Standardized Testing and Employment Equity Career Counselling: A Literature Review of Six Tests

Prepared for

The Employment Equity Career Development Office
Public Service Commission of Canada

by Gerald P. Gruber, M.A.

March, 2000

Contents

Note: This document uses hyperlinks. Click on words and phrases in the report that are in blue type to jump directly to their associated glossary entry or web site. To go back to the original point in the text, use the Back arrow, if using a web browser, or the Back button on the Hypertext/Web Links tool bar in WordPerfect.

Executive Summary

This report presents evidence from the published literature on the use of standardized assessment tools in career counselling with Employment Equity group members (i.e., women, visible minorities, persons with disabilities, and Aboriginal peoples). The purpose is to provide information to career counsellors on what is known and not known about several assessment tools when used with members of diversity groups. The report summarizes research on these tests both as it pertains to the general population and to diversity groups. Most of the research to date pertains to general population test usage (i.e., Canadians of European descent) and is not necessarily generalizable to EE group members.

Six assessment instruments (tests) were chosen for the focus of the report. These include the Values Scale, as a measure of work related values, the Jackson Vocational Interest Survey, as a measure of occupational interest, the Myers-Briggs Type Indicator and the NEO Personality Inventory-Revised, as two measures of personality, and finally the Multidimensional Aptitude Battery-II and the Test d'aptitudes informatisé, as two measures of aptitude. These tests were chosen to be generally representative of standardized assessment tools. However, their inclusion in this report should not be interpreted as meaning that they are any better or worse than other tests not chosen for the review.

The Values Scale was developed by a Canadian team as part of the Work Importance Study, a multi-national effort to better understand work values in different cultures around the world. The scale is comprised of 100 items and takes only about 30 minutes to administer, generating results for 20 different values, including Cultural Identity. There are several different versions of the Values Scale, as each participating country developed their own version. However, each version includes 60 percent of the same items, allowing for comparative research across cultures. The reliability and validity evidence for the scale is adequate. A strength of the Values Scale is the breadth of normative data provided, including French versus English, male versus female, and several age categories as well. No research was located specific to the use of the tool with diverse populations.

The Jackson Vocational Interest Survey has received positive reviews, summarized in the body of the report, and although not as popular as the Strong Interest Inventory or the Self-Directed Search, it warrants serious consideration as a general population measure of occupational interest. Reliability and validity data are adequate. The noted lack of predictive validity for the tool is more a matter of research yet to be done than of failed attempts to obtain such evidence. Some authors have cautioned that the reading level is higher than the published grade seven level. Others raise concerns about the representativeness of the normative data, although this concern is an American perspective on a test with strong Canadian representation in the normative sample. There is no appreciable literature on the JVIS with diversity groups.

The Myers-Briggs Type Indicator and the NEO Personality Inventory-Revised were focused upon as examples of personality measures. The MBTI is one of the most used assessment tools in the world. Based upon Carl Jung's personality theory, the instrument identifies four bi-polar types or preferences: Extraversion/Introversion; Sensing/Intuition; Thinking/Feeling; and

Judgement/Perception. Considerable research has been done on the instrument. Although the stability for 4-type combinations fluctuates somewhat over time, the reliability and validity data for the MBTI are similar to other main stream personality tools. The primary criticism of the MBTI has been its use of bi-polar preferences, and the concern that this oversimplifies personality and promotes a view that personality is fixed rather than fluid and subject to personal control. The strengths of the instrument include its ease of administration and scoring and the substantial normative database to assist in interpretation. A limitation in the use of the MBTI with diversity groups is the lack of research on the validity of the tool for use with these populations. It is not available in alternative format.

The NEO Personality Inventory-Revised is the focus of a great deal of personality inventory research today. Based upon the increasingly accepted five-factor model of personality, the NEO PI-R has strong reliability and validity evidence, for example, through comparisons to Holland's RIASEC model and to the MBTI. Intercultural research also suggests that both the five-factor model and the NEO PI-R as a means to assess against this model have increasingly universal applicability. This research also has generated language versions for the NEO PI-R that permit first language administration in many instances where other tests would require assessment in a client's second language or through an interpreter. The NEO PI-R also includes both a self and other report format, allowing flexibility where the client is unable to complete the inventory or where it is desirable for a third party to provide an additional perspective.

Two aptitude measures, the Multidimensional Aptitude Battery-II (MAB-II) and the Test d'aptitudes informatisé (TAI), both of which were developed in Canada, were reviewed. The MAB-II was developed as a measure comparable to the Wechsler Adult Intelligence Scale-Revised (WAIS-R) but which could be administered in group settings (the WAIS-R must be administered one-on-one). Although the instrument has received less research attention than some other general aptitude measures, the reliability and validity evidence for the test are satisfactory. Jackson, the test author, cautions against administering the Verbal component of the test to clients in their second language. Because the Verbal and Performance (non-verbal) components are bound in separate test booklets, the test user has the flexibility to administer just the Performance component in instances where the full test cannot be administered in the client's first language.

Finally, the TAI offers a professionally sound general aptitude measure for a general population of clients whose first language is French. The reliability and validity research, which is limited, indicates that the TAI is psychometrically strong. The TAI uses computer administration and computer adaptation (i.e., with each subtest, items are presented from least to most difficult, and when three consecutive items are answered incorrectly, the program skips to the next subtest, thereby reducing the length of the overall test). As with all previous tests reviewed in this report, validity research specific to the applicability of the TAI and the MAB-II to diversity groups is limited at this time.

Some general principles with respect to the use of standardized assessment with members from Employment Equity groups are offered. Test users are encouraged to look for evidence of reliability and validity for the intended test use. Client perception of test validity should be taken into consideration in choosing assessment tools. Caution is advised in assessing in languages other

than a client's first language, in particular when assessing for verbal aptitude. Various accommodation strategies are discussed, in the context of ensuring that the construct is measured as accurately as possible. Different strategies for interpreting test results, including normative, criterion-referenced, and ipsative interpretation, are discussed, with the point offered that if there is any concern with the applicability of available norms, an ipsative interpretation of results is usually a lower-risk approach. The use of general versus subgroup norms is discussed, as are alternatives to standardized assessment.

Finally, three recommendations are offered. The first is that research into the reliability and validity of career assessment tests for Employment Equity designated groups be encouraged. The second is that, given that test norms for designated groups are unlikely to be made available by test publishers in the short to medium time period, it is recommended that the Federal Public Service take steps to develop "local" test norms for tests used more frequently with these groups. And, the third recommendation is that the Federal Public Service look for collaborative opportunities with commercial test publishers to foster research and test development work into career assessment test usage for diversity groups.

Overall, the report concludes that there is a lack of validity research specific to the use of these tests with diversity groups in a career assessment context. This is not to say that such tests should not be used with members of EE groups, but rather, if used they should be administered with understanding based on knowledge of the ultimate purpose of the test (e.g., using ipsative interpretation, putting the results in the context of other career and cultural data, and involving the client in the assessment process as much as possible).

Introduction

The purpose of this document is to summarize research evidence pertaining to the use of standardized assessment (testing) in career counselling of Employment Equity (EE) designated group members (i.e., Aboriginal peoples, persons with disabilities, visible minorities, and women). Due to the variety of terminology used in the literature to describe diversity groups, for purposes of this report the terms used by the authors of the original research will likewise be used in this report when summarizing their research. Career counsellors have requested such information in various needs surveys and during pilot presentations of the Employment Equity Career Counselling Course.

It is important to understand that this document is *not* a "how to" guide for career counselling in general or for career counselling of EE group members in particular. Rather, it is a presentation of information for career counsellors to consider in deciding on the use of commercially available standardized assessment tools as part of the career counselling process. That is, it is assumed that the reader is familiar with the general process of providing career counselling to members of diversity groups.

It is likewise important to understand that this document looks at standardized assessment tools for purposes of *career counselling only*, and *not* for personnel selection. Standardized assessment for career counselling provides information to the *individual client* being assessed to assist him/her in making *career choices*, whereas standardized assessment for personnel selection provides information to the *organization* to make *staffing decisions*. Because such staffing decisions can have major impacts on peoples' lives, tests used for such a purpose are subject to certain standards and legal scrutiny that they would not be if used for other purposes. For example, in the Public Service, the use of any psychological test for personnel selection must first be approved by the Public Service Commission. Keep in mind, that what may appear as career counselling could be interpreted by others as personnel selection. For example, if an aptitude measure is used to select candidates to a training course, and the course is needed for a particular job, then by failing the aptitude measure, the candidate is effectively eliminated from consideration for the job. In such a case, the application of the aptitude measure could be interpreted as personnel selection. Consequently, this report should be used to guide the use of tests for career counselling purposes only. If a reader of this report wishes to determine the suitability of a test for personnel selection, they should consult their Human Resources directorate or the Public Service Commission for advice.

The Employment Equity Career Development Office (EECDO) of the Public Service Commission of Canada is the sponsoring agency for this project. Its mandate is to facilitate the effective career development of Employment Equity group members with the Federal public service. The EECDO is funded through Treasury Board's Employment Equity Positive Measures Program. The EECDO includes a Centre of Excellence (Resource Centre) where counsellors, Employment Equity coordinators, and managers can access and share current research and best practices. The EECDO has developed a Master's level course on EE career counselling, which is offered in over eight universities and colleges across Canada. The EECDO also conducts analyses of needs of EE career counsellors across the Federal public service.

The Personnel Psychology Centre (PPC) of the Public Service Commission has had a significant role in the preparation of this report, and will continue to do so in following up on the report's findings. The Personnel Psychology Centre is the centre of expertise within the Public Service with respect to standardized assessment. Although primarily involved in the use of standardized assessment for the purposes of personnel selection, the PPC has partnered with the EECDO in developing the frame of reference for this report and will no doubt continue as a key stakeholder in the appropriate application of standardized assessment within the Public Service.

Method

How We Chose the Tests to be Reviewed

There are many more tests available to career counsellors than could be reviewed for this report. Consequently, a strategy had to be adopted for choosing tests for the review. Two options were considered: including as many tests as possible but going into less detail on each, or including fewer tests and going into more detail on each. The latter approach was chosen, with the view that the project could be supplemented in future years with additional tests as time and resources permit. The four major categories of tests for career counselling that were considered are:

- values scales;
- interest inventories;
- personality inventories; and
- aptitude tests.

Although the relationship amongst these four categories of assessment is complex, a simplified way of expressing the key interrelationships from a career assessment point-of-view is that values measures help to understand underlying motivators and cultural context. Interests flow from, and can be understood in the context of these underlying values. If the results of interest measures are ambiguous, for example if there are two competing interest streams, the counsellor may refer back to the values results to provide a basis for selecting between them. Measures of personality and aptitude complete the assessment picture, providing the person a view of their strengths and weaknesses relative to their interests. If occupational interests are supported with appropriate personality and aptitude strengths, then the career choices for the candidate are more straight forward. If interests are *not* supported by demonstrated personality and/or aptitude strengths, then the career counselling process becomes more complex, in terms of understanding the divergency of the data, and in considering the needed development steps to achieve career aspirations.

Although the measurement of interests may appear the most critical assessment variable in the counselling process, it has been argued that a comprehensive approach to assessment is optimal: "The limited overlap between interests, personality, and ability suggest that assessment of each domain provides unique information about the client and that neither form of assessment provides a dependable replacement for the other" (Carless, 1999, p.139). Klein, Wheaton, and Wilson (1997), in discussing career counselling for persons with disabilities, likewise advocate a comprehensive approach to assessment, including vocational interests, vocational aptitudes and intelligence factors, career-related personality factors, career maturity, achievement level, and work evaluation. Katz, Joyner, and Seaman (1999) also argue for multiple assessment, finding that combining the MBTI and the Strong Interest Inventory have greater impact on career planning than using just one or the other tool separately. Suzuki and Kugler (1995) echo the need for a multi-faceted assessment approach for clients from diverse backgrounds.

Within each assessment category, one or two tests were chosen for the review, with the thinking that conclusions for the tests being reviewed may generalize to some degree to other tests within

the same category. Tests were chosen taking into consideration: their perceived popularity amongst career counsellors; the amount of information already known about the test; the future potential for the test, the language availability of the test, and the source/location of the test publisher (i.e., Canadian versus internationally based).

For the values scales category, the *Life Roles Inventory Values Scale* (LRI VS) was selected because it was developed in Canada for the Canadian cultural context and has been used by career counsellors in the public service in the past. For the interest inventory category, the Jackson Vocational Interest Inventory (JVIS) was chosen, as it is currently used by career counsellors and was developed by a Canadian psychologist, Doug Jackson, who has a considerable history in developing high quality tests. In the personality inventory category, two tests were chosen for the review: the *NEO Personality Inventory-Revised* (NEO PI-R) and the *Myers-Briggs Type Indicator* (MBTI). The NEO PI-R was chosen because it is based upon an emerging theory of personality referred to as the five-factor theory. The NEO is also garnering considerable attention amongst personality researchers. It was felt that conclusions about the five factors of personality may generalize to other personality measures which also assess the same five factors. The MBTI was chosen due to its considerable popularity within the career counselling community. Finally, two tests were chosen from the aptitude test category: the *Multidimensional Aptitude Battery-II* (MAB-II) and the *Test d'aptitudes informatisé* (TAI). The MAB-II was chosen because it represents a Canadian-based, professionally developed aptitude measure, and the TAI was likewise chosen because of its Canadian roots, and because it was developed within a French Canadian cultural context.

It should be noted that a test's inclusion in this review does not mean that it is superior to other tests not included in the review. Nor does the exclusion of a test from this review mean that it is not a good test. For a less detailed review, but of a wider range of tests that may be applicable to career counselling of diverse populations, the reader is directed to Eby, Johnson, and Russell (1998) and to Samuda (1998). Finally, for a review of the applicability of Public Service Commission tests for personnel selection (primarily tests of abilities and skills), as they apply to diversity groups, see Cronshaw (1999).

Sources for the Literature Review

The literature review was conducted with the assistance of the Public Service Commission (PSC) Library. Searches were conducted of the PsychInfo database and the PSC Library holdings using combinations of the following three keyword search term categories:

1. The actual test names or typical abbreviations;

2. One of the following terms/phrases: career counselling/counseling; career development; vocational counselling/counseling; assessment; or measurement; and

3. One of the following terms/phrases: employment equity; affirmative action; protected groups; persons with disabilities; visible minorities; Aboriginal peoples; women; diversity; multicultural, or equal employment opportunity.

Additionally, test review serials, including the *Mental Measurement Yearbook* series and the *Test Critiques* series were also searched for published reviews of the six tests in question.

Means by Which Tests Were Reviewed

For each test category, we first present relevant information relating to all tests in the category, and address the issue of why one would use such a generic test in career counselling of Employment Equity members. We then provide information specific to the test(s) reviewed for the category. This test-specific information is organized into four sections: (1) an "At-a-Glance" summary of the test, which offers the reader a snap-shot of the test's critical characteristics; (2) a summary of reliability evidence; (3) a summary of validity evidence; and finally (4) a "Critique and Application" section is provided, which focuses more directly on literature describing the test's usage with diversity groups. After the sections describing each test category and the specific test(s), a "General Principles" section discusses issues and procedures that are applicable to standardized assessment of diversity groups in general, regardless of test category. Finally, some recommendations are offered, primarily with respect to future test development and validation work within the Canadian public service.

Values Assessment

Why Assess Values

We shall begin with a distinction between values and interests. Macnab, Fitzsimmons, and Casserly (undated) state that, "[v]alues are the objectives sought in behaviour, and interests are the activities in which the values are sought", (p.6). For example, someone may value creativity, and as a result may demonstrate interest in behaviours such as creative writing or designing new products. Therefore, values underlie interests, and provide important contextual information for understanding and interpreting interests. Macnab et al. note that this is especially important when the client's assessed or expressed interests are unclear. If, for example, a client is pulled in different directions by competing interests, the identification of underlying values may help the client to chose amongst these competing interests in terms of which will help attain the underlying values.

One value that takes on increased importance in counselling EE members is cultural identity - the degree to which the client's values are aligned with their culture of origin versus that of the dominant society. Bowman (1993) notes that visible minority members may vary considerably in terms of their cultural identity. Someone who recently immigrated to Canada may have stronger ties to the culture of their country of origin than someone who is a third or fourth generation Canadian. Further, Bowman observes that within a person's lifespan, cultural identity may also vary. Initially, the person may conform to the predominant culture. Later they may resist the predominant culture and immerse themselves in the culture of their own subgroup. Later yet, they may adopt a cultural value that embraces their own subgroup's uniqueness and also which accepts and works well within the predominant culture.

Kim, Atkinson, and Yang (1999) note that the process of acculturation includes two components: behaviours and values, and that behavioural acculturation may occur more quickly than values acculturation, simply because behavioural acculturation may be needed in order to survive economically in the dominant society, whereas there is no compelling reason to adopt the dominant society's values. Consequently, values assessment is a critical component of assessing a client's cultural identity. Kim et al. summarize research that indicates that not only are clients' values important to understand, but also that (Asian American) clients prefer counselors with similar values to their own. Ridley, Li, and Hill (1998) likewise draw attention to the interaction amongst three variables: the client's culture, the counsellor's culture, and the cultural context within which the assessment takes place. (Note: The Asian Values Scale, presented by Kim et al. is *not* part of the Work Importance Study, cited below, but was developed independently.) In summary, the use of a values assessment tool is an effective way to help determine a client's cultural identity as well as other values that impact and shape career interests.

Life Roles Inventory Values Scale

Life Roles Inventory Values Scale At-a-Glance

Item-types: 100 statements describing the person's current or future activities/accomplishments, rated on a scale with four anchors: of little or no importance; of some importance; important; or very important
Subscales: 20 scales (5 items per scale) assessing the values of: Ability Utilization, Achievement, Advancement, Aesthetics, Altruism, Authority, Autonomy, Creativity, Economics, Life Style, Personal Development, Physical Activity, Prestige, Risk, Social Interaction, Social Relations, Variety, Working Conditions, Cultural Identity, and Physical Prowess
Languages available: Canadian English and Canadian French. Other versions with 60 percent of the same items are available in various languages, developed in other countries, including Australia, Belgium, Italy, Japan, Poland, Portugal, South Africa (English, Afrikaans, and African language), United States, and Croatia.
Reading level: Not indicated. Manual indicates Values Scale appropriate from junior high through adulthood.
Administration time: 30 minutes (untimed)
Test user qualifications: no professional credential requirements listed in Test Documentation
Publisher: Psychometrics Canada Ltd.
Author: Fitzsimmons, G.W., Macnab, D., and Casserly, C.
Forms available: One form
Norms available: All Canadian norms: French Adult Female, French Adult Male, English Adult Female, English Adult Male, French Grade 10 Female, French Grade 12 Female, French Adult Professional Occupations, French Adult Clerical Occupations, French Adult Skilled Occupations, French Adult Unskilled Occupations, English Adult Professional Occupations, English Adult Clerical Occupations, English Adult Skilled Occupations, English Adult Unskilled Occupations, French Grade 10 Males, French Grade 12 Males, English Grade 10 Females, English Grade 12 Females, English Grade 10 Males, English Grade 12 Males, English Post-Secondary Females, and English Post-Secondary Males
EE group norms available: Women, broken down by language (French/English) and by age (current academic level)

The Life Roles Inventory Values Scale (LRI Values Scale) is a measurement tool to assist in the assessment of values. It was developed as part of a multinational project called the Work Importance Study. Each participating country (e.g., Australia, Canada, Portugal, Spain, U.S.A.) developed their own values scale that included 100 items assessing 20 values. Sixty percent of the items were common across all international versions of the scale, to facilitate comparative research. The project, including the Canadian contribution, is well documented by Super and Šverko (1995). Overall, much of the Work Importance Study concluded that people from different cultures around the world tend to have similar values.

The Canadian scale is referred to as the Life Roles Inventory Values Scale or LRI Values Scale, to distinguish it from the U.S.A. and other versions. (If you come across a reference to the Values Scale, check to find out which national version is being cited.) One of the twenty values subscales, the Cultural Identity scale, was included in all international versions of the Values Scale as a result of the efforts of the Canadian project team, supported by two Canadian work values literature reviews, one covering French Canadian literature (Bujold, 1980), and the other English literature

(Casserly & Cote, 1980).

Reliability

The reliability of the LRI Values Scale is reported in the Technical Manual (Fitzsimmons, Macnab, & Casserly, 1985). The internal consistency measure of reliability is reported for the 20 values subscales for English and French versions, for adults as well as grade 10 and 12 students. Additionally, coefficients are provided for English post-secondary students. Coefficients range from a low of .60 for French grade 10 and 12 Life Style and French grade 10 Cultural Identity subscales to a high of .90 for the English grade 12 Altruism subscale. The median average internal consistency for both French and English adults is .80, which is acceptable.

In terms of test-retest and parallel forms reliability, grades 10 and 12 students were administered the LRI Values Scale twice with 4 to 6 weeks between testings. The median test-retest correlation for the 20 subscales was .65 for the French sample, and .69 for the English sample, with the French Cultural Identity scale the lowest at .53 and the French Physical Prowess scale the highest at .83. Finally, the degree to which the French and English versions were parallel was investigated by administering both language forms to a group of bilingual students. The median subgroup correlation was .74, ranging from .62 for the subscale of Achievement to .88 for Physical Prowess.

Validity

Fitzsimmons et al. reports construct validity evidence for the LRI Values Scale via a factor analysis. They conclude that the factor analysis results support the construct validity of the Values Scale, in that on 14 of 20 resultant factors, the highest loading items (those that correlated most closely with the factor) were from the corresponding subscale which appeared to define the factor. In contrast, Rousseau (1989) questions the construct validity of the U.S. version of the Values Scale, on the basis of factor analytic results showing considerable overlap amongst the scales.

Fitzsimmons et al. report convergent validity research where on eight of the LRI Values Scale subscales, results are similar to those obtained on comparable subscales for three other measures (i.e., the Minnesota Importance Questionnaire, the Work Aspect Preference Scale, and the Work Values Inventory). Additionally, they report discriminant validity research that supports the notion that values and interests are related but different. Finally, they report criterion-related validity evidence with a study that distinguished amongst business, education, and rehabilitation medicine students on the basis of the students' results on the Values Scale. The subscales of Advancement, Altruism, Autonomy, Creativity, Economics, Life Style, Physical Activity, Social Interaction, Variety, and Physical Prowess all contributed significantly to the discriminant function.

]

Critique and Application

No research was located that looked at the use of the LRI Values Scale, specifically with designated group members. Walsh, Vacha-Haase, Kapes, Dresden, Thomson, and Ochoa-Shargey (1996) studied the applicability of the American version of the Values Scale for African Americans and Hispanic Americans, investigating differences between 9th graders, 12th graders and college students. Their results indicated that values tended to differ across age groups. For example, college age students placed less emphasis on the two values of Physical Prowess and Prestige than did 9th graders. Walsh et al. also caution against using national norms for ethnic minority college students. Although the American version of the Values Scale has 60 percent overlap with the Canadian version, care must be taken in generalizing Walsh et al.'s results to Canada. Consequently, no firm conclusions can be drawn as to whether and/or how this tool should be used with designated group members.

One of the strengths of the Values Scale is that it provides several normative samples against which to compare individual results (i.e., English vs. French, adult vs. post-secondary vs. grade 12 vs. grade 10, and male vs. female). Consistent with Walsh et al.'s cautions about using general population norms for the American version, it also advisable to use the appropriate normative sample for interpreting results with the Canadian version. If the appropriate norms are unavailable, an ipsative approach to interpreting results could be used (i.e., doing a self-comparison amongst the 20 subscale scores to identify those that are highest or lowest for the respondent, regardless of how they compare to others' scores).

Such an approach (i.e., ipsative interpretation) is recommended by Rousseau (1989), in a review of the U.S. version of the values scale, on the basis of her perception that the normative sample for the U.S. version is not representative of the American population. However, in another review of the U.S. version, Slaney (1989) questions the use of ipsative interpretation on the basis of reported weak test-retest reliability data. Green (1998) and Schoenrade (1998) also note the weak reliability data. As a result, Green recommends group use only, for example for research purposes, (versus individual assessment). It should be noted that for the Canadian version, the test-retest reliability is acceptable, although for some subscales the reliability is marginal.

There is no data available on issues of accommodation for administering the Values Scale. It is not a timed test. Consequently, offering a client more time to complete the test would be an acceptable accommodation.

Other

When contacted by the author, the test publisher Psychometrics Canada indicated that a replacement instrument for the LRI Values Scale is in the process of being developed, and is estimated to be ready for use by early 2001.

Interest Inventories

Why Assess Interests

The assessment of interests is generally considered to be critical in the career counselling process, even though interest as a predictor is only mildly related (.10) to training and job proficiency criteria (e.g., Schmidt & Hunter, 1996). In addition to the Jackson Vocational Interest Survey (JVIS), which we focus upon below, other common vocational interest measures include: the Self-Directed Search (SDS), which classifies people according to Holland's six categories of vocational personality types; the Kuder Occupational Interest Survey, and the Strong Interest Inventory (SII). Research (e.g., Haverkamp, Collins & Hansen, 1994; Ryan, Tracey, & Rounds, 1996; Fouad, Harmon & Borgen, 1997; Getreu, 1997; Day & Rounds, 1998) suggests that there is a certain universality to vocational interest structure, such as that expressed by Holland's vocational personality types, cutting across ethnicity, gender, and socioeconomic status. This conclusion is not unanimous however (e.g., Rounds & Tracey, 1996). Nevertheless, standardized interest inventories, such as the JVIS, may apply across diverse populations, if used with caution and in the context of other client information. For example, Fouad and Spreda (1995) advise that, "[c]ounsellors must understand the client's cultural context and socialization history and make sure that they use the inventory to expand vocational options, not limit them" (p.465).

Although interest inventories are designed to be completed without taking into consideration one's training/experience, or lack thereof, respondents may still allow their perceptions of their skills and abilities to affect their expressed interests. For example, Jackson (2000) relates how certain interest inventory data has correlated with scholastic aptitude: scores on the Verbal subtest (SAT-V) of the Educational Testing Service Scholastic Aptitude Test correlated (.21) with the JVIS Author-Journalism subscale, and scores on the Quantitative subtest (SAT-Q) correlated (.37) with the JVIS Mathematics subscale. One cannot say whether specific interests motivated these respondents to excel in related academic studies, or whether success in their studies prompted them to report interest in these occupational areas. The implication of these findings is not clear, as this problem is endemic to all interest inventories. At a minimum, counsellors should be aware that clients may mistakenly complete interest inventories on the basis of their current skills and abilities, and if this is suspected, the counsellor may wish to explore other occupational themes with the client that may not have emerged in the interest inventory results.

Jackson Vocational Interest Survey (JVIS)

Jackson Vocational Interest Survey At-a-Glance

Item-types: 289 pairs of work-related activities; subject must choose that activity from each pair which they most prefer
Subscales: 34 Basic Interest scales, comprised of 26 work roles (Creative Arts, Performing Arts, Mathematics, Physical Science, Engineering, Life Science, Social Science, Adventure, Nature-Agriculture, Skilled Trades, Personal Service, Family Activity, Medical Service, Teaching, Social Service, Elementary Education, Finance, Business, Office Work, Sales, Supervision, Human Relations Management, Law, Professional Advising, Author Journalism, and Technical Writing) and 8 work styles (Dominant Leadership, Job Security, Stamina, Accountability, Academic Achievement, Independence, Planfulness, and Interpersonal Confidence); also 10 occupational themes (Expressive, Logical, Inquiring, Practical, Assertive, Socialized, Helping, Conventional, Enterprising, and Communicative)
Languages available: English, French, Spanish
Reading level: 7th grade level, as reported in the manual
Administration time: 45 to 60 minutes
Test user qualifications: relevant courses or training in psychology, counselling, or testing, or related work experience
Publisher: Research Psychologists Press
Author: Douglas N. Jackson
Forms available: available as hard copy or can be completed over the internet at www.jvis.com (with a credit card payment) or on a stand alone windows-based PC
Norms available: male, female and combined, Canada and U.S. data
EE group norms available: women

The Jackson Vocational Interest Survey (JVIS) consists of 289 items. The test can be administered in hard copy, stand alone windows-based PC, or via the internet. The results are broken down by 34 Basic Interest scales, comprised of 26 work roles and 8 work styles. The same items can also be scored to produce 10 occupational themes (comparable to, and an expansion on, Holland's six occupational personality types).

Reliability

Measures of internal consistency and test-retest reliability for the 34 Basic Interest scales are provided in the test manual. Measures of internal consistency ranged from .56 for Accountability and Independence to .88 for Mathematics and Medical Service. One-week test-retest reliability coefficients ranged from .72 for Independence to .91 for Social Service. Four-week test-retest reliability coefficients ranged from .69 for Independence to .92 for Social Service.

Additionally, reliability is reported for the 10 General Occupational Themes. Measures of internal consistency range from .70 for Socialized to .93 for Logical. One-week test-retest reliability coefficients range from .82 for Assertive to .92 for Communicative. Four-week test-retest reliability coefficients range from .83 for Assertive, Socialized, and Communicative to .93 for Enterprising. Overall, these data suggest that the JVIS has acceptable reliability.

Finally, individual reliability is reported in Berk and Fekken (1990). Described as a "within session profile stability index", the measure is obtained by splitting the subtests in half, and calculating the correlation between the two half-scores using the 34 pairs of half-scores (one pair for each subtest). Berk and Fekken report mean reliability coefficients of .84 and .87 (corrected by the Spearman-Brown formula) for two testings of a sample of 95 university students.

Validity

Jackson presents construct validity findings by correlating subscale scores on the JVIS with the Strong Vocational Interest Blank. The results are generally supportive of the JVIS assessing similar occupational interests as the Strong. For example, Medical Service and Physical Science scales can be found on both the JVIS and the Strong Vocational Interest Blank. For a female sample these subscales correlated .75 and .74 respectively. Another study investigating construct validity demonstrated that university students were more inclined to volunteer to participate in experiments that were related to work roles that they preferred (as measured by the JVIS). On a more critical side, Jepsen (1992) notes the absence of predictive validity evidence, for example, that JVIS scores predict future occupational satisfaction.

Critique and Application

Beyond normative data for males versus females, no research is reported in the manual on how the JVIS performs with EE designated populations. Research on other interest inventories, such as the Strong Interest Inventory (Lattimore & Borgen, 1999), and its predecessor, the Strong Vocational Interest Blank (Borgen & Harper, 1973) shows little difference between cultures in these tools' abilities to predict occupation. Nevertheless, Jackson, aware of the lack of research on the JVIS with diversity groups, calls for future research into several areas including: interests of persons with disabilities, career interests of women, and cross-cultural and ethnic diversity of career interests.

Overall, in spite of the lack of predictive validity research, test reviewers comment favourably on the JVIS. Brown (1989) who reviewed the JVIS for the 10th Mental Measurements Yearbook, compliments the developmental process for the JVIS, "The care with which the JVIS was standardized is exemplary... This level of factor integrity is excellent for a test of this type". Brown also finds the computer-generated results very useful, in that it links to potential occupations and provides references for additional research into identified occupational titles. Shepard (1989) indicates that the JVIS "has the potential to be one of the best measures of educational/occupational interests".

On the more critical side, Shepard also notes that the general norms of the JVIS are inadequate, although this criticism may apply more to use of the test in the U.S. than in Canada, in that a sizeable proportion of the normative sample came from Canada. Shepard also offers the opinion that the vocabulary for the test items is too sophisticated for many high-school students. Jepsen (1992) likewise, expresses caution with respect to the vocabulary level of the items (reported in the manual as seventh-grade reading level). Juni and Koenig (1982) criticize the JVIS on two

counts. First, in constructing the forced-choice item pairings, not all combinations of scale items were used (there are 34 subscales and 17 items per subscale). Consequently, certain outcomes are unlikely to occur as a result. Second, whether items were presented first or second in the pairing was not varied within certain subscales. If someone has a response set, for example, to choose the first or second in a pairing, their results may be in error. In spite of these shortcomings, there is much to like about the JVIS, as a general measure of occupational interest. With respect its use with EE group clients other than women, there is no discernable research at this time.

Personality Measures

Why Assess Personality

Previously, literature was cited as supporting a multi-measure approach to career assessment. In considering the assessment of personality, the career counsellor may question what it will provide beyond a good career interest inventory. Research (e.g., Costa, McCrae & Kay, 1995) suggests that measures of personality traits and interests tend to correlate with one another, but not enough to suggest that they are interchangeable. Consequently, given the increasing evidence for the validity of personality in predicting training and job proficiency as summarized below, the inclusion of a valid personality inventory in the career assessment battery is appropriate.

In the past, the validity of personality measures was called into question. More recently, however, evidence is building to indicate that personality measures do have a direct relationship to training success and job performance. In particular, two meta-analyses (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991) have looked at the validity of the five-factor model of personality (i.e., Neuroticism/Emotional Stability, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness). Barrick and Mount meta-analyzed 117 studies, with a combined total sample size of 23,994. They observed corrected mean validity coefficients ranging from .03 (for Openness to Experience predicting job proficiency), to .26 (for Extraversion predicting training success). Tett et al. meta-analyzed 97 studies, with a combined sample size of 13,521, and found corrected validity coefficients ranging from .16 (Extraversion) to .33 (Agreeableness). Judge, Higgins, Thoresen, and Barrick (1999) in a primary validity study, found that Conscientiousness predicted both intrinsic (e.g., job satisfaction) and extrinsic (e.g., income and occupational status) success, while Neuroticism negatively predicted extrinsic success. Although, these findings tend to vary according to the personality dimension and the criterion predicted, the validity evidence for personality testing is gaining strength.

Research on personality inventories, in particular those which assess the five-factor model, generally supports the universal nature of personality spanning divergent cultures. For example, Collins and Gleaves (1998) found that the five-factor model, as assessed by the 80 Bipolar Adjective Checklist, fit Caucasian and African American prison guard applicants equally well. Jackson, Guthrie, Astilla, and Elwood (1983) found that the Personality Research Form effectively assessed personality in a Phillipine sample. On the other hand, Ghorpade, Hattrup, and Lackritz (1999), offer evidence that personality structure is *not* equivalent, at least when comparing samples of American and (East) Indian respondents. Nevertheless, Hogan, Hogan, and Roberts (1996) argue that “there is no evidence whatsoever that well-constructed personality inventories systematically discriminate against any ethnic or national group”.

Finally, research suggests that at least some designated groups *perceive* personality tests to be more valid than other assessment measures. For example, Chan (1997) assessed the impact of test content (personality versus cognitive ability) on racial sub-group perceptions of predictive validity, when holding test method constant (i.e., paper-and-pencil rather than video-based or situational). The sample consisted of university students. The NEO-FFI personality test and the

Wonderlic Personnel Test were administered prior to a five-item measure of perceived predictive validity. The results indicated that blacks' perceptions of the cognitive ability test's predictive validity were significantly lower than whites' perceptions for the cognitive ability test, but there were no differences in perception for the personality test. It is difficult to say whether Chan's results generalize to other measures of personality. However, in this instance, perceptions of validity, on the behalf of minority group members, tend to favor personality over cognitive ability tests.

Myers-Briggs Type Indicator

Myers-Briggs Type Indicator At-a-Glance

Item-types: forced-choice
Subscales: Preferred style on four bi-polar scales: Extraversion-Introversion; Sensing-Intuition; Thinking-Feeling; and Judgement-Perception, yielding 16 possible "types"
Languages available: English. Unofficial Japanese and Spanish translations exist.
Reading level: eighth grade level, as reported in the manual
Administration time: 20 to 30 minutes (untimed)
Test user qualifications: MBTI Qualification Course
Publisher: Consulting Psychologists Press, Inc.
Author: Katharine C. Briggs and Isabel Briggs Myers
Forms available: Form G (126 items), Form AV (Abbreviated Version, first 50 items only), Form K (Step II, 131 items)
Norms available: percent scores on the four bi-polar scales for 182 different occupations
EE group norms available: none published in the test manual

The Myers-Briggs Type Indicator (MBTI) is made up of items describing pairs of activities, feelings, and words, between which respondents must chose one or the other which most appeals to them. Based upon C.G. Jung's theory of psychological types, the MBTI measures four dichotomous preferences: Extraversion versus Introversion (EI); Sensing versus Intuition (SN); Thinking versus Feeling (TF); and Judgement versus Perception (JP). Jung's theory suggests that for each of the first three variables, we demonstrate one of two possible preferences. (The fourth variable, JP, was added by Briggs and Myers). Each of the four types act independently of the other. Consequently, there are 16 possible combinations of these four types. It is important to note that these four types are not seen as scales or continuums on which a respondent is located. Rather, the person chooses one preference or the other for each of the four types. (This is the basis for some criticism of the MBTI which will be discussed later.)

Reliability

Because the MBTI generates dichotomous results (types), the preferred approach to reliability is to consider how often the same types emerge in repeated testing. However, for comparative purposes, the MBTI manual (Myers & McCaulley, 1985) also presents split-half and coefficient alpha measures of internal consistency using continuous scores for various samples. The reported reliability coefficients appear acceptable, with split-half coefficients ranging from .82 for EI to .86 for JP for Form G overall, spanning age and gender groups. Coefficient alpha reliability coefficients for Form F were in the same range: .76 for TF to .83 for EI and SN.

Test-retest results are reported in the manual as well for various samples over various time periods. In one study, using continuous data, test-retest correlations for Form G over four weeks resulted in reliability coefficients of .93 for EI, .80 for SN, .89 for TF, and .88 for JP. In another study, using dichotomous data with 329 university sophomores (both males and females combined), after two years, the percent of agreement (those who chose the same type) was 74% for EI, 71% for SN, 73% for TF, and 77% for JP. Considering this same sample, 31% retained

the exact same 4-type combination over the two-year period. Although it is somewhat difficult to interpret these reliability data, given the unusual methods used, the results appear reasonable, demonstrating that such preferences are relatively stable over time.

Alternatively, Willis (1984) recommends caution in using the AV form of the MBTI due to low reliability. Also, Druckman and Bjork (1991) in reviewing reliability data for the MBTI express concern that the fluctuation in type over time suggest that caution should be exercised in making career decisions on such data.

Validity

The Manual reports construct validity evidence by way of correlations between continuous scores on the four types with 31 various other measures of personality and interest, including such recognizable personality measures as the California Psychological Inventory and the Minnesota Multiphasic Personality Inventory, and interest measures including the Kuder Occupational Interest Survey and the Strong-Campbell Interest Inventory. There are many more validity coefficients reported in the Manual than can be adequately summarized in this report. However, by way of example, for a sample of 1,218 males and females combined, the EI scale correlated .66 with the Sociability scale on the California Psychological Inventory, as one would expect. Likewise, for a sample of 100 males, the SN scale correlated .34 with the Architecture scale, again in accordance with expectations. However, for the majority of the data presented in the Manual, there is no obvious expected relationship between MBTI type scores and measures of personality and interest. Consequently, the vast number of correlational studies reported in the Manual attest more to the popularity of the MBTI as a focus for research than to its psychometric strengths.

Concurrent validity evidence is provided by Healy and Woodward (1998), who correlated continuous scores on the MBTI with counsellor ratings of clients' career obstacles. Some of the relationships observed included that men with higher Thinking scores tended to be more obstructed by external demands, have inadequate career information, but have less difficulty relating to the career counsellor. There were fewer relationships between women's MBTI scores and career obstacles, which the authors were unable to explain.

Critique and Application

The MBTI is an exceptionally popular assessment tool. Willis (1984) notes that the MBTI is referenced in one or more articles in 106 different journals, including a journal dedicated to the MBTI, the *Journal of Psychological Type*. Additionally, 40 theses and 260 dissertations focus on the MBTI, as do 15 books or chapters in books. (These numbers have no doubt increased since 1984.)

One of the strengths of the MBTI is the comparative data included in the Manual on many different occupations. Sample data for 182 different occupations is presented in the Manual showing the percentage of people in each occupation who prefer one type over another. For

example, some of the occupations that show clear preferences for one type over another include: marketing personnel - 75 percent prefer the Extravert (versus Introvert) type; police and detectives - 85 percent prefer the Sensing (versus Intuition) type; managers - 80 percent prefer the Thinking (versus Feeling) type; and nursing administrators - 79 percent prefer the Judgement (versus Perception) type.

One of the weaknesses of the MBTI is the presentation of type preferences as either/or (e.g., Wiggins, 1989; McCrae & Costa, 1989). Most personality theorists view personality traits as continuous, with individuals possessing more or less of each trait. The concern with presenting traits (preferences in MBTI terminology) as either present or absent, is that it suggests a finality to the results, such that the individual cannot change the way he or she behaves. Willis (1984) states, "There is some concern, however, that an examinee can accept in a too literal sense the description of type as a command for action rather than as another mode for self-understanding. In these cases, behavior may be viewed as static rather than developmental". The MBTI publisher has responded to this concern more recently by providing continuous as well as preference scoring.

Related to this, Coe (1992) recommends against using the MBTI for personnel selection. He reasons that because the MBTI does not measure the "shadow functions" - the preferences opposite on the bi-polars to the chosen types, it cannot indicate how strong someone is in a particular shadow function. For example, if a position requires dealing with people, someone with an Introvert preferred type may still be more capable in dealing with people than someone with an Extravert type, particularly if the Introvert made the effort to develop this non-preferred shadow function. Coe concludes, "because of its limitations and because it is beatable, the MBTI should *not* be used in any part of the selection process".

Druckman and Bjork (1991) in reviewing the MBTI, likewise are somewhat critical of the tool. As indicated in the reliability section above, they argue that the fluctuation in type over time is too high to use as the basis for career counselling. In reviewing validity research of the tool, for example the lack in diagnostic accuracy, they conclude that the evidence does not support the MBTI's widespread use.

The Manual does not address the issue of how or whether the approach to using the MBTI should be modified for designated group members. It does summarize some research indicating that the MBTI types are not culture-specific but are found in various cultures around the world. On the other hand, other research suggests that the MBTI may be less accurate with African Americans than with European Americans (e.g., Posey, Thorne, & Carskadon, 1999), and that different type frequencies emerge between cultures (e.g., Nuby & Oxford, 1998). Posey et al. do not go so far as to advise against using the MBTI with African American clients, nor do they advocate African American norms, but rather advise to be sensitive to cultural factors when interpreting the results, and "ensuring an atmosphere of comfort, trust, and informed, truly voluntary participation before administering the MBTI" (p.21).

There is a literature that looks at predominant MBTI types for various subpopulations. For example, Simmons and Barrineau (1994) considered type for Native Americans (U.S.). The authors found that for a sample of 210 Native American first-year university students, Ss

(sensing) were over-represented amongst males, and Ss and SFs (sensing and feeling) were over-represented amongst females in comparison to normative data. The authors further cite research that college faculty tend to favour intuition as a perceiving function, and suggest that this may result in them instructing their courses using a learning style different from and less conducive to learning for Native Americans. The authors note that this lack of attention to learning styles may account for the higher drop-out rates for Native Americans and the lower enrollment rate of Native Americans in university level programs. They suggest that a decrease in competition and an increase in cooperation, such as small-group class discussions might help align teaching styles to Native American learning styles. Unfortunately, the paper presents no empirical data that such an approach would actually impact drop-out rates or enrollment rates for Native Americans.

Overall, the MBTI is a very popular assessment instrument, which is easy to administer and score. There have been concerns raised about the dichotomous nature of the four types, and some authors have pointed to fluctuating results over time. Finally, preliminary research suggests that MBTI type frequencies tend to vary according to cultural group.

NEO Personality Inventory-Revised

NEO PI-R At-a-Glance

Item-types: 240 statements describing the person which are rated on a scale with 5 anchors: strongly disagree, disagree, neutral, agree, and strongly agree
Subscales: Five domain scores and six facet scores per domain: Neuroticism (including facets of Anxiety, Angry Hostility, Depression, Self-Consciousness, Impulsiveness, and Vulnerability); Extraversion (including facets of Warmth, Gregariousness, Assertiveness, Activity, Excitement-Seeking, and Positive Emotions); Openness (including facets of Fantasy, Aesthetics, Feelings, Actions, Ideas, and Values); Agreeableness (including facets of Trust, Straightforwardness, Altruism, Compliance, Modesty, and Tender-Mindedness); and finally Conscientiousness (including facets of Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, and Deliberation).
Languages available: In addition to English, available "for research purposes" in: Arabic, Chinese, Czechoslovak, Dutch, French, German, Hebrew, Japanese, Norwegian, Polish, Portuguese, and Swedish
Reading level: 6th grade
Administration time: 30 to 40 minutes
Test user qualifications: training in psychological testing and measurement required to interpret results
Publisher: Psychological Assessment Resources, Inc.
Author: Paul T. Costa, Jr. , Ph.D. and Robert R. McCrae, Ph.D.
Forms available: Form S (self-report), Form R (observer-report), both paper and computer versions
Norms available: men, women, and combined, stratified for age, gender, and race (U.S. data)
EE group norms available: women

The Revised NEO Personality Inventory (NEO PI-R) is quickly emerging as one of the most popular personality assessment tools, both from a user and a researcher perspective. This is due, in no small part, to the test's adoption of the five-factor model of personality (i.e., neuroticism/emotional stability; extraversion; openness to experience; agreeableness; and conscientiousness). Prior to the five-factor model, personality theorists had many and varied views as to how many personality traits there were and how those traits were defined. The five-factor model provides a simplified approach to understanding personality. Psychologists have not unanimously endorsed the five factor model to date (e.g., Child, 1998). However, the acceptance of the model has increased substantially over the past 10 years.

Reliability

Internal consistency reliability for the facet scales is reported in the manual as ranging from .56 to .81 for self-reports, and from .60 to .90 for observer ratings. The domain scales, which have more items as they are comprised of six facet scales each, have slightly higher internal consistency reliability coefficients ranging from .86 to .95.

Test-retest reliability for a sample of 31 men and women ranged from .66 to .92 for the facet scales, and were .87, .91, and .86 for the N, E, and O domain scales respectively. In another study, the three-month retest reliability of the NEO Five-Factor Inventory (a 60 item abbreviated

version of the NEO PI-R) for 208 college students was .79, .79, .80, .75, and .83 for the N, E, O, A, and C domains, respectively. Six-year long-term reliability for the N, E, and O domains ranged from .68 to .79 for both self-reports and spousal ratings. Additionally, a study found seven-year retest reliability for the 18 N, E, and O facet scales ranging from .63 to .81, and for the five domain scales ranging from .63 to .81.

In another assessment of the stability of the constructs measured by the NEO over long periods of time, Soldz and Vaillant (1999) reported that for 163 men followed over a 45-year period, measures of Neuroticism, Extraversion, and Openness, assessed at the conclusion of college and then again 45 years later (the latter assessment done with the NEO-PI), significantly correlated with one another. Overall, these data indicate that the NEO PI-R has reasonably strong reliability.

Validity

The manual reports considerable validity evidence. Several studies are cited which report convergent and discriminant validity of the five domain scales with various adjective measures. Others report correlations with other measures of personality and interest including the Self-Directed Search, Myers-Briggs Type Indicator, Personality Research Form, Minnesota Multiphasic Personality Inventory, the revised California Psychological Inventory, Guilford-Zimmerman Temperament Survey, Adjective Check List, and the Interpersonal Adjective Scale, Revised. Barrick and Mount (1993) found that for 205 students, the five domain scores on the NEO-PI correlated positively with corresponding subscales on the Personal Characteristics Inventory. Additionally, several studies cited in the manual investigated the relationships amongst self-report, spousal, and peer assessment scores. Generally, patterns of correlations are supportive of the construct validity of the NEO PI-R.

Several studies have compared results on the NEO PI-R (and previous versions) with assessments based on Holland's RIASEC vocational personality types (i.e., Realistic, Investigative, Artistic, Social, Enterprising, and Conventional). For example, in a correlational study of the NEO (precursor to the NEO PI-R) and the Self-Directed Search (SDS), Costa, McCrae and Holland (1984) found that, for 217 men and 144 women, the Extraversion domain scores correlated primarily with the Artistic, Social and Enterprising SDS subscales, while the Openness scores correlated primarily with the Investigative and Artistic SDS subscales. The Neuroticism domain scores tended to correlate less well with SDS results. Gottfredson, Jones, and Holland (1993) considered the same issue with a later version of the NEO (the NEO-PI) and a different measure of Holland's vocational types (the Vocational Preference Inventory). They found that the Extraversion domain tended to correlate with the Social and Enterprising vocational types, but less so with Artistic, in contrast to Costa et al.'s earlier findings; the Openness domain correlated with Investigative and Artistic types, consistent with Costa et al.; and the Conscientiousness domain correlated with the Conventional type (Conscientiousness was not yet part of the NEO when Costa et al. conducted their comparison to the SDS). Holland, Johnston, and Asama (1994) found similar patterns between the NEO-FFI and the SDS: Extraversion, Openness, and Conscientiousness scales correlated positively with the SDS overall; the Neuroticism scale correlated negatively with the SDS, and the Agreeableness scale did not correlate at all with the SDS. Tokar and Swanson (1995) found that for males and females, Extraversion and Openness

predicted RIASEC types, and for females Agreeableness also predicted RIASEC types. Gottfredson et al. (1993) caution that their observed relationships between the NEO PI-R and Holland's vocational types are not so strong as to warrant using one assessment tool as a substitute for the other.

Finally, De Fruyt and Mervielde (1999) added to the understanding of the relationship between the five factors and Holland's vocational types, by observing that the big-five factors of Extraversion and Conscientiousness predicted employment *status* (i.e., whether employed or not), whereas none of the RIASEC types added significant variance to the prediction. However, the RIASEC scores significantly predicted the *nature* of employment, prompting De Fruyt and Mervielde to conclude, "Although interests might be important to explain applicants' attraction to jobs, their chances of getting jobs are more influenced by their personality traits" (p.719).

Several comparisons have been made between the NEO PI-R and the Myers-Briggs Type Indicator. The Manual reports that Extraversion is comparable in both the NEO and the MBTI; Openness is closely related to Intuition (versus Sensation); Agreeableness is like the Thinking/Feeling type dichotomy; and Conscientiousness parallels the Judging/Perceiving type. MacDonald, Anderson, Tsagarakis, and Holland (1994) report a pattern of correlations between the MBTI type scores and the NEO-PI domain scores, with a sample of 161 female and 48 male university students, which support the Manual's comparisons. Additionally, Furnham (1996) found similar results, although he found that Conscientiousness correlated with both Judging/Perceiving *and* Thinking/Feeling. And, he found that Openness was not only related to Intuition/Sensation, but to a lesser extent, to all four MBTI types.

These data suggest that the NEO PI-R has acceptable reliability and validity. In reviewing the precursor to the NEO PI-R, the NEO-PI, Hess (1992) draws similar conclusions, "The NEO-PI seems to be a reliable and valid measure".

Critique and Application

The NEO PI-R and its predecessors have been favourably reviewed for the most part (e.g., Hess, 1992; Widiger, 1992; Tinsley, 1994; Botwin, 1995). Widiger, for example, notes how reliability and validity research is not simply published in the test manual, but subjected to peer-reviewed journals, as a means to verify the accuracy of the results. He notes that the strength of the NEO lies in its use of the five-factor model, and that at the time of his review, there was still some debate as to whether Openness to Experience was a legitimate factor, or whether it was better described as Intellect. Botwin comments on the improvements made between the earlier version, the NEO-PI, and the current NEO PI-R, "There are few shortcomings in the NEO-PI-R. This version of the instrument admirably addresses many criticisms of the previous version... These scales [the NEO PI-R] should be considered a standard set of useful tools for personality assessment...". Finally, Tinsley concludes, "that the NEO PI-R is one of the best instruments available for the assessment of normal adult personality".

Less favourably, Juni (1995) suggests that the support garnered by the NEO PI-R is not warranted. For example, he questions the translation techniques of the non-English versions of the

test, and whether it is even appropriate/possible to translate the NEO PI-R, which is based upon descriptors of traits primarily grounded in the English language. Juni also criticizes the test authors for drawing linkages between NEO results and psychopathology. He notes that the NEO is presented as a normal range personality measure, and is not designed to assess evidence of psychopathology. This criticism may explain partially why a NEO-4, omitting the Neuroticism scale, was recently released by the test publisher. Juni also offers several examples of where, in his opinion, the wording of items is misleading, or otherwise inappropriate. On the issue of the NEO inappropriately overlapping into areas of psychopathology, Juni's point is valid. There are several well designed tools for assessing psychopathology, such as the Minnesota Multiphasic Personality Inventory-2, or the Millon Clinical Multiaxial Inventory-III, that have served the clinical community well. However, Juni's other criticisms seem to be out of step with other reviewers' comments on the NEO.

The test manual notes that women tend to score slightly higher than men on the Anxiety, Straightforwardness, and Altruism facets. The manual suggests that these differences are small and combined sex norms can be used for most applications. However, separate female versus male norms are available should the counsellor wish to use gender-specific norms. Widiger (1992) in reviewing the NEO-PI, commented favourably on the fact that the NEO-PI's norms included a full range of adults rather than just college students.

Costa & McCrae (1992) note that the NEO PI-R can be administered orally (e.g., to persons with visual impairments) without loss of validity. Another feature about the NEO PI-R that allows for flexibility in administration is that there is a Form S (self-report) and a Form R (observer-report). For example, if it was desirable for a family member to be involved in the career counselling process, that family member could be invited to complete a Form R NEO on the client, as a way to supplement the client's Form S self-report, or instead of a self-report, if the client was unable to complete the form. Unfortunately, as Tinsley has pointed out, the test manual does not report parallel form reliability data between Form S and R. Consequently, caution must be exercised in using Form S and R interchangeably.

We have previously discussed research which is generally supportive of the universality of personality structures in general, and of the five-factor model in particular. With respect the NEO, research is likewise supportive of it's universal applicability. For example, De Fruyt and Mervielde (1996) found that the Conscientiousness scale, using a Dutch/Flemish adaptation of the NEO PI-R correlated positively with educational achievement in a university setting, while Neuroticism correlated negatively with educational achievement for males. In another example, McCrae, Costa, and Yik (1996) administered a Chinese translation of the NEO PI-R to 352 Chinese University of Hong Kong students. The factor analyzed results supported the five-factor model of personality, with 29 of 30 facet scores loading highest on the intended factor. However, Chinese students had lower Extraversion scores overall, and had lower Competence scores combined with higher Vulnerability scores. Consequently, with some noted exceptions, McCrae et al. conclude that personality structures are much more similar across cultures than they are different.

Other

A NEO-4 Inventory has recently (1998) been released. Although little literature is available on it at this time, it would appear to be essentially the same instrument as the NEO PI-R with the Neuroticism scale omitted. The Research Psychologists Press catalogue describes the NEO-4 as appropriate "for use in employment and personal counselling settings". It would appear that the NEO-4 is an attempt to appeal to a segment of the personality assessment market that values a simplified approach to personality, such as used by the NEO PI-R, but which does not wish to expose their respondents to items which assess neuroticism. Consequently, when emotional stability is not an issue for a particular client, the NEO-4 may hold potential for career assessment.

Overall, the NEO PI-R has strong potential for the assessment of personality in a career counselling context. The theoretical underpinnings of the tool (i.e., the five-factor model) are clear and appear universally applicable. Reliability and validity data are generally strong; test reviewers speak favourably about the instrument; and it provides flexibility in administration with both self and third party perspectives. Some authors have gone beyond the intended purposes of the tool, and suggested it can be used in assessing psychopathology, which it clearly is not designed to do. The NEO PI-R has had more intercultural research conducted on it than most commercial tests, and much of this research is favourable. However, with the exception of Women, no normative data is available for Canadian designated EE groups.

Aptitude Measures

Why Assess Aptitude

Measures of aptitude and general ability are consistently shown to be amongst the best predictors of future training and job success (Schmidt & Hunter, 1996). A challenge in using aptitude measures with diversity groups is that some groups, including Blacks (e.g., Bobko, Roth, & Potosky, 1999) and Aboriginal peoples (e.g., McShane & Berry, 1988), have been found to have lower mean test scores (often as much as one standard deviation), in comparison to the majority population. The reasons for these results are not fully understood. For example, some (e.g., Samuda, 1985) argue that tests that produce such differences are unfair, while others suggest that the tests are displaying the impact of group differences in socio-economic opportunities. Frisby (1999) notes that alternative aptitude measures (e.g., portfolio, performance, or authentic assessment) for culturally diverse groups have failed to eliminate such subgroup differences.

Use of ipsative interpretation of standardized measures (see discussion on different test interpretation approaches below) may be part of the solution to this problem. Given that in the career counselling process, there is greater emphasis on learning about the client's strengths and weaknesses relative to him/herself, and less emphasis on comparisons to others, there is less need for normative interpretation, where the problem of subgroup differences emerges. However, measures of aptitude tend to assess a singular construct. Consequently, the expectation that strengths and weaknesses relative to him/herself may predict differential success in various occupations has not been fully realized. Jackson (1998) notes, "[a]lthough factor based theories of intellectual abilities have offered the promise of differential prediction, this promise has not, in general, been realized in practice".

Perceptions of aptitude tests also are a factor to be considered. Because of the higher profile of aptitude measures, clients are more likely to have an opinion, whether positive or negative, on the usefulness of such measures. As summarized above in the section on personality measures, some research indicates that Blacks' perceptions of the validity of aptitude measures is lower than Whites' perceptions. To the degree that the counsellor knows the client's perceptions of the proposed assessment process, information can be offered to overcome misconceptions held by the client about the measure. Or, if the perception is deeply held, the counsellor may chose a different assessment tool, given that this negative perception may itself invalidate the results (i.e., the client may not try as hard to correctly answer questions, or may intentionally falsify responses).

Multidimensional Aptitude Battery II

Multidimensional Aptitude Battery II At-a-Glance

Item-types: Verbal subscales: 174 multiple-choice items; Performance subscales: 161 multiple-choice items
Subscales: Full Scale, Verbal (comprised of Information, Comprehension, Arithmetic, Similarities, and Vocabulary), and Performance (comprised of Digit Symbol, Picture Completion, Spatial, Picture Arrangement, and Object Assembly)
Languages available: English, French, Spanish
Reading level: Not specified. The manual states the MAB was designed for ages 16 and up.
Administration time: approximately 100 minutes (50 minutes for the Verbal half and 50 minutes for the Performance half)
Test user qualifications: "advanced level university course in psychological testing at the Master's level, as well as training under the supervision of a qualified psychologist" (RPP Catalogue)
Publisher: Research Psychologists Press
Author: Douglas N. Jackson
Forms available: hard copy or computer administered
Norms available: sample of 1600 from Canada and U.S.; nine age categories
EE group norms available: none

The Multidimensional Aptitude Battery II (MAB-II) is a measure of verbal and nonverbal aptitudes (cognitive abilities). It was designed as a group-administered alternative to the Wechsler Adult Intelligence Scale-Revised (WAIS-R). (The WAIS-R must be individually administered, and as such is more costly to administer than a comparable group-administered test.)

Reliability

Measures of internal consistency were obtained for the Full, Verbal, Performance and the 10 subscales for a sample of adolescents including 230 males and 285 females. Subscale reliabilities ranged from .70 for 16-year old test-takers on the Arithmetic subscale to .96 for 19 and 20-year old test-takers on the Spatial subscale. Full scale reliabilities ranged from .96 for 16 and 17-year old test-takers to .98 for 18, 19, and 20-year old test-takers.

Test-retest reliability data over a 45-day period is presented in the test manual for a sample of 52 young adult psychiatric patients. This population is different from the general population and different from the clientele that receive career counselling in the Federal public service. However, Jackson, in presenting this data, notes that this is a "stringent evaluation", suggesting that reliability may be higher with normal populations. Test-retest reliability coefficients ranged from .83 for the Similarities subscale to .97 for the Information subscale. The Full scale test-retest reliability was .97. These data suggest that the MAB-II has good reliability.

Validity

Two factor analyses of the MAB-II reported in the test manual support the distinction between the Verbal and Performance subscales. Likewise, Wallbrown, Carmin and Barnett (1988 & 1989) confirmed the identification of general intelligence (g), verbal and performance factors from a sample of 300 adult male felons, and Lee, Wallbrown and Blaha (1990) found additional evidence to support such a factor structure in analyzing the subscale intercorrelation matrix for 3,121 high-school students, as provided in the test manual.

Additional construct/convergent validity evidence is presented in the manual in the way of correlations between the MAB-II and the WAIS-R, using a sample of 145 people. Correlations between comparable subscales on the MAB-II and the WAIS-R ranged from .44 for the Spatial/Block Design subscale to .89 on the Arithmetic and Vocabulary subscales. The Verbal, Performance, and Full scale correlations were .94, .79, and .91 respectively. Further, a factor analysis of the MAB-II and WAIS-R resulted in clearly defined verbal and performance factors, adding additional support to the verbal/performance distinction made in the MAB-II. In contrast, Kranzler (1991) did a factor analysis of the MAB-II with university students, and concluded that the verbal/performance factor structure was not supported. He suggested that these negative results were obtained because, even for the performance tests, certain verbal skills were required, such as the test instructions.

Finally, Kranzler (1991) presents results of how the MAB-II correlates with the Raven's Advanced Progressive Matrices (RAPM), a popular non-verbal measure of general intellectual ability. The overall correlation between the RAPM and the MAB-II was .45. The 10 MAB-II subscales correlated with the RAPM in the range of .13 for the Similarities subscale to .56 for Object Assembly. The results showed higher correlations between the RAPM and the MAB-II Performance subscales than with the MAB-II Verbal subscales, as expected.

Critique and Application

One of the strengths of the MAB-II test manual (Jackson, 1998) is that it provides information on occupations that may be suitable for strengths displayed on the various subscales. However, counsellors must use caution in differentiating amongst career options on the basis of aptitude measures alone, given that aptitude measures tend to assess a singular construct (e.g., general cognitive ability), or at most two or three constructs (e.g., verbal, numeric, and spatial/mechanical ability), which taken as a whole predict training success and job proficiency, but which do not effectively predict more success in one career stream versus another.

Jackson notes in the manual that the Verbal subtests are unsuitable as a measure of verbal aptitude for persons who's first language is not that of the language in which the test is being taken. Consequently, if a client's first language is other than English, French, or Spanish, the MAB-II would not be appropriate for verbal aptitude assessment (nor would most other verbal aptitude measures administered in a client's second language). Jackson notes, however, that the MAB-II verbal subtests, administered in a client's second language, could serve as a measure of a client's current level of functioning in the language being tested.

No literature on how designated groups perform on the MAB-II was located. However, research

(McShane & Berry, 1988) indicates that various Aboriginal groups across North America tend to perform less well on verbal versus visual and spatial components of aptitude measures. There is no conclusive explanation offered for this, although testing in one's second language may contribute to these results. McShane and Berry also point out that there are many different Aboriginal cultures in North America, and therefore, it is *not* appropriate to draw conclusions about individuals on the basis of equivocal group data. Clearly, if someone is taking an aptitude test in their second language, this may impact verbal subscore performance. Consequently, if there is a need to assess aptitude for an Aboriginal client (or any client for that matter), and that person's first language is other than a language in which the assessment tool is available, then it may be prudent to administer only the non-verbal component (which can be done with the MAB-II - not all aptitude tests separate verbal and non-verbal components).

Because each of the 10 subtests are timed at 7 minutes, accommodation by providing extra time to complete the test would be problematic in terms of interpreting the results, especially if interpreting using the published norms (see section below on normative versus ipsative test interpretation). Finally, some reviewers (e.g., Krieshok & Harrington, 1985) have criticized the MAB-II for the limited norms available. As with the JVIS, this criticism may be a function of Krieshok & Harrington taking a U.S. perspective - the Canadian representation in the norms is stronger than the American.

Overall, the available reliability and validity data are strong for the MAB-II. Care must be taken in administering the test to second-language clients, and must also be taken in interpreting the results from the point-of-view of directing the client into certain occupational directions and not others. Its best use in a career counselling environment may be in helping the client to understand their aptitude to succeed in a challenging academic/training program. Care must be taken, however, in using general norms for such interpretation, especially for African Canadians and norms for verbal aptitude for Aboriginal peoples, where research summarized above indicates that mean differences have been found between these groups and the majority population.

Test d'aptitudes informatisé

Test d'aptitudes informatisé At-a-Glance

Item-types: multiple-choice (324 items maximum)
Subscales: Overall, Verbal, Non-verbal, as well as Vocabulary, Verbal Reasoning, Spatial Relations, Mathematical Operations, Arithmetic, Knowledge, Spatial Visualization, Comprehension, Series, Memory, and Perception
Languages available: French
Reading level: Not indicated. Designed for ages 12 and older.
Administration time: approximately 1 hour and 30 minutes
Test user qualifications: advanced training in testing and measurement
Publisher: Le réseau psychotech inc.
Author: Michel Pépin & Michel Loranger
Forms available: computer version
Norms available: sample of 675 people (French speaking) from the province of Québec
EE group norms available: none

The Test d'aptitudes informatisé (TAI) is a computer-administered and scored, as well as computer-adapted aptitude test. The items within each subtest are ordered according to difficulty, and when three items are failed consecutively, the program moves onto the next subtest, thereby reducing administration time. Scores on the overall test, on verbal and non-verbal portions, as well as on the eleven sub-tests are provided as percentiles. The authors advise against interpreting the scores as IQ scores.

Test administration timing is different from typical paper and pencil tests. In order for a test item to be successfully completed, the correct option must be chosen within 60 seconds of the item being displayed on the screen. A test completion time is calculated from the time taken to respond to 30 selected items in the test (6 items from each of 5 subtests).

Reliability

Measures of internal consistency for the subtests range from .79 to .95. Portugais, Daudelin, Loranger, and Pépin, (1995) report test-retest reliability over a five-week period with 44 members of the Canadian Armed Forces (40 percent of the sample were female). Individual subtest reliability coefficients ranged from .33 for Memory and .39 for Spatial Relations to .90 for Comprehension. Verbal, Non-verbal, and General scores achieved test-retest reliability coefficients of .92, .74 and .90 respectively. Larue, Pépin, and Loranger (1996) also report test-retest reliability coefficients for a version of the TAI for children. They report test-retest reliability coefficients of, .82, .72, and .86 for Verbal, Non-verbal, and General scores respectively. These data indicate slightly higher reliability for the Verbal component than the Non-verbal. Of particular concern are the low test-retest reliability coefficients for the Memory and Spatial Relations subtests. On the basis of these results, it would be advisable to exercise caution in interpreting individual subtest scores. The Verbal, Non-verbal and General scores are more stable, as they sum across the individual subtest scores. (Not overly interpreting subtest scores for the

TAI is consistent with previous discussions on the MAB-II, in that differences on aptitude subtests typically do not effectively differentiation between career options.)

Validity

The test manual (Pépin, M. et Loranger, M., 1996) reports criterion-related validity evidence in that Verbal, Non-verbal, and General scores on the TAI correlate with their counterparts in l'Épreuve individuelle d'habileté mentale (EIHM), at the .63, .51 and .75 levels respectively. Larue, Pépin and Loranger (1996) also report correlations between the Verbal, Non-verbal and General scores on the TAI for children and their counterpart scores on the Wechsler Intelligence Scale for Children-Revised (WISC-R), at .77, .54, and .70 respectively. Larue et al. also report significant correlations between TAI Verbal, Non-verbal, and General scores and results in French and Mathematics. Additionally, the manual reports that a factor analysis was conducted, the results of which support the constructs measured by the TAI. The manual also shows correlation coefficients between time to complete the test and the subtest scores. For each of the eleven subtests, there is a negative correlation between average item completion time and the subtest score (i.e., high scorers on the test tended to complete each item more quickly than low scorers). As a result, the test authors argue that item completion time is a measure of success on the test.

Critique and Application

For clients who prefer or need to be assessed in French and for whom computer administration is not a problem, the TAI is a reasonable means to assess general aptitudes. As with most other tests, there is no published literature on the performance of the TAI with designated groups. Consequently, the literature provides no direction or guidance on the use of the TAI specifically with designated group members.

To the degree that the TAI is computer administered, accommodation may be facilitated. For example, it is relatively easier to produce a large print version of a computerized test than of a hard copy test. Likewise, if a test document already exists in computer form, it may be possible to use specialized software that can read the items out loud to a client, thereby facilitating administration for some people with visual impairments. Consequently, there may be higher potential for the TAI and other computer based test instruments in terms of accommodation options.

On the other hand, the scoring system for the TAI takes into consideration the time to respond to each item. Consequently, adjusting the amount of time to complete the TAI could be problematic, as an accommodation option.

Overall, the same qualifiers discussed previously for the MAB-II apply to the TAI, in terms of not over-interpreting subtest scores, assessing verbal aptitude in a client's first language, and using general population norms with caution.

General Principles in Using Standardized Tests in Career

Counselling with Employment Equity Groups

There are some principles and criteria that apply to the use of standardized tests in career assessment that apply whether one is measuring values, interests, personality, or aptitudes. These principles are outlined below.

Reliability

Any measure, in order to be useful, must be reliable, that is, it must consistently produce similar results under repeated measures. Reliability is reported in several different ways. One such way is through measures of internal consistency, which are relatively easy to obtain, but are seen as upper-bound measures (i.e., high-side estimates) of reliability. Test-retest reliability measures are considered to be better measures of reliability, but are more difficult to obtain. Most often reliability research on an assessment tool is summarized in the technical manual, which may be combined with the administration manual. If there is no technical manual, or if the technical manual makes no reference to reliability or only to internal consistency measures of reliability, this suggests a potential weakness in the tool, and may prompt you to consider alternative assessment tools. Additionally, the reliability data should be generated with samples of test-takers that are representative of the populations of interest. For example, test-retest reliability data obtained with a predominantly European Canadian sample, will not necessarily generalize to test-takers from other cultural backgrounds.

Validity

Validity is an equally important characteristic of a professionally developed assessment tool. Essentially, validity refers to the degree to which an assessment tool measures what it purports to measure. There are many different kinds of validity. The most often cited include content, construct, and criterion-related validity. Content validity refers to evidence that a test adequately represents the domain or entirety of the construct being measured. For example, if an interest inventory is shown to cover all major career streams and items on the inventory are logically related to important activities carried out in these career streams, this would be an example of content validity. Construct validity refers to evidence that supports the conclusion that the measure assesses the construct, factor, or variable that it purports to measure. For example, if the MBTI EI type truly measures extraversion then it should correlate positively with other measures of extraversion. Criterion-related validity refers to evidence that a measure predicts other phenomena happening either concurrently (concurrent validity) or in the future (predictive validity). For example, if a mechanical aptitude measure is shown to predict future job performance in mechanical occupations, this would be considered evidence of criterion-related validity, or more specifically - predictive validity.

Tests that have been around for many years will have more validity evidence than newer tests, simply because there has been more opportunity for research to be conducted on them. When

considering a test for use in career counselling with EE group members, look for evidence of the test's validity, which will most often be found in the technical manual near the reliability data. If there is no mention of validity evidence for a relatively new test, or if validity evidence appears quite limited for a test that has been around for some time, this may be an indicator that you should consider alternative assessment tools. Likewise, check the appropriateness of the samples used in the validity research. Validity evidence collected on a sample representative of the general population may not necessarily generalize to EE groups.

It should be noted as well, that reliability and validity criteria apply to all measures, not just tests. For example, an interview must also be reliable and valid. If evidence is collected that different interviewers obtain similar data when interviewing the same clients this supports the reliability of the interview process used. Likewise, if results from an interview process correlate with future job success, this supports the predictive validity of the interview process. Without such data, the interview is no less prone to error than other measurement tools.

Perception

In a career counselling context, prior to administering a assessment tool, it is a good idea to find out clients' perceptions of the assessment tool. If their perceptions are that the tool is fair and accurate, they are more likely to take the time to carefully complete it and to give careful thought to their responses. On the other hand, if they feel that a test is "biased" and will not assess them fairly, they may not complete the instrument with the care and attention required. Effectively, their perception of the validity of the test may be a self-fulfilling prophecy. In these instances, it may be advantageous to consider alternative methods of obtaining the same or similar information. For example, if a client resisted completing an aptitude test, you may be able to ascertain their strengths and weaknesses through questioning them about past work and life experiences. Similarly, if they felt that an interest inventory was not a valid assessment, a measure of personality or of values may get at similar information.

Related to this, Schmit and Ryan (1997) found that withdrawal from a testing situation for personnel selection was greater for African Americans than Caucasians. However, in investigating the reasons for withdrawal, neither test anxiety nor perceptions of fairness were major factors. Major reasons for withdrawal had to do with conflicts with Mother's Day commitments and the time of day during which the testing was scheduled. Consequently, to the degree that this study is generalizable to a career counselling context, the use of standardized tests will not necessarily cause clients to withdraw from the process.

Language of Testing

In general, if at all possible, the test and the test administration should be conducted in the language of the client's choice. This goes beyond offering testing in one or the other of Canada's official languages. Unfortunately, alternative language versions will often not be available. However, for most of the tools described above, there are several other language versions beyond English and French. Additionally, for some assessment needs, non-verbal test formats are

available. For example, the MAB-II, previously described, includes a Verbal and a Performance (non-verbal) component, one of which or both can be administered to the client. There also are non-verbal personality tests. However, validity data on such well known non-verbal tests as the Rorschach, or the Thematic Apperception Test, are not as strong as for verbal personality measures.

If an alternative language version is not available, the only option available may be to have an interpreter administer the test orally to the client. Such an interpreter must be fully trained in both the language of the test and the client's language. The interpreter must have expertise in language interpretation, and have a good understanding of the assessment process (Standards for Educational and Psychological Testing, 1999). Of course, the use of an interpreter does not eliminate the potential problem of ensuring test equivalence across cultures. Also, when an interpreter is used the test-taker is not provided the same level of privacy and independence that other test-takers would receive.

If an alternative language version is available, it is good practice to check the test documentation (e.g., the technical manual) to determine how the alternative version was developed. If the test was simply translated from the original version, even by a subject-matter-expert, the validity of the instrument may be compromised in the process. A better, but still not ideal, process, is back translation, where a test is translated into the target language, and then re-translated back into the original language, with a comparison made to the original test. (A variation of this is reported by Kaufert and Shapiro (1996) in adapting the Mental Status Questionnaire for use with Manitoba Aboriginal elders. They used back translation repeatedly, supplemented by a panel of Aboriginal consultants.) The ideal process, however, is to develop language versions in the target language, parallel to the original language. That is, items are developed in both languages concurrently (Standards for Educational and Psychological Testing, 1999). The same test development plan is used for the different language versions, but the items are not simply translations of the original items. Finally, look for empirical evidence that the different language versions are indeed parallel, that is, measuring the same construct at the same level. Technical manuals will not always indicate how an alternative language version was developed. However, if this information is available, it will help in deciding whether to use the test, or in choosing amongst tests.

Even if a client has "adequate" English or French, if their first language is neither English nor French, testing (again, both the test administration and the test document itself) in their first language is preferable. Li (1999) reports evidence that, even when both parties are using the same language at the same level of proficiency, *inter*-cultural communication conveys significantly less information (two-thirds) that of *intra*-cultural communication.

The above points relating to language apply equally to all client interactions, not just assessment. To the degree that the entire career counselling process is dependent upon effective communication, all interactions with the client will benefit from the use of the client's first language.

Accommodation for Persons with Disabilities

The purpose of accommodation is, "to minimize the impact of test-taker attributes that are not

relevant to the construct that is the primary focus of the assessment" (Standards for Educational and Psychological Testing, 1999). Consequently, when it is determined that the standard test administration format will not accurately assess the construct in question, accommodation options should be considered.

Most standardized tests lend themselves to at least some kinds of accommodation. A good first strategy for determining the kind of accommodation appropriate for a client is to ask them whether they received accommodation of any kind in the past, what the nature of the accommodation was, and how well it worked for them (i.e., Do they feel it allowed them to be accurately assessed? Were they comfortable with the level of independence the accommodation afforded them?). The client may not have experienced any form of accommodation previously. However, if they did, it may be a simple matter of implementing the same strategy again. For example, someone with a visual impairment may have written other tests previously through computer-assisted administration (e.g., a program that verbalizes test items). It is important to understand that there is not one specific kind of accommodation corresponding to each form of disability. It will depend upon the tools and procedures with which the client is most familiar and comfortable, and the construct that is being measured.

Most commercial tests are *not* readily available in large print, Braille, or auditory tape versions due to limited demand. However, test publishers may be aware of past clients who have translated the test into alternative formats, and who may be willing to share the version with others. It is important to coordinate such efforts through the test publisher in these instances to avoid copyright violations. (Note also, that issues with respect to accuracy of translations, discussed in the previous section, also apply to tests in Braille and sign language.) Also, a client may be able to read the standard version of a test, but due to a disability, may not be able to record his/her responses on a standard answer sheet. Consequently, an accommodation in responding format may be necessary (e.g., a computer-assisted response system).

There are many different accommodation options to consider. Some include: adjusting time to complete the test (for non-speeded tests only); providing breaks during the test (should client fatigue be an issue); providing individual rather than group administration; allowing the test to be completed at home (not appropriate where the security of the test material is an issue, such as with aptitude and personality measures); or providing sample test booklets in advance and/or pretest counselling to reduce test anxiety. Key principles in choosing amongst accommodation options are: that the approach assesses the construct of interest; that it neither provides an advantage or disadvantage to the client; and that it maintains the client's dignity through independence and privacy in the test administration and results communication process. Technological advances have provided accommodation options not previously available, and will continue to generate new options in the future. For more information on ways in which technology can assist in accommodation options for testing, contact the Enabling Resource Centre, Public Service Commission, Government of Canada.

Normative, Criterion-Referenced or Ipsative Interpretation

In using standardized assessment tools for career counselling, there is less need to compare results

to normed samples (normative interpretation) or to an absolute standard (criterion-referenced) as there is to compare subscale scores within a client's own results (ipsative interpretation). When hiring to a job, normative interpretation is often used, as the applicant is typically in a competitive situation with other applicants for the same job, and comparing results to other job applicants or to normed samples is often necessary to identify the most qualified candidate. Alternatively, criterion-referenced score interpretation (Owings and Siefker, 1991) is used to determine whether a candidate meets minimal competency levels to perform the job in question. However, in career counselling, the focus is more on determining the best course of action for the individual client, and comparisons to others or to external standards may not assist in this process.

Consequently, whether assessing values, interests, personality, or aptitudes, in career counselling contexts the first step in interpreting the results is to consider the candidate's pattern of subtest scores. For example, with respect to values, on which of the measured values does the client score highest? On which does he/she score lowest? How does this compare to other information known about the client's values (i.e., from other measures or from discussions with the client)?

Lewis (1998) argues for the identification of contextual factors in interpreting test results. These include "language facility, familiarity with test procedures, approach to completing the test (i.e., speed vs. power), and noticeable level of anxiety" (p.235). Additionally, Lewis advocates noting any departures from normal standardized testing procedures as well as how more dynamic procedures (those which may change from one testing situation to the next) were handled. Lewis further recommends consulting with another professional from the same culture as the candidate being assessed, or someone who has extensive knowledge of the culture, as a final quality check. By considering such contextual factors as these, the interpretation will not only identify how well the candidate performed on the test, but may provide information on why the results were as they were, which in turn may suggest follow-up interventions, such as remediation, counselling for test anxiety, etc.).

If a candidate is being assessed for their likeliness of successfully completing a training course, using normative or criterion-referenced interpretation may be more appropriate for aptitude measures, in that you wish to determine whether the client has sufficient aptitude to successfully complete the training course. However, if the issue is not whether the candidate can succeed in a training course, but rather choosing amongst differing training programs, then the objective becomes one of determining the client's stronger aptitudes, regardless of how they compare to other people's results or to required minimum standards. Consequently, ipsative interpretation of test results is almost always a good starting point for career counselling situations.

It should be noted that for some kinds of tests, the distinction between normative and ipsative interpretation is blurred. For example, personality test subscores are typically provided as "T-scores", which reflect the candidate's results as a deviation from the mean average score for a normative sample. Consequently, T-scores are normative in that sense. However, they can still be interpreted in an ipsative manner by considering the candidate's highest T-scores relative to his/her lowest T-scores.

General or Subgroup Norms

If it is determined that comparison of scores to normative data is appropriate for a particular client, then the question arises whether to use standard norms (i.e., those that represent the general population) or norms specific to the subgroup(s) of which the client is a member. For some tests, norms are provided separately for males and females. Periodically, especially for aptitude measures, they are also provided for age groups. However, in reviewing the six tests above, the author came across no normative data for any of the other three designated groups (i.e., visible minorities, Aboriginal persons, and persons with disabilities). Consequently, although some authors have advocated using subgroup norms (e.g., Martin, date unknown, with respect to the assessment of occupational interest amongst American Indians), in most testing situations, subgroup norms are not available.

If it is determined that general norms are to be used, it is a good idea to check the test documentation to determine the composition of the general normative sample. Most newer tests use appropriate sampling strategies (i.e., which reflect the actual demographics of society). However, older tests may have been normed against a predominant white male sample, which may not provide an adequate comparison.

In deciding whether to use general versus subgroup norms, just as in deciding whether to use normative versus ipsative interpretation, the first step is to determine the purpose of the assessment. If the purpose is to estimate how well the candidate may perform a particular job relative to the general population, then general norms may be appropriate. If, however, the objective is to determine how well suited the candidate is to a particular occupation relative to others from the same subgroup, then subgroup norms may be appropriate. Parker & Schaller, (date unknown) suggest that, if available, subgroup norms should be used. If subgroup norms are not available, and an agency anticipates using an assessment tool frequently, it may be worthwhile developing local (e.g., Federal Government) subgroup norms.

Alternatives to Standardized Assessment

At present, there are few professionally developed standardized assessment tools that have sufficient validity evidence to be confidently used with members of designated groups. Likewise, if normative interpretation of results is desirable, subgroup norms are unlikely to be available. Are there alternatives to standardized assessment tools?

Parker and Schaller suggest four alternatives: self-ratings; criterion assessment; ecological assessment; and qualitative assessment. The concept of self-ratings is to simply ask the client what their interest is in various occupations, or to ask them what their aptitudes are. Parker and Schaller cite research showing that self-ratings tend to be positively correlated with standardized measures of the same constructs, and therefore argue that self-ratings might be a less expensive way to obtain the same information.

Criterion assessment refers to offering the client a job-tryout. If a major objective of career assessment is to predict job success, then the best predictor is to have the person actually try out the job, and measure the criterion of job success directly. The third alternative that Parker and

Schaller offer, ecological assessment, relates more to assessing persons with severe disabilities. They describe it as assessing the individual and the environment, and then looking for a possible fit between the client's interests and abilities and the job market. Finally, qualitative assessment is less clearly described by Parker and Schaller. One of the tenets behind it, however, is the notion that quantitative approaches (i.e., standardized assessment) do not take into account the subtle variables. A qualitative approach would look at the unique combination of factors that impact each instance of career counselling.

None of the four options that Parker and Schaller suggest are without their limitations. Self-ratings are less expensive than standardized assessment tools. However, the use of self-assessment assumes that the candidate already understands their interests, personality, and abilities. Job-tryouts are much more time consuming. If the employer is prepared to offer job-tryouts, directly measuring performance rather than predicting it from a paper-and-pencil instrument has merit. The last two options that Parker and Schaller suggest are less well operationalized and consequently, it is difficult to anticipate their potential utility.

Samuda (1998) summarizes other efforts to produce assessment processes appropriate to diverse populations. He describes efforts dating from the 1950's and 1960's to produce "culture fair" ability tests, including the Cattell Culture-Free Intelligence Test, the Davis-Ells Games Test, Raven's Progressive Matrices, Goodenough Draw-a-Man Test, and Rulon's Semantic Test of Intelligence. The approach used in many of these tests was to minimize the verbal component of the assessment. However, as Samuda concludes, these efforts did little to reduce mean group differences in test performance between minority and majority students.

Samuda also describes several projects to produce "culture specific" tests, that is, rather than attempting to eliminate cultural factors from the test, develop the test to assess specifically within a designated culture. Two examples of attempts to assess specifically within the Black culture include the Dove Counterbalance General Intelligence Test and the Black Intelligence Test of Cultural Homogeneity. Some of the drawbacks to this approach include the cost to develop tests specific to each culture, the challenge of determining who belongs to a certain culture, and the applicability of such testing results beyond the cultural community within which the assessment occurs.

Finally, Samuda outlines two other testing alternatives: efforts to measure environmental factors that could affect intellectual development; and criterion-referenced testing. The premise behind measuring environmental factors is that if a good measure of such factors were available, it could be used to supplement traditional ability tests to generate a more precise prediction of academic success. Criterion-referenced testing, on the other hand, refers to efforts to develop tests that assess candidates against a defined standard of performance rather than against the mean average performance of all other test-takers. Advantages of criterion-referenced testing include the focus on how to close the gap between current performance and the desired standard, as well as the independence of the performance standard from how well others (e.g., the majority population) perform on the test. A key challenge to criterion-referenced testing is to accurately define the standard or criterion to be achieved.

Recommendations

The primary objective of this paper was to better understand the strengths and weaknesses of standardized measurement tools for career assessment of members of diversity groups. However, in the course of reviewing the available literature, several issues were identified that warranted recommendations. Consequently, three recommendations are offered below.

1. It is recommended that research into the reliability and validity of career assessment tests for Employment Equity designated groups (i.e., women, visible minorities, persons with disabilities, and Aboriginal peoples) be encouraged.

As an example of how sparse the literature is, Carter and Swanson (1990) reviewed the validity research literature for the Strong Interest Inventory (SII) with Blacks. Even though the SII is one of the most popular interest inventories and Black Americans are one of the most researched cultural groups, the reviewers found only eight studies in the published literature, most of which focused on earlier versions of the SII.

Other reviewers of career assessment tools for diverse populations (e.g., Eby, Johnson, and Russell, 1998) have likewise called for research into the applicability of standardized assessment tools for diverse populations in a career counselling context.

2. Given that test norms for designated groups are unlikely to be made available by test publishers in the short to medium term future, it is recommended that the Federal Public Service take steps to develop “local” test norms for tests used more frequently.

Some authors have previously called for local normative databases (e.g., Carter & Swanson, 1990, with respect to the use of the SII with Black Americans). From the point of view of the career counsellor, it would be desirable to have the option to use subgroup norms, even if in some instances, it was determined that general norms were satisfactory. Given that test publishers are unlikely to generate such subgroup norms, at least in the short to medium term future, and given that the Federal Public Service administers many tests due to its large size, it may be more feasible for the Public Service to take the initiative to develop such subgroup norms than to wait for test publishers to make such norms available. This would have to be done on a selective basis. However, if four or five tests (such as those reviewed in this paper) were chosen, within a year or two there would be sufficient data to allow other counsellors within the Public Service to use the data to assist in interpreting results for their clients.

3. It is recommended that the Federal Public Service look for collaborative opportunities with commercial test publishers to foster research and test development work into career assessment test usage for diversity groups.

In the course of preparing this report, the author happened upon a test publisher that was in the

process of developing a new test. In informal discussions with this test publisher, the author learned that this test publisher normally would not place test development efforts for diversity groups high on the priority list, simply due to the economics of the matter. However, this same publisher expressed interest in the possibility of a collaborative test development relationship with the Federal Government (which does not necessarily translate into Federal funding of private sector test development work). For example, an arrangement might be made where career counsellors in the Public Service administer a research version of a test at the same time that they administer their regular battery of career assessment tools, thereby facilitating the collection of normative data on a new test, or normative data on an existing test for a subgroup of interest.

Conclusion

The amount and quality of research located that was specific to the use of these tests with diversity groups was very limited. Research indicates that for general population use, the six tests reviewed for this report are reliable and valid assessment tools for career assessment. However, this report cannot conclude at this time that these tests are valid career assessment indicators for members of diversity groups, due to the lack of research focusing on the use of these tools with these groups. The Values Scale represents an impressive international effort to better understand work values in Canada and around the world. It is scheduled to be replaced shortly. This may provide the opportunity for much needed research into the applicability of such a tool to diversity groups. The Jackson Vocational Interest Survey, although not as popular as the Strong Interest Inventory or the Self-Directed Search, is psychometrically every bit the equal of these tools. With its Canadian roots, and thorough computer results, the JVIS has potential to become an effective interest measure for diversity groups. Required is the development of a research base demonstrating its reliability and validity with such groups.

In terms of personality inventories, the MBTI has a rich past and an impressive normative base upon which to build. Once again, however, research looking directly at the validity of the instrument when used with Canadian EE groups is lacking. The NEO PI-R is built on the five-factor model, arguably the personality theory of the future. Given the promising data collected thus far relating to the intercultural applicability of the inventory, it would be timely for specific research to be conducted into the NEO PI-R's applicability to Canadian EE groups in a career assessment context.

Finally, the aptitude measures considered in this report represent two examples of professionally developed and validated tools (for the general population) to round out the career assessment picture. The TAI has sound psychometrics, and offers the convenience of computer assisted and adaptive administration. The MAB-II continues the tradition of psychometrically strong assessment tools produced by Dr. Jackson and his colleagues. Confidence in the use of both of these tools would be strengthened with a research program looking specifically at the validity of these instruments for career assessment with Canadian EE groups.

In the use of any standardized assessment tool with members of EE groups, the counsellor and the

client will be well served if some key principles are applied. Consider a comprehensive approach to assessment, which includes not only standardized assessment measures, but also other information that will help to put the test results in context. Always satisfy yourself that the assessment instruments chosen have adequate reliability and validity evidence based on appropriate samples to support their intended use. If using a verbal aptitude measure, ensure that the test is administered in the client's first language. Where appropriate, consider accommodation options that provide for a more accurate assessment of the construct of interest. Consider the options for interpretation of test results, including ipsative, criterion-referenced, and general population/subgroup normative interpretation. Finally, because the career assessment process is primarily for the client's benefit, look for ways to increase the client's involvement in the assessment process, in particular during the interpretation of the assessment results.

References

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (1999) *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Barrick, M.R. and M.K. Mount (1991) The Big Five Personality Dimensions and Job Performance: A Meta-Analysis, *Personnel Psychology, 44*(1), 1-26.

Barrick, M.R. and M.K. Mount (1993) Autonomy as a moderator of the relationships between the big five personality dimensions and job performance. *Journal of Applied Psychology, 78* (1), 111-118.

Berk, L.E. and Fekken, G.C. (1990) Person reliability evaluated in the context of vocational interest assessment. *Journal of Vocational Behavior, 37*, 7-16.

Bobko, P., Roth, P.L., and Potosky, D. (1999) Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561-590.

Borgen, F.H., and Harper, G.T. (1973) Predictive validity of measured vocational interests with Black and White college men. *Measurement and Evaluation in Guidance, 6* (1), 19-27.

Botwin, M.D. (1995) Review of the Revised NEO Personality Inventory. In *The Twelfth Mental Measurements Yearbook.* Conoley, J.C. and Impala, J.C. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Bowman, S.L. (1993) Career intervention strategies for ethnic minorities. *The Career Development Quarterly, 42* (1), 14-25.

Brown, D.T. (1989) Review of the Jackson Vocational Interest Survey. In *The Tenth Mental Measurements Yearbook.* Conoley, J.C. and Kramer, J.J. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Bujold, C. (1980) Signification du travail et valeurs de travail: revue de la littérature canadienne de langue française. *L'Orientation professionnelle, 16* (1), 5-47.

Carless, S.A. (1999) Career assessment: Holland's vocational interests, personality characteristics, and abilities. *Journal of Career Assessment, 7* (2), 125-144.

Carter, R.T. and Swanson, J.L. (1990) The validity of the Strong Interest Inventory with black Americans: a review of the literature. *Journal of Vocational Behavior, 36,* 195-209.

Casserly, M.C., and Cote, L. (1980) *The Work Importance Study in the Canadian Context.* Ottawa: Canada Employment and Immigration Commission.

Casserly, C., Fitzsimmons, G., and Macnab, D. (1995) Chapter in, *Life Roles, Values, and Careers.* D.E. Super and B. Šverko, Ed., San Francisco: Jossey-Bass, 117-127.

Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82*, 311-320.

Child, D. (1998) Some technical problems in the use of personality measures in occupational settings illustrated using the 'Big Five'. In *Directions in Educational Psychology.* Shorrocks-Taylor, S. (Ed.), London: Whurr Pub., 346-364.

Coe, C.K. (1992) "The MBTI: Potential Uses and Misuses in Personnel Administration", *Public Personnel Management, 21*(4), 511-522.

Collins, J.M. and Gleaves, D.H. (1998) Race, job applicants, and the five-factor model of personality: implications for Black psychology, industrial/organizational psychology, and the five-factor theory. *Journal of Applied Psychology, 83* (4), 531-544.

Costa, P.T., Jr., McCrae, R.R., and Holland, J.L. (1984) Personality and vocational interests in an adult sample. *Journal of Applied Psychology, 69* (3), 390-400.

Costa, P.T. and McCrae, R.R. (1992) *Neo PI-R Professional Manual*. Odessa, Florida: Psychological Assessment Resources, Inc.

Costa, P.T., Jr., McCrae, R.R., and Kay, G.G. (1995) Persons, places, and personality: career assessment using the Revised NEO Personality Inventory. *Journal of Career Assessment, 3* (2), 123-139.

Cronshaw, S.F. (1999) *Draft Final Report of External Review of Selected Instruments of the Personnel Psychology Centre.* Ottawa: Technical report prepared for the Public Service Commission.

Day, S.X. and Rounds, J. (1998) Universality of vocational interest structure among racial and ethnic minorities. *American Psychologist, 53* (7), 728-736.

De Fruyt, F. and Mervielde, I. (1996) Personality and interests as predictors of educational streaming and achievement. *European Journal of Personality, 10*, 405-425.

De Fruyt, F. and Mervielde, I. (1999) RIASEC types and big five traits as predictors of employment status and nature of employment. *Personnel Psychology, 52,* (3), 701-727.

Druckman, D. and Bjork, R.A. (1991) *In the Minds's Eye: Enhancing Human Performance.* Washington, D.C.: National Academy Press.

Eby, L.T., Johnson, C.D., and Russell, J.E.A. (1998) A psychometric review of career assessment tools for use with diverse individuals. *Journal of Career Assessment, 6* (3), 269-310.

Fouad, N.A. and Spreda, S.L. (1995) Use of interest inventories with special populations: women and minority groups. *Journal of Career Assessment, 3* (4), 453-468.

Fouad, N.A., Harmon, L.W., and Borgen, F.H. (1997) Structure of interests in employed male and female members of U.S. racial-ethnic minority and nonminority groups. *Journal of Counseling Psychology, 44* (4), 339-345.

Frisby, C.L. (1999) Straight talk about cognitive assessment and diversity. *School Psychology Quarterly. 14* (3), 195-207.

Furnham, A. (1996) The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Personality and Individual Differences, 21* (2), 303-307.

Getreu, M.R. (1997) Structural interrelationships between vocational interests and personality traits in those who are male sex-typed, female sex-typed, and androgynous. *Dissertation Abstracts International, 58* (3), 1577-B.

Ghorpade, J., Hattrup, K., and Lackritz, J.R. (1999) The use of personality measures in cross-cultural research: a test of three personality scales across two countries. *Journal of Applied Psychology, 84*(5)

Gottfredson, G.D., Jones, E.M., and Holland, J.L. (1993) Personality and vocational interests: the relationship of Holland's six interest dimensions to five robust dimensions of personality. *Journal of Counseling Psychology, 40* (4), 518-524.

Green, K.E. (1998) Review of The Values Scale, Second Edition. In *The Thirteenth Mental Measurements Yearbook.* Impala, J.C. and Plake, B.S. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Grimm, S.D. and Church, A.T. (1999) A cross-cultural study of response biases in personality measures. *Journal of Research in Personality, 33,* 415-441.

Haverkamp, B.E., Collins, R.C., and Hansen, J.C. (1994) Structure of interests of Asian-American college students. *Journal of Counseling Psychology, 41* (2), 256-264.

Healy, C.C. and Woodward, G.A. (1998) The Myers-Briggs Type Indicator and Career Obstacles. *Measurement and Evaluation in Counseling and Development, 31,* 74-85.

Hess, A.K. (1992) Review of the NEO Personality Inventory. In *The Eleventh Mental Measurements Yearbook.* Kramer, J.J. and Conoley, J.C. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Hogan, R., Hogan, J. and Roberts, B.W. (1996) Personality measurement and employment decisions. *American Psychologist, 51,* 469-477.

Holland, J.L., Johnston, J.A., and Asama, N.F. (1994) More evidence for the relationship between Holland's personality types and personality variables. *Journal of Career Assessment, 2* (4), 331-340.

Jackson, D.N. (1998) *Multidimensional Aptitude Battery II: Manual*. London, Ontario: Sigma Assessment Systems.

Jackson, D.N. (2000) *JVIS Manual (Second Edition)*. London, Ontario: Research Psychologists Press.

Jackson, D.N., Guthrie, G.M., Astilla, E., and Elwood, B. (1983) The cross-cultural generalizability of personality construct measures. In *Human Assessment and Cultural Factors,* Irvine, S.H. and Berry, J.W. (Eds.), NATO Conference Series: Series III: Human Factors. New York: Plenum Press, 365-376.

Jepsen, D.A. (1992) In *Test Critiques*, Volume IX. Keyser, D. and Sweetland, R. (Eds.), Austin, TX: PRO-ED, 308-318.

Judge, T.A., Higgins, C.A., Thoresen, C.J. and Barrick, M.R. (1999) The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology, 52* (3), 621-652.

Juni. S. and Koenig, E.J. (1982) Contingency validity as a requirement in forced-choice item construction: a critique of the Jackson Vocational Interest Survey. *Measurement and Evaluation in Guidance, 14* (4), 202-207.

Juni, S. (1995) Review of the Revised NEO Personality Inventory. In *The Twelfth Mental Measurements Yearbook.* Conoley, J.C. and Impala, J.C. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Katz, L., Joyner, J.W., and Seaman, N. (1999) Effects of joint interpretation of the Strong Interest Inventory and the Myers-Briggs Type Indicator in career choice. *Journal of Career Assessment, 7*, (3), 281-297.

Kaufert, J.M. and Shapiro, E. (1996) Cultural, linguistic and contextual factors in validating the Mental Status Questionnaire: the experience of Aboriginal elders in Manitoba. *Transcultural Psychiatric Research Review, 33* (3), 277-296.

Kim, B.S.K., Atkinson, D.R., and Yang, P.H. (1999) The Asian Values Scale: Development, Factor Analysis, Validation, and Reliability. *Journal of Counseling Psychology, 46* (3), 342-352.

Klein, M.L., Wheaton, J.E., & Wilson, K.B. (1997) The career assessment of persons with disabilities: a review. *Journal of Career Assessment, 5*, 203-211.

Kranzler, J.H. (1991) The construct validity of the Multidimensional Aptitude Battery: a word of

caution. *Journal of Clinical Psychology, 47* (5), 691-697.

Krieshok, T.S. and Harrington, R.G. (1985) A review of the Multidimensional Aptitude Battery. *Journal of Counseling and Development, 64* (1), 87-89.

Larue, H., Pépin, M. and Loranger, M. (1996) Une étude de validité et de stabilité du Test d'aptitudes informatisé (TAI) pour enfants. *Revue québécoise de psychologie, 17* (2), 3-9.

Lattimore, R.R. and Borgen, F.H. (1999) Validity of the 1994 Strong Interest Inventory with racial and ethnic groups in the United States. *Journal of Counseling Psychology, 46* (2), 185-195.

Lee, M.S., Wallbrown, F.H., and Blaha, J. (1990) Note on the construct validity of the Multidimensional Aptitude Battery. *Psychological Reports, 67,* 1219-1222.

Lewis, J.E. (1998) Nontraditional uses of traditional aptitude tests. In *Advances in Cross-Cultural Assessment,* R.J. Samuda, R. Feuerstein, A.S. Kaufman, J.E. Lewis, and R.J. Sternberg (contributing authors), Thousand Oaks: Sage.

Li, H.Z. (1999) Communicating information in conversations: a cross-cultural comparison. *Great Britain: Pergamon, Elsevier Science Ltd. (Journal? Book?), Vol.?,* 387-?.

MacDonald, D.A., Anderson, P.E., Tsagarakis, C.I., and Holland, C.J. (1994) Examination of the relationship between the Myers-Briggs Type Indicator and the NEO Personality Inventory. *Psychological Reports, 74,* 339-344.

Macnab, D., Fitzsimmons, G., and Casserly, C. (undated) *Administration Manual for the Life Roles Inventory Values Scale and the Salience Inventory*. Edmonton: PsiCan Consulting Ltd.

Macnab, D., Fitzsimmons, G., and Casserly, C. (1985) *Technical Manual for the Life Roles Inventory Values Scale and the Salience Inventory*. Edmonton: PsiCan Consulting Ltd.

Martin, Jr., W.E. (Date unknown) Career development and American Indians living on reservations: cross-cultural factors to consider. *The Career Development Quarterly, Vol?,* ?-273-283.

McCrae, R. R. and Costa, P. T., Jr. (1989) Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. *Journal of Personality, 57* (1), 17-40.

McCrae, R. R., Costa, P. T., Jr., & Yik, M. S. M. (1996). Universal aspects of Chinese personality structure. In M. H. B (Ed.), *Handbook of Chinese Psychology.* Hong Kong: Oxford University Press.

McShane, D. and Berry, J.W. (1988) Native North Americans: Indian and Inuit abilities. In *Human Abilities in Cultural Context.* Irvine, S.H. and Berry, J.W. (Eds.). Cambridge: Cambridge University Press.

Myers, I.B. and McCaulley, M.H. (1985) *Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Nuby, J.F. and Oxford, R.L. (1998) Learning style preferences of Native American and African American secondary students. *Journal of Psychological Type, 44,* 5-19.

Owings, S. and Siefker, J.M. (1991) Criterion-referenced scoring vs. norming: a critical discussion. *Vocational Evaluation and Work Adjustment Bulletin,* Fall 109-111.

Parker, R.M. and Schaller, J.L. (date unknown) Issues in vocational assessment and disability. In *Work and Disability: Issues and Strategies in Career Development and Job Placement.* Szymanski, E.M. and Parker, R.M. et al. (Eds.), Austin, TX: Pro-Ed., 127-164.

Pépin, M. et Loranger, M. (1996) *Le TAI - Adolescents et adultes: Guide d'utilisation.* Saint-Foy (Québec): Le Réseau Psychotech inc.

Portugais, C., Daudelin, G., Loranger, M. and Pépin, M. (1995) La stabilité du test d'aptutude informatisé (TAI). *L'orientation, 8* (3), 6-8.

Posey, A.M., Thorne, B.M., and Carskadon, T.G. (1999) Differential validity and comparative type distributions of blacks and whites on the Myers-Briggs Type Indicator. *Journal of Psychological Type, 48,* 6-21.

Ridley, C.R., Li, L.C., and Hill, C.L. (1998) Multicultural assessment: reexamination, reconceptualization, and practical application. *The Counseling Psychologist, 26* (6), 827-910.

Rounds, J. and Tracey, T.J. (1996) Cross-cultural structural equivalence of RIASEC models and measures. *Journal of Counseling Psychology, 43* (3), 310-329.

Rousseau, D.M. (1989) Review of the Values Scale, Research Edition. In *The Tenth Mental Measurements Yearbook.* Conoley, J.C. and Kramer, J.J. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Ryan, J.M., Tracey, T.J.G., and Rounds, J. (1996) Generalizability of Holland's structure of vocational interests across ethnicity, gender, and socioeconomic status. *Journal of Counseling Psychology, 43* (3), 330-337.

Samuda, R.J. (1983) Cross-cultural testing within a multicultural society. In *Human Assessment and Cultural Factors,* Irvine, S.H. and Berry, J.W. (Eds.), NATO Conference Series: Series III: Human Factors. New York: Plenum Press. 591-606.

Samuda, R.J. (1998) *Psychological Testing of American Minorities: Issues and Consequences (Second Edition).* Thousand Oaks: Sage.

Schmidt, F.L. and Hunter, J.E. (1996) Measurable personnel characteristics: stability, variability, and validity for predicting future job performance and job related learning. Chapter to appear

in Kleinmann, Martin, and Stauss, Bernd (Eds.), *Instruments for Potential Assessment and Personnel Development.* Gottingen, Germany: Hogrefe.

Schmit, M.J. and Ryan, A.M. (1997) Applicant withdrawal: the role of test-taking attitudes and racial differences. *Personnel Psychology, 50* (4), 855-876.

Schoenrade, P. (1998) Review of The Values Scale, Second Edition. In *The Thirteenth Mental Measurements Yearbook.* Impala, J.C. and Plake, B.S. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Shepherd, J.W. (1989) Review of the Jackson Vocational Interest Survey. In *The Tenth Mental Measurements Yearbook.* Conoley, J.C. and Kramer, J.J. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Simmons, G. and Barrineau, P. (1994) Learning style and the Native American. *Journal of Psychological Type. 28,* 3-10.

Slaney, R.B. (1989) Review of the Values Scale, Research Edition. In *The Tenth Mental Measurements Yearbook.* Conoley, J.C. and Kramer, J.J. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Soldz, S. and Vaillant, G.E. (1999) The big five personality traits and the life course: a 45-year longitudinal study. *Journal of Research in Personality, 33*, 208-232.

Super, D.E. (1995) Values: Their nature, assessment, and practical use. Chapter in, *Life Roles, Values, and Careers.* D.E. Super and B. Šverko, Ed., San Francisco: Jossey-Bass, 54-61.

Suzuki, L.A. and Kugler, J.F. (1995) Intelligence and personality assessment. In *Handbook of Multicultural Counseling,* Ponterotto, J.G., Casas, J.M. Suzuki, L.A., and Alexander, C.M. (Eds.), Thousand Oaks: Sage, 493-515.

Tett, R.P., D.N. Jackson, and M. Rothstein (1991) "Personality Measures as Predictors of Job Performance: a Meta-Analytic Review." *Personnel Psychology, 44.* 703-742.

Tinsley, H.E.A. (1994) Review of the NEO Personality Inventory-Revised. In Keyser, D. and Sweetland, R. (Eds.) *Test Critiques*, Vol. X. Austin, TX: PRO-ED.

Tokar, D.M. and Swanson, J.L. (1995) Evaluation of the correspondence between Holland's vocational personality typology and the five-factor model of personality. *Journal of Vocational Behavior, 46* 89-108.

Wallbrown, F.H., Carmin, C.N., and Barnett, R.W. (1988) Investigating the construct validity of the Multidimensional Aptitude Battery. *Psychological Reports, 62,* 871-878.

Wallbrown, F.H., Carmin, C.N., and Barnett, R.W. (1989) A further note on the construct validity of the Multidimensional Aptitude Battery. *Journal of Clinical Psychology, 45,* 429-

433.

Walsh, B.D., Vacha-Haase, T., Kapes, J.T., Dresden, J.H., Thomson, W.A., and Ochoa-Shargey, B. (1996) The Values Scale: differences across grade levels for ethnic minority students. *Educational and Psychological Measurement, 56* (2), 263-275.

Widiger, T.A. (1992) Review of the NEO Personality Inventory. In *The Eleventh Mental Measurements Yearbook.* Kramer, J.J. and Conoley, J.C. (Eds). Lincoln, NE: Buros Institute of Mental Measurements.

Willis, C.G. (1984) Review of Myers-Briggs Type Indicator. In *Test Critiques*, Volume I. Keyser, D. and Sweetland, R. (Eds.), Austin, TX: PRO-ED, 482-490.

Wiggins, J.S. (1989) Review of the Myers-Briggs Type Indicator. In *The Tenth Mental Measurements Yearbook.* Conoley, J.C. and Kramer, J.J. (Eds.) Lincoln, NE: Buros Institute of Mental Measurements.

Glossary

Computer-administered versus computer-adaptive tests - A computer administered test means simply that the test is taken via computer. A computer-adaptive test is also taken on the computer, but in addition, the test content is adjusted according to the respondent's answers to the earlier items on the test. For example, test questions in each section of the test may be presented in order of difficulty with easier items first. Once respondents answer a predetermined number of items incorrectly, the computer automatically takes them to the next subtest, thereby reducing the administration time for the overall test. (Other test adaptive strategies can apply as well.)

Construct validity - evidence that a measure is assessing the construct or variable that was designed to be measured.

Convergent validity - evidence showing that a measure tends to correlate with other measures that assess similar constructs. Convergent validity evidence is often used to support the construct validity of a measure.

Criterion-referenced interpretation - interpreting test results by determining whether they achieve a level that is predetermined to be the desirable or needed level of the construct. For example, through research it may be determined that in order for a trainee to succeed in a particular course of studies, they must score at a certain level on an aptitude measure. The test results are interpreted in terms of whether or not they meet the "criterion" score previously identified.

Criterion-related validity - evidence showing that a measure correlates with some outcome measure that is taken more or less at the same time (concurrent validity) or in the future (predictive validity). For example, if an interest measure correlates with satisfaction with one's current job, this is criterion-related validity - more specifically, it is concurrent validity evidence for the interest measure. Alternatively, if an interest measure correlates with satisfaction with one's job five years later, this is a different form of criterion-related validity called predictive validity.

Discriminant function analysis - a statistical procedure, related to factor analysis, which attempts to determine whether existing groupings can be explained (i.e., mathematically predicted) by other quantitative data.

Discriminant validity - evidence showing that a measure tends *not* to correlate with other measures that are logically/theoretically unrelated. For example, personality is typically understood as something different from and unrelated to cognitive ability. Consequently, one would expect that a measure of personality not correlate well with a measure of cognitive ability. This would be discriminant validity evidence. In a way, discriminant validity is the opposite of convergent validity.

Factor analysis - a statistical procedure that identifies how well similar variables group together, such as items that comprise a scale on a test.

Individual reliability - an assessment of the stability of assessing interest for an individual respondent (versus stability of the test).

Internal consistency - a measure of reliability which reflects the degree to which the various items in a scale tend to measure the same construct. It is considered to be an upper bound estimate of the measure's true reliability (i.e., the true reliability is unlikely to be any higher). It also is easy to obtain but is perceived to be less useful than test-retest reliability measures.

Ipsative test interpretation - an approach to interpreting test results that draws comparisons only amongst subtest scores for the individual being tested. No comparisons are made to other people's test results.

Meta-analysis - a category of analytical procedures which involve combining results from many different individual studies. The advantage of a meta-analysis over a single study is that the meta-analysis is based upon a much larger sample (the combined samples of all of the individual studies included in the analysis) and therefore is typically interpreted as a more robust assessment of validity.

Norm-based interpretation - interpreting test results on the basis of how they compare to results obtained from other test-takers, either from the general population, or from a subset of the population, defined by age, gender, culture, occupational group, or any other factor perceived to be important in understanding the test results.

Parallel forms reliability - a measure of the reliability or consistency of a measure, obtained by correlating two different but equivalent forms or versions of the same test.

Public service - The "public service", written without capitalization, refers to all federal departments and agencies that receive their funding through Treasury Board appropriations. A subset of this group, the "Public Service", with the "P" and "S" capitalized, refers to those departments and agencies, that are subject to the Public Service Employment Act.

Reliability and validity - the two characteristics, above all others, which must be present in order for a standardized assessment tool to be effective. Reliability refers to the degree to which something can be measured consistently over repeated measures or over time. For example, a tape measure is a reliable measure of a person's height in that it consistently provides the same information. However, reliability is a necessary but insufficient characteristic for a good measure. That is because something can be *consistently* measured *inaccurately*. For example, a typical bathroom weigh scale that is adjusted incorrectly may consistently show a person's weight as 5 pounds too heavy, every time. Consequently, a measure must not only be consistent (reliable), but also valid, meaning that it accurately measures what it purports to measure.

Standardized assessment - measurement instruments, typically paper and pencil tools (although computer administration is becoming more common), which are usually referred to as "tests".

They are “standardized” in that all test-takers read the same instructions, complete the same test items, and the results are scored using the same process.

Standard deviation - a measure of variance within a distribution of scores (i.e., how spread out a group of scores are), calculated as the average deviation of all the scores from the mean average score. A score at the mean average point on a distribution is higher than approximately 50 percent of all other scores. If the score is one standard deviation below the mean, it exceeds only approximately 16 percent of all other scores in the distribution.

Test-retest reliability - a form of reliability where the measure is administered to the same sample twice with some time between the first and second administrations. The two sets of scores are then correlated to determine how similar they are over the elapsed time.