi & Urbina, 1997; Arvey & Faley, 1988). Characteristically, culture-fair tests minimize the emphasis on verbally-based test items, requiring no explicit use of language or reading (Anastasi & Urbina, 1997; Arvey & Faley, 1988; Jensen, 1980). Time constraints are often also removed (Anastasi & Urbina, 1997). Elimination of these aspects supposedly reduces the impact of differences in native language, literacy level, and values attached to rapid performance between cultural groups (Anastasi & Urbina, 1997; Arvey & Faley, 1988; Jensen, 1980). Instead, greater emphasis is placed on visuospatial or abstract

reasoning. The assumption is that group differences will be eliminated because factors pertaining to ethnocentric experiences have been removed from the test (Arvey & Faley, 1988).

The Raven Progressive Matrices, one widely used group of culture-fair tests of intelligence, are available in three forms each appropriate for different ranges of ability. The Standard Progressive Matrices (SPM) has been designed for individuals of average ability and is the focus of the current research. The test manual states that the SPM is a "test of observation and clear thinking" (pg. 3, Raven, Court & Raven, 1978). It is designed to measure the educative component of "g" (a general intelligence factor). In other words, past experiences and verbal communication are not assessed by the test.

Although there is a large body of research on the Raven tests, the studies have focused largely on assessment in educational settings rather than employment selection (Llabre, 1990). Arvey and Faley (1988) suggest that culture-fair tests, in general, have failed to reduce adverse impact. However, studies on the SPM are dated and are not specific to the use of the SPM in the context of employment selection. As such, research that clarifies the validity and adverse impact of the SPM in relation to a traditional cognitive ability test such as the GATB would be informative. The current study was designed to address the following questions:

1. Does the SPM eliminate the influence of verbal skills on test performance?

2. Does the SPM demonstrate less adverse impact than the GATB?

Method

A total of 770 applicants (n = 672 Majority group members – white applicants; n = 98 Minority group members) for the position of police constable completed all measures for the study. The applicants came from two large Canadian cities, and were combined into one sample due to a lack of differences between the applicant groups. Minority group candidates represented 13% of the final sample and were from heterogeneous backgrounds: 15 West Asian/North African, 13 South Asian/Indo-Pakistani, five other South East Asian, 24 Black, five Central/South American, 14 mixed, eight Filipino, 11 Chinese, one Oceanic, one Korean, and one Japanese. The minority data were organized into two groups: in one, all minority members were grouped together because of their small individual sample sizes and Canadian employment equity guidelines require combining these groups when evaluating adverse impact statistics; the other group excluded Orientals (Korean, Chinese and Japanese) as some research supports their superior visuospatial skills (see Vernon, 1982 for an excellent review).

All participants were over the age of 18 and had a minimum of Grade 12 high school education. The GATB (Verbal, Spatial and Numeric subtests) and SPM were administered in accordance with test publisher guidelines (Nelson Canada, 1986), with the exception of a one and a half hour time limit imposed for practical reasons on the SPM.

Data Analyses and Results

Descriptive statistics are presented in Table 1 separated by group membership. The overall mean score on the two tests were 263.47 (*SD* = 34.93) for the GATB – GVN and 50.16 (*SD* = 5.66) for the SPM. Both individual and composite scale scores on the GATB were normally distributed. The overall score on the SPM was significantly negatively skewed ($z = 17.45$, $p < .001$) and kurtotic ($z = 26.63$, $p < .001$) with 63% of scores falling in the 50 to 60 range (60 being the highest possible score).

Question 1

The SPM scores were regressed on the three GATB subscales (Verbal, Spatial and Numeric) with the complete minority sample to determine whether the SPM eliminates the impact of verbal abilities. All together, 23% of the variability in SPM scores was predicted by the IVs ($R = .48$, $F(3, 751) = 75.08$, $p < .001$) with spatial scores ($\exists = .27$, $t(751) = 7.71$, $p < .001$) and numeric scores ($\exists = .31$, $t(751) = 8.96$, $p < .001$) contributing significantly but not verbal scores ($\exists = .05$, $t(751) = 1.32$, $p = .19$). These results did not change significantly when the analysis was conducted with the Oriental participants (Korean, Chinese and Japanese) removed from the minority group, $R = .48$, $F = (3, 738) = 73.21$, $p < .001$.

Question 2

Adverse impact comparisons for the GATB-GVN and SPM were tested against two different criteria: (a) mean differences in performance on the two tests, and (b) the four-fifths rule applied to selection ratios of .10, .50 and .90 (consistent with a top-down selection process). The results of these comparisons are shown in Tables 1 and 2.

Table 1: Differences in Majority and Minority Applicant Test Scores

	Majority Applicants			Minority Applicants			Comparison	
Test	Mean	SD	N	Mean	SD	N	t	d
Standard Progressive Matrices	50.43	5.41	672	48.26	6.85	98	3.01**	.39
General Aptitude Test Battery								
Verbal	87.70	13.58	672	79.98	16.00	98	5.13***	.55
Numerical	88.22	13.70	672	84.48	15.10	98	2.49*	.27
Spatial	98.72	16.82	672	92.57	20.10	98	2.88**	.36
V + N + S	265.80	33.89	672	247.51	37.88	98	4.92***	.53

Note: *** $p < .001$, ** $p < .01$, * $p < .05$; t values are based on equal variances for all measures except SPM and GATB – Spatial which are based on unequal variances due to significant results for Levin's test of equality of variances

Two MANOVAs were conducted to test for significant mean differences in performance between majority and minority groups (IV) on the SPM and the GVN (DVs). The first MANOVA contained the complete sample of minorities whereas the second MANOVA used the reduced sample of minorities (with Koreans, Chinese and Japanese participants removed).

For the complete minority sample, the combined dependent variables were significantly affected by group membership, F (2, 755) = 15.74, p< .001 (using Pillai's Trace criterion). Univariate analyses of the dependent variables revealed a significant main effect of group membership for both the SPM, F (1, 756) = 18.15, p < .001 and the GVN, F(1, 756) = 25.53, p < .001. With the GATB subtests separated, univariate analyses revealed a significant main effect of group membership for the verbal (F (1, 764) = 24.57, p < .001), and spatial subtests (F (1, 764) = 10.75, p < .001) but not the numeric subtest (F (1, 764) = 4.60, p > .0125). For the reduced minority sample the combined dependent variables were significantly affected by group membership, F (2, 742) = 19.91, p< .001 (using Pillai's Trace criterion). Univariate analyses of the dependent variables revealed a significant main effect of group membership for both the SPM, F (1, 743) = 23.66, p < .001 and the GVN, F(1, 743) = 31.43, p < .001. With the GATB subtests separated, univariate analyses revealed a significant main effect of group membership for all three subtests: Verbal (F (1, 752) = 31.18, p < .001), Spatial (F (1, 752) = 17.29, p < .001), and Numeric (F (1, 752) = 8.93, p < .01). The exclusion of Orientals raised the numeric score difference to significance.

The effect size of the mean performance differences between groups was estimated by calculation of the standardized difference score (d score) for both tests. The resulting d values for the two tests were compared to determine which test demonstrated greater adverse impact. Standardized d scores and selection ratios were calculated for the complete sample (both majority and minority groups). According to Howell (1997), d scores can be interpreted within the following ranges: 0 - .20 = small, .21 - .50 = moderate, .51 - .80 = considerable, > .80 = severe. Based on these guidelines, the standardized d scores indicate a moderate and comparable degree of adverse impact on the SPM and GATB-GVN composite, Verbal and Spatial subtests. Selection ratios as high as .90 for either test are necessary to avoid violation of the four-fifths rule.

Table 2: Predicted Adverse Impact Statistics Using the 4/5s Rule

Selection Ratio	Test Battery	
	SPM	GATB
90%	.91	.87
50%	.76	.61
10%	.71	.57

Note: Values represent the minority hiring rate divided by the majority hiring rate.

Discussion

The primary focus of the current study was to examine the viability of the SPM as an alternate test of cognitive ability relative to the GATB in the context of employment selection. The SPM does eliminate the impact of verbal skills in test performance. However, the SPM did demonstrate mean differences in performance between groups, which translates into a moderate degree of adverse impact within a top-down hiring process (with the exception of the 90% selection ratio).

Implications

Based on the lack of improvement with respect to adverse impact and the negative skewness for overall SPM scores (i.e. ceiling effect), the SPM is not an adequate substitute for the GATB. This is not to say that the GATB is the "hands down" winner by comparison, because it still has adverse impact. The search for an alternative to the GATB needs to become a priority for researchers and practitioners.

For practitioners, the GATB represents a viable tool but given its adverse impact it needs to be defended as a bona fide occupational qualification (BFOQ). Therefore, organizations using the GATB must demonstrate its job relatedness through validation (*Uniform Guidelines on Employment Selection*, 1978). Considering the lack of available alternatives for the assessment of cognitive ability, to not use the GATB is to decrease the effectiveness of the selection process. However, because neither the SPM nor the GATB are ideal solutions to the problem of adverse impact, an alternative procedure with considerably less adverse impact should be identified (The Uniform Guidelines on Employee Selection, 1978). Alternatives already exist that have demonstrated validity and less adverse impact (for the structured interview see Campion, Pursell & Brown, 1988; for biodata and personality see Schmitt, Clause & Pulakos, 1996).

Limitations

The results of the current investigation must be interpreted in light of several limitations. First, the small minority group sample size, which required placing individuals from heterogeneous groups together, may have masked variation in performance between different sub-samples of minorities. Prior research has indicated that Orientals (Koreans, Chinese and Japanese) have superior visuospatial skills (see Vernon, 1982) and so these participants were removed from a portion of the analysis. Their removal resulted in the change from a non-significant to significant group difference for the GATB Numeric subtest. This suggests that the analysis for the complete minority sample underestimates group differences when Oriental group members are included and thus this study's results can be considered conservative.

Despite this limitation, placing minority group members into one group is consistent with the Canadian Human Rights legislation for the assessment of adverse impact. This example highlights some of the challenges researchers face when conducting research with categories established through public policy or legislation.

Despite this limitation, these findings represent new data that compare the SPM with the GATB in an employment context, which is informative for both researchers and practitioners. Furthermore, these findings may be useful for future meta-analytic research on the GATB and SPM.

Secondly, as with any field study, the data were not sampled at random from the population. All participants had self-selected to be part of the hiring process for the position of police constable. Therefore, these results need to be replicated in another selection context, using a broader sample. Future research should focus on developing a better understanding of the causes of adverse impact, which can include: test content, the test taker themselves, and even the testing context.

Finally, adverse impact only addresses group differences on the predictor and does not consider group differences on the job. In other words, although adverse impact represents prima facie evidence of discrimination, it does not signify that a test is biased. Future research needs to determine if these group differences translate into differential prediction (Cleary, 1968) or mean differences on job performance (Thorndike, 1971), particularly for the SPM. Despite the absence of job performance data, adverse impact is important in its own right from both Human Rights and organizational diversity perspectives.

Conclusion

The SPM does eliminate the impact of verbal skills on overall test performance; however, this elimination of verbal content is not sufficient to eliminate adverse impact. In addition, although the SPM demonstrates comparable adverse impact to the GATB, its ability to differentiate amongst applicants is weak because of a demonstrated ceiling effect. More research needs to be conducted to determine the causes of adverse impact and to create employment assessment tools, which meet the criteria outlined in the *Uniform Guidelines*.

References

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

Arvey, R.D., & Faley, R.H. (1988). *Fairness in selecting employees (2nd ed.).* New York: Addison-Wesley Publishing Company.

Campbell, J.P. (1996). Group differences and personnel decisions: Validity, fairness and affirmative action. *Journal of Vocational Behaviour, 49,* 122-158.

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41,* 25-42.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115-124.

Hartigan, J.A. & Wigdor, A.K. (1989). *Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery*. Washington, DC: National Academy Press.

Howell, D.C. (1997). *Statistical Methods for Psychology, 4th Ed.* Belmont, CA: Duxbury Press.

Hunter J.E. & Hunter, R.F. (1984). Validity and utility of alternate predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Jensen, A. R. (1980). *Bias In Mental Testing*. New York, NY: Free Press.

Kirnan, J.P. & Geisinger, K.F. (1990). General Aptitude Test Battery. In J. Hogan & R. Hogan (Eds.) *Business and Industry Testing: Current Practices and Test Reviews*. Austin, TX: Pro-ed.

Llabre, M.M. (1990). Standard Progressive Matrices. In J. Hogan & R. Hogan (Eds.) *Business and Industry Testing: Current Practices and Test Reviews*. Austin, TX: Pro-Ed.

Nelson Canada (1986). *Manual for the General Aptitude Test Battery Section 1: Administration and Scoring (Canadian Edition).* Scarborough, ON: Author.

Raven, J.C., Court, J.H., & Raven, J. (1978)**.** *Manual for Raven's Progressive Matrices and Vocabulary Scales.* London, UK: H.K. Lewis.

Sackett, P.R. & Wilk, S.L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49 (11),* 929-954.

Schmidt, F. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behaviour, 33,* 272-292.

Schmitt, N., Clause, C.S., & Pulakos, E.D. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In C.L. Cooper & I.T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology*, (pp. 115 – 140). New York: Wiley.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8 (2),* 63-70.

United States Department of Labour (1978). *Uniform Guidelines on Employee Selection Procedures.* Washington, DC: Author.

Vernon, P. E. (1982). *The abilities and achievements of Orientals in North America*. Toronto, ON: Academic Press.

Wigdor, A.K. & Sackett, P.R. (1993). Employment testing and public policy: The case of the General Aptitude Test Battery. In H. Schuler, J.L. Farr, & M. Smith (Eds.). *Personnel Selection and Assessment: Individual and Organizational Perspectives.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Author Notes

1. This study was conducted while the first author was employed at Hay Management Consultants in Toronto, Ontario. The authors would like to thank Kevin Kelloway and Caroll Robinson for comments on earlier drafts of this paper. The views expressed in this paper are those of the authors and not of the Ontario Ministry of the Solicitor General who commissioned the studies from which the data were obtained.

2. Correspondence should be addressed to:
Dr. Peter Hausdorf
Department of Psychology
University of Guelph
Guelph, Ontario
N1G2W1
E-mail: phausdor@uoguelph.ca
Ph: (519) 824-4120 ext.53976