Independent Research Report

The following report is authored by Tracy Tipene (BMS, Dip Soc Sci) who is a part owner of Career Development Systems Pacific. Tracy is currently studying his Honours papers as a prelude to Doctorate study in 2004. This report is a compilation of current research on the CareerScope Assessment & Reporting Tool and has been designed to:

- Provide some technical data on CareerScope.
- Provide additional information on psychometric testing principles.
- Encourage the developers of competing products to make validity data available.
- Assist potential customers in making a purchasing decision.

TABLE OF CONTENTS

GENERAL INFORMATION

Test Name: CareerScope Release 4.0 ANZ

Test Author: Vocational Research Institute (VRI) - JEVS

Publication Date: CareerScope User Guide for Release 4.0 ANZ (2000)

Time:		
	Orientation	10 minutes
	Interest Inventory (no time limit)	20 minutes
	Aptitude Battery (time limits)	25 minutes

NATURE AND PURPOSE

Type: CS is a computer-based tool capable of assessing individuals or groups. It combines interest inventory test results with a seven-test aptitude battery to produce occupational recommendations. A database management module is also included.

Norm Groups: Interest Inventory - Youth 18 years and younger
Interest Inventory - Adults 19 years and older
Interest Inventory – Male and female
Aptitude Battery - Youth 15 years and younger
Aptitude Battery - Youth 16 & 17 year
Aptitude Battery - Adults 18 years and over
Aptitude Battery – Male and female

Populations: High school students and adults.

Test Content: CS measures for twelve career interests including artistic, scientific, plants / animals, protective, mechanical, industrial, business detail, selling, accommodation, humanitarian, leading / influencing and physical performing. The six aptitudes measured include general learning, numerical, verbal, spatial, form perception and clerical perception.

Scores: The interest inventory uses percentile ranks displayed both graphically and as an idiographic individual profile. The aptitude battery uses standardised scores with a mean of 100 and standard deviations of 20. It is reported in histogram format and as numerical percentile ranks.

Item Types: The 145 question interest inventory includes a three-point scale forcing the test taker to indicate a "Dislike, Don't know and Like" opinion of the displayed job tasks. The aptitude battery uses multiple-choice to encourage the test-taker to choose from one of 4 options. Each aptitude exercise has a time limit (speeded).

PRACTICAL EVALUATION

Qualitative Features

A reliability study by Lustig, Brown and Lott (1998a) summarised the features of CS. They noted a number of attractive features including: (i) the assessment can be completed in an hour, (ii) aptitudes and interests are cross-referenced based on existing job areas and (iii) the produced reports are readable and informative for both the practitioner and the test taker. A later review by Clarence Brown (2001) also noted that another attractive feature is that a person with a year six (NZ equivalent of US 4th grade) reading level can complete it. Participants in the Lustig et al (1998a) study also reported that CS is easy to use, has clear & concise instructions and is an appropriate length. The participant ratings would suggest test taker support.

Other features not mentioned in the study pertain to the management module. The management module is designed to make life easier for the practitioner by providing data storage options, data filtering, assessment templates, result output options, and allows the input of some additional assessment results.

Ease of Administration

CS is a very easy to administer to clients. It uses a basic computer set-up including a pentium speed processor, monitor, keyboard, mouse and printer. It does not require special tools and once installed the test practitioner needs only to ensure that paper and two pens are available for each test-taker. During the assessment the test taker uses the keyboard to enter demographic information. The test-taker is easily guided through the rest of the assessment process and uses only the mouse. In general, the practitioners' time is freed up but someone must be present during group assessments to ensure that factors affecting test validity such as cheating and interference[1] do not occur. However, practitioners using laptops (mobile testing) need to ensure they are using TFT active matrix screens. The quality of passive matrix screens can seriously affect the quality of graphics required for form perception and spatial ability tests. Additionally, adequate desktop space is required to allow enough room for the adequate use of a keyboard, mouse, paper and pen.

[1] Examples of interference can rage from tapping on others shoulders etc to noise such as excessive clicking of a pen.

Scoring Procedures

VRI have made the most of computer technology and completely automated the scoring procedures. The full automation leads to a successful standardisation of the scoring procedure and is inline with the Murphy & Davidshofer (2001) advice to standardise as much of the assessment procedure as possible. The most obvious advantage of complete automation is that scoring variance is eliminated (Anastasi & Urbina, 1997). This procedure is rigid but competent practitioners can modify report recommendations.

Murphy & Davidshofer (2001) also point out the problem of test developers who do not release scoring algorithms embedded in their assessment software. Scoring algorithms need to be available for scrutinising by researchers. The CS manual contains explanations of how both the interest inventory and aptitude scores are calculated. The interest inventory has 145 questions that are coded to one of the twelve interest areas. A tally is made of the Likes etc and presented in table format. The “Like percentile score” is compared to American Department of Labor (DOL) norms for each interest area and displayed in table format. This allows all parties to view above average interests.

The most useful scoring algorithm is the Individual Profile Analysis (IPA). The IPA is an idiographic profile, presented in histogram format, which is used to determine career interest “like” scores that are relatively higher than the others. Table one and the following steps demonstrate how an IPA is calculated for an example adult (Vocational Research Institute, 2002, p65).

1. Calculate the mean of all “like” percentages (example mean = 35.9).
2. Calculate the deviation from the mean for each interest area.

3. Divide deviation scores by one of two average intra-individual score profile variability coefficients (Adults = 23, Youth 21.6). The CareerScope Deviation Index (CSDI) column displays those scores that are above the mean.
4. The outcome column indicates scores that are above a 0.5 benchmark.
5. Rank those scores above the 0.5 benchmark.

Table 1: Steps to calculating an IPA.

Interest Area	"Like" (%)	Deviation	CSDI	Outcome (> .5)	IPA Rank
01	21	-14.9	Below Mean	NO	
02	77	41.1	1.79	YES	1
03	36	0.1	.004	NO	
04	17	-18.9	Below Mean	NO	
05	17	-18.9	Below Mean	NO	
06	8	-27.9	Below Mean	NO	
07	67	31.1	1.35	YES	3
08	40	4.1	.18	NO	
09	30	-5.9	Below Mean	NO	
10	27	-8.9	Below Mean	NO	
11	71	35.1	1.53	YES	2
12	20	-15.9	Below Mean	NO	

A closer look at the CSDI equation reveals that they are essentially z-test scores. The only difference from normal z-test routine is that the CSDI uses a normed standard deviation (termed "average intra-profile variability") rather than the test takers standard deviation. VRI posit that the use of a normed standard deviation reduces problems that occur due to profile flatness (when most of the scores are similar).

Test Administrator Qualifications

VRI have not published information to indicate that any specific level of education or expertise is required to be able to administer CS. CS has been developed primarily for use by high school careers advisors and private career practitioners. Thus by default only degree-qualified people are administering CS. The restricting of psychometric tools is embedded in “best practice” ethos that helps to maintain the validity of assessment tools. This restriction is especially suited to methods such as projective testing that require the practitioner to score and interpret data (Anastasi & Urbina, 1997). In support of test restriction Davidson (1997) argued that technical and contextual competence is required to decrease inaccurate results. A technically competent practitioner is well versed in scales, scoring, interpretation, testing environment, reliability, validity and most other psychometric issues. Contextually competence refers to the practitioner’s ability to treat issues that arise from test taker behaviour and background.

The CS user guide (2000) is very informative and comprehensive training guide for practitioners. Some of the jargon used assumes a basic knowledge of psychometrics but in other areas the guide also provides explanations designed to communicate to test administrators who are not psychometrically trained. Untrained practitioners are common in high schools where a subject-qualified teacher concurrently serves as the school careers advisor. CS is fully automated and as such does not require an expert to oversee the assessment.

Perhaps one of the finer features of CareerScope is the use of an expert system to assist practitioners in the interpretation of results and reports. Sampson (2001) suggested that an expert system be a part of computer assessment packages. VRI achieve this via an after sales service that offers practitioners a second opinion on hard to interpret reports. The process requires the practitioner to download the Counselor Report and then forward this to the distributor by print or electronic means. The distributor will examine the report, make recommendations and if necessary get further verification from VRI. The expert opinion is then sent back to the practitioner. This system protects CS validity and accuracy by providing expert feedback that has been provided by qualified personnel.

Therefore while CS is not formally restricted to psychometrically qualified personnel, it can still reap protectionist benefits by providing a package that does not require the practitioner to score and interpret data. A Counselors Report is available if further inspection is required.

Reliability

Reliability refers to the measurement of consistency between a persons test scores (Murphy & Davidshofer, 2001 and Anastasi & Urbina, 1997). It can also measure the stability of a test over time. A reliable test has consistent scoring resulting in accurate comparisons of scores while an unreliable test will result in significant score variations. As mentioned by Murphy and Davidshofer (2001), a valid test is also reliable but a reliable test is not necessarily valid.

Even in the most perfect of conditions no test will ever be 100% reliable (Murphy & Davidshofer, 2001 and Anastasi & Urbina, 1997). Every test is exposed to a variety of factors that affect its ability to deliver consistent scores. Reliability coefficients allow these factors to be identified. The causes of variations in scores are:

- Temporal / Time Factors – Variance caused by differences between the first and subsequent test times. Examples include uncontrollable conditions (sudden noise or a broken pencil) and those conditions attributable to the test taker such as illness, fatigue, stress, and recent experiences. Other factors mentioned by Murphy & Davidshofer include test-taker memory and learning over time.
- Content Factors – Variance due to the selection of items in the test
- Heterogeneity factors - Variance due to test items differing in similarity.
- Scorer Variance – Variance that occurs when more than one observer is involved in scoring. The full automation of tests using computer technology, such as has CS, eliminates scorer variance (Anastasi & Urbina, 1997).

Different methods exist to measure reliability. Reliability studies on the CS tool have used the Test-Retest and Internal Consistency methods. Each method uses correlation coefficients to quantify the degree of consistency between results. A reliability coefficient of 1.0 is the perfect result and a score of 0.8 is considered very good. A score of 0.8 still indicates that the other 0.2 is due to variance. The aim of test developers is to decrease the level of variance in a test and the best way to reduce variance is to understand which variance factors are affecting the test. The most common correlation methods are:

- Test-retest. Found by having the same group of people sit the same test twice with a time delay. The scores of both tests are used to determine the coefficient. This method determines variance due to time factors but also introduces uncontrollables such as test-taker memory and time spent learning. It is generally the least preferred method because if its limited detection ability but is preferred for speed tests such as are found in the CS aptitude battery. It is also time consuming since two test occasions are required.

- Alternate forms. This method requires a group of people to sit two forms of a test. The second test can be done immediately or delayed. This method determines variance due to content factors and also time factors if the testing is delayed. And since the content is different, the time related problem of memory is controlled for. However, temporal problems due to test fatigue can be introduced if the second is done immediately. A high alternate form coefficient measures more variance and is therefore more robuste than a test-retest coefficient. However developing an alternate test form and conducting two testing sessions does make it a more costly option.
- Split-half. This method requires a group of people to sit two halves of a test. The second test can be done immediately since test fatigue should not be a problem with a single test length. This method determines variance due to content factors. Variance due to time factors is controlled for since only one testing session is used. A high split half coefficient is more robuste than a test-retest coefficient. However splitting the test in half must be done with extreme care if results are to be accurate. Preparation time is less than alternate form and the costs reduced because only one testing session is required.
- Internal consistency. One test is administered once to one group of people. This determines variance due to content and heterogeneity factors. Internal consistency is valued because it indicates how homogenous the test items

are. Generally, the more homogeneous the results, the greater the chances for consistent scores. It is also time and cost effective.

Scorer variance can also be measured but for the purposes of this essay is an unnecessary addition since CS's fully automated scoring system results in no scorer variance. Also discussed in Murphy and Davidshofer (2001) is the use of analysis of variance techniques that can quantify specific sources of error.

Reliability research on the CS interest inventory as been conducted by VRI (1999c, d) and Lustig et al (1998a). For whatever reason CS does not include this information in their manuals. The latter group administered CS, in a test-retest fashion, to 46 education students in a southern United States university and used a congruency method to estimate reliability. As previously mentioned the IPA is a rank of those interests that were significantly higher than the others. The two highest IPA ranks for each student were compared in both tests. IPA rank sets that matched regardless of order were considered in full agreement (54%). Sets that partially matched (i.e. one of the IPA ranks was common to both tests) were considered in partial agreement (44%). No match up between the top two IPAs was a no match (2%). Overall there was some form of agreement in 98% of the IPA sets. This result indicates reliability but a more robuste measure could have been gained from using IPA sets of three. In addition range restrictions discourages the generalisation of the results.

The VRI (1999, c&d) studies utilised test-retest and internal consistency methods. The test-retest involved administering CS to a mixture of 307 high school students in a Louisiana, USA parish. The internal consistency research involved 115 employees of the Philadelphia Jewish Employment and Vocational Service (JEVS). A range of people and abilities were represented in the study. Table 2 displays the results from both methods.

These results combined with the Lustig et al (1998a) indicate that the CS interest inventory is a reliable measure for high school students. The results also show high levels of internal consistency demonstrating homogeneous test items. The Brown (1999) review suggests that more robuste investigation is required in order to generalise the results. Had alternate-form coefficients been calculated it would have been possible to portion out the amount of error variance.

Table 2: Internal Consistency (α) and Retest Reliability (r_{tt}) and Retest Summary Statistics

Test	α	r_{tt}	*SE*	*Mean (SD)*
Artistic	.88	.83	1.8	5.2 (4.3)
Scientific	.89	.84	1.6	4.2 (4.1)
Plants/Animals	.86	.79	1.5	3.3 (3.2)
Protective	.84	.81	1.6	3.8 (3.6)
Mechanical	.89	.86	1.4	3.0 (3.7)
Industrial	.90	.73	1.0	1.1 (1.9)
Business Detail	.91	.86	1.7	4.4 (4.6)
Selling	.82	.74	1.2	2.0 (2.3)
Accommodating	.81	.73	1.2	2.2 (2.4)
Humanitarian	.89	.84	1.8	4.5 (4.5)
Leading/Influencing	.84	.82	1.7	5.5 (4.1)
Physical Performing	.91	.87	1.5	4.9 (4.1)

Table 3: Retest Reliability and Summary Statistics

Aptitude Test	$r_{tt\text{ - Lustig}}$	$r_{tt\text{ - VRI}}$	*SE*$_{\text{- VRI}}$	*Mean (SD)* $_{\text{- VRI}}$
General Learning Ability	0.81	.80	3.7	91.8 (8.3)
Verbal Ability	0.83	.79	6.6	85.4 (14.3)
Numerical Ability	0.74	.73	4.3	97.7 (8.3)
Spatial Ability	0.52	.70	7.2	89.6 (13.1)
Form Perception	0.72	.70	5.4	98.9 (9.8)
Clerical Perception	0.75	.72	10.7	112.7 (20.2)

Reliability research on the CS aptitude battery has been conducted VRI (1999b) and Lustig et al (1998a). Both authors used test-retest methods. This method was the most suitable since the CS aptitude battery has a mixture of speed and power tests. The two most speeded tests are Form Perception and Clerical Perception (VRI, 1999a). Table 3 combines the results from both these studies allowing easy comparison of the results.

For five of the six aptitudes both VRI and Lustig et al got similar reliability coefficients. Otherwise all the VRI coefficients are over 70 with general learning above the magic 0.80 benchmark. Lustig reported higher but similar coefficients for five of the six aptitudes. The only exception was spatial ability with a lower coefficient of 0.52 compared to VRI's (1999b) 0.70. The low reliability coefficient could have resulted from the limited sample of 46 students but this argument is moderated by the fact that all the other aptitudes were closely matched. The study suggests that the aptitude tests are moderately - highly reliable.

Validity

Validity attempts to analyse and report on test accuracy and usefulness. Murphy and Davidshofer (2001, p.145) describe validity as: “(1) the validity of measurement and (2) the validity for decisions”. The four methods of measuring validity are:

1. Content Validity –Validity of measurement by ensuring that a test contains a representative sample of items. For example a mechanics test will include a representative sample of items that would normally be included in a typical mechanics job. Work Sample testing is an example of careers assessments that require solid content validity. However, content validity is an inappropriate measurement for aptitude batteries that are concerned mainly with predicting job performance (Anastasi & Urbina, 1997).
2. Construct Validity – Validity of measurement by analysing the extent to which a test measures a construct. A number of methods a used to measure constructs. Factor analysis can be used during test construction to psychometrically develop constructs from closely correlated traits. Internal consistency (also used to measure for reliability) can be used to determine the homogeneity of test items. Convergent and discriminant validation can be conducted to ensure that test constructs correlate with other similar tests and, importantly, do not correlate with tests that it should not. A basic and less costly method is to choose constructs that have already been highly scrutinised and validated (so much for intellectual property). Relying wholly on past data results in reduced costs but also reduces innovation and competitive advantage.

3. External Validity – Ensuring decision-making validity by conducting a proper correlation study. Most external validity studies are costly, time expensive, and considered impractical.
4. Concurrent Validity – Ensuring decision-making validity by correlating test results with either people already in a situation or with assessment tools that have already been validated.

As mentioned previously, content validity research is deemed unnecessary for test instruments that are concerned mainly with predicting job performance (Anastasi & Urbina, 1997). CS fits this criteria, however, this only serves to ensure that the other validity coefficients are strong enough. An internal consistency study conducted by VRI (1999d), and reported previously, does suggest construct validity for the CS interest inventory scales. Internal consistency estimates construct validity by measuring the homogeneity of test items which in turn indicates that all test items are testing one construct (Anastasi & Urbina, 1997).

Concurrent validity research on the CS interest inventory as been conducted by Lustig et al (1998b). The latter used three indices to measure congruency between the CS interests and those found in the Holland Self Directed Search (SDS) (sample was 47 college of education students). The SDS is widely used and has been found to be an impressive instrument (Daniels, 1994, p.210, cited in Lustig et al, 1998b). On all three indices the CS interests exceeded the SDS mean indicating a high level of congruence.

Concurrent validity research on the CS aptitude battery as been conducted by VRI (1999a) and Lustig et al (1998b). Both groups correlated CS aptitude scores with equivalent aptitudes found in the General Aptitude Test Battery (GATB). The GATB is an appropriate choice against which to validate due to the following factors:

- CS is a second-generation development based on the GATB. CS test items are very similar to the GATB excepting each sub test is shorter and CS only measures six aptitudes. CS does not measure psychomotor abilities.
- The GATB has had an enormous number of validation studies conducted on it. It is one of the most widely used aptitude batteries with recognised and accepted validity.
- A study by Howe (1975) indicates that the GATB has been normed successfully in Australia.

Table 4 combines the results from both these studies allowing easy comparison.

Table 4: Concurrent Validity Results – Pearson Product Moment (*r*) and Significance Levels (*p*).

Aptitude Test	$r_{\text{- Lustig}}$	*p*	$r_{\text{- VRI}}$	*p*
General Learning Ability	0.86	.01	0.81	.01
Verbal Ability	0.73	.01	0.74	.01
Numerical Ability	0.67	.01	0.82	.01
Spatial Ability	0.68	.01	0.71	.01
Form Perception	0.36	.05	0.59	.01
Clerical Perception	0.45	.01	0.52	.01

Both studies found that CS and the GATB aptitudes were correlated at least moderately. Both showed that the most important aptitude, general learning, was greater than 0.80 while the VRI study displayed four coefficients over 0.70. For four of the six aptitudes, the correlation coefficients were similar with the two exceptions showing sizable differences (numerical and form). The VRI study showed higher coefficients except in general leaning. The reason for the higher coefficients in the VRI study, particularly for form and clerical perception, could be due to sample size differences. The Lustig et al study had only 47 participants while the VRI study contained 115 people. Additionally, range restrictions likely affected the Lustig et al study because it involved only college education students while the VRI study contained a range of educational achievement.

Additionally, the VRI study also included inter-correlation results for both CS and the GATB. The results displayed similar coefficients and just as importantly, displayed the same discriminant patterns. The studies indicate that CS is a valid tool.

Norms

The CS user guide does not contain norm information except to say that tests are based on VRI norms developed in the USA. A recent study by Rodriguez, Treacy, Sowerby and Murphy (1998) suggests that ability results from adapted tests are much the same as those found in USA norms. This further suggests that if the USA norms are well developed then professionally adapted or customised tests do not need to be re-normed in NZ & Australia.

A review of both the Lustig et al studies and also the Clarence Brown review indicate that CS has been normed on a predecessor product called Apticom. Ingram (1987) in a dissertation abstract indicates that Apticom was successfully validated against the GATB.

SUMMARY

The CS assessment and reporting tool is an exciting state of the art tool. It offers both assessment and database modules to make life easier for the careers practitioner. Its computer-automated approach encourages a high level of reliability as supported by the developer and Lustig et al (1998a). CS validity has also been indicated by research from the same study authors. Although the adapted ANZ version of CS has not been re-normed in either Australia or NZ, there are studies that suggest re-norming is not necessary. CS is developing its technical information with favourable initial results.

In addition, CS's computer assisted assessment of aptitudes is unparalleled in NZ & Australia. This is an important consideration since many South African immigrants are asking that their children have aptitude assessments while at high school (apparently the norm in South Africa). Thus CS is poised in an advantageous position.

REFERENCES

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

Brown, C. (1999). CareerScope assessment & reporting system. In J.T. Kapes, M.M. Mastie, & E.A. Whitfield (Eds.), A counselor's guide to career assessment instruments (4th Ed), (pp. 116-122). Columbus, OH: National Career Development Association.

Davidson, G. (1997). The ethical use of psychological tests: Australia. European journal of psychological assessment, 13 (2), pp.132-139.

Howe, M. (1975). General aptitude test battery – an Australian empirical study. Australian psychologist, 10 (1), March 1975, pp.32-44.

Ingram, G. (1987). A comparative investigation of the Apticom and the general aptitude test battery. Dissertation abstracts international, 48 (8), February 1988, pp.2495b-2496b.

Lustig, D., Brown, D., & Lott. (1998a). Reliability of the CareerScope career assessment and reporting system. Vocational evaluation and work adjustment journal bulletin. Spring 1998, pp. 19-21.

Lustig, D., Brown, D., Lott, A., & Larkin, V. (1998b). Concurrent validity of the CareerScope career assessment and reporting system. Vocational evaluation and work adjustment journal. Summer 1998, pp.28-31.

Murphy, K., & Davidshofer, C. Psychological testing, principles and applications (5th Ed). Englewood Cliffs, NJ; London, UK: Prentice Hall, 2001.

Rodriguez, C., Treacy, L., Sowerby, P., & Murphy, L. (1998). Applicability of Australian adaptations of intelligence tests in New Zealand with a Dunedin sample of children. New Zealand journal of psychology, 27 (1), pp.4-13.

Sampson, J. (2000). Computer applications. In C.E. Watkins & V.L. Campbell (Eds.),Testing and assessment in counseling practice (2nd Ed) (pp.517-544). Mahwah, NJ: Lawrence Erlbaum Associates.

Vocational Research Institute. (1999a). CareerScope research brief #1: concurrent validity of the careerscope aptitude battery. Philadelphia, PA: Vocational Research Institute.

Vocational Research Institute. (1999b). CareerScope research brief #2: retest reliability of the careerscope aptitude battery. Philadelphia, PA: Vocational Research Institute.

Vocational Research Institute. (1999c). CareerScope research brief #3: retest reliability of the careerscope interest inventory scales. Philadelphia, PA: Vocational Research Institute.

Vocational Research Institute. (1999d). CareerScope research brief #4: internal consistency of the careerscope interest inventory scales. Philadelphia, PA: Vocational Research Institute.

Vocational Research Institute (2000). CareerScope user guide for release 4.0 anz. Philadelphia: Vocational Research Institute.