Hand-Scoring Error Rates in Psychological Testing

Roland Simons, Richard Goddard and Wendy Patton

Assessment 2002; 9; 292
DOI: 10.1177/1073191102009003008

Hand-Scoring Error Rates in Psychological Testing

Roland Simons
Richard Goddard
Wendy Patton
Queensland University of Technology

Despite the comprehensive treatment of test validity in most technical manuals, test authors appear to routinely assume that clients and professionals will score their instruments without error. Recently Allard and Faust challenged this assumption by suggesting that error rates "may not be rare or benign" and demonstrated that tests with more complex scoring procedures were associated with a greater number of scoring errors. This study investigated error rates that resulted from hand scoring seven psychometric tests commonly employed in psychological practice. Significant and serious error rates were identified for both psychologist and client scorers across all tests investigated. Scoring complexity was found to predict the base rate of scorer errors. The findings suggest that greater development in and attention to test-scoring procedures is required to restrict the likelihood of scorer error.

Keywords: error rates, hand scoring, psychological tests, validity, reliability, test design

Despite comprehensive treatment of validity, with a few notable exceptions, test authors appear to assume that their instruments will always be scored perfectly (Allard & Faust, 2000). This assumption appears to hold whether the instrument is to be scored by a professional psychologist, a trained technician or research assistant, or if the instrument is to be self-scored by the respondent. Although the science of human performance, including human error, is a well recognized and legitimate area of industrial research (Lin & Salvendy, 1999; Reason, 1990; Woltz & Gardner, 2000), statistical analyses of error rates and their relationship to basic parameters such as scoring complexity and repetition are only rarely addressed within the empirical research literature describing confidence intervals of test use. Allard and Faust (2000) pointed out the paucity of attention to scoring errors on personality tests, and the considerable body of work investigating clerical errors on the Wechsler Scales (e.g., Connor & Woodall, 1983; Franklin, Stillman, Burpeau, & Sabers, 1982; C. K. Miller, Chansky, & Gredler, 1970; Sherrets, Gard, & Langner, 1979; Slate & Chick, 1989; Sullivan, 2000; Whitten, Slate, Shine, & Raggio, 1994) stands in stark contrast to the relative neglect of career interest inventories and other tests generally.

Of the error rate research that has been conducted in the career interest area, most if not all has been concentrated on self-scoring error rates by clients using the various editions of Holland's Self-Directed Search (Bickham, Miller, O'Neal, & Clanton, 1998; Christensen, Gelso, Williams, & Sedlacek, 1975; Cummings & Maddux, 1987; M. J. Miller, 1997; Tracey & Sedlacek, 1980). Although limited to test takers themselves, this body of research is particularly instructive as it has repeatedly demonstrated, across a series of revisions, that hand-scoring errors by clients who self-score are a significant phenomenon. In recognition of the importance of these findings, Holland, Powell, and Fritzsche (1994) have addressed the issue of scoring errors made by test takers in their current professional user's guide for this instrument (pp. 15-16). Although Holland et al.'s discussion of client scoring errors within their professional user's manual is commendable, such a discussion within a professional manual is unusual and even then

Please address requests for reprints to Roland Simons, Australian Centre in Strategic Management, Queensland University of Technology, GPO Box 2434 Brisbane 4001, QLD, Australia; e-mail: r.simons@qut.edu.au.

Assessment, Volume 9, No. 3, September 2002 292-300

does not extend to a discussion of errors that could be anticipated from scoring undertaken by professional users. It is notable that the topic of hand-scorer error rate for both professionals and nonprofessionals alike is usually absent from most manuals of career interest tests in common use today. At best, a technical manual will encourage that scoring is performed carefully (e.g., Bass & Avolio, 1995; Myers, McCaulley, Quenk, & Hammer, 1998). Likewise, in test guidelines provided by professional associations (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Australian Psychological Society, 1997; Canadian Psychological Association, 1996), there is little acknowledgement of scoring errors made by professionals. Indeed, there appears to be a general assumption that professionals will make no scoring errors. The possibility for professional error is only touched on in codes of conduct; for example, the Australian Psychological Society (1997) guidelines indicate, "psychologists must ensure that assessment procedures are chosen, administered and interpreted appropriately and accurately" (p. 2). Few provisions for reducing scoring errors have been put forward by these associations, with the only form of error discussed being measurement error of the instruments themselves.

Of the relatively limited number of studies investigating hand-scoring error rates that have been conducted and reported, all have concluded that for both client and professional scorers, the percentage of errors that can be expected when psychometric tests are scored by hand will be significantly greater than zero. Possibly reflecting a complicated scoring system, research into the Wechsler scales has consistently found a high incidence of clerical errors. For example, Sherrets et al. (1979), in their investigation of errors on 200 Wechsler Intelligence Scale for Children (WISC) protocols, reported that nearly 90% of examiners were found to have made at least one clerical error. Similarly, Levenson, Golden-Scaduto, Aiosa-Karpas, and Ward (1988) reported a 57% clerical error rate in 162 WISC-R protocols. Both in studies involving the Wechsler scales (e.g., Sullivan, 2000) and other instruments (e.g., Allard, Butler, Shea, & Faust, 1995; Bickham et al., 1998; Charter, Walden, & Padilla, 2000; M. J. Miller, 1997), research has consistently shown that scoring errors can often result in significant changes in standardized scores or shifts in career interest profiles or preferences from that which would have been reported had scoring errors not occurred.

In the course of addressing the dearth of empirical work investigating errors in objective personality tests, Allard and Faust (2000) have suggested that error rates "may not be rare or benign" (p. 120). Their study subdivided error rates into categories according to scoring procedure complexity and scorer commitment to accuracy and found a 29% error rate for the high complexity and low commitment category of scoring and scorers. These findings are important as they raise concern about probable error rates on other instruments for which hand-scoring error rates are unreported (e.g., Vocational Interest Survey for Australia [VISA], Rothwell Miller Interest Blank [RMIB]) and suggest that the applied context in which instruments are scored may significantly influence hand-scoring error rates by clients and professionals alike.

Clearly, along with other technical data describing validity, the likelihood of errors that arise from hand scoring is a pertinent consideration when selecting and evaluating the suitability of an instrument for a specific use, particularly, one would think, where client self-assessment is involved. As a response to the dearth of empirical research into error rates, this study has sought to investigate hand-scoring error rates in a range of psychometric test instruments that are commonly used in the field of occupational psychology today.

THE PRESENT STUDY

For the purposes of this study, the Australian unemployment industry was considered an appropriate source of data. This industry is characterized by large-volume psychometric testing and occupational evaluations. Based on the results of testing, interviews with professionals, and group training exercises focusing on job acquisition skills, employment service organizations may recommend individuals to complete further training and development or even recommend disability/welfare arrangements as an alternative to employment. Incorrectly applied, these courses of action can have a marked impact on how individuals are treated by professionals, their peers, and on their psychological well-being. In this context, therefore, there are strong professional and legal imperatives to ensure that scoring errors do not occur and perhaps equally strong motivations for believing that they do not occur frequently.

The present study has investigated hand-scoring errors by both client test takers and professional psychologists by reexamining a large volume of psychometric test results pertaining to the work of several occupational psychologists conducted over a 3-year period to the year 2000. From this work, only test results pertaining to seven commonly used psychometric assessment tools were reexamined to determine average error rates. The central hypothesis for this study was that hand-scoring error rates for all seven instruments evaluated would be significantly greater than zero. This was Hypothesis 1. In addition, a number of subsequent hypotheses were evaluated.

Using the classification system of human errors proposed by Reason (1990), that is, the knowledge-based,

rule-based, and skill-based mistake trilogy, this study also formulated the following hypothesis: As professional psychologists, on average, could be expected to have greater skills and knowledge of scoring each of the psychological tests evaluated, it was hypothesized that the hand-scoring error rates of psychologists would be significantly lower than the self-scoring error rates of client respondents. This was Hypothesis 2. Finally, as other researchers (Allard & Faust, 2000; McGrew, Murphy, & Knutson, 1994; Whitten et al., 1994) have advocated the importance of the relationship between reduced scoring complexity and scoring accuracy, this study hypothesized that the number of items and relative scoring complexity of each of the seven tests evaluated would be positively associated with errors rate. This was Hypothesis 3.

METHOD

Sample

The results of 1,453 psychological test results collected over a 3-year period by an Australian private sector employment agency operating nationally were made available for this research. Original completed test booklets were chosen from the psychologist case files of eight psychologists who were responsible for all psychometric testing of agency clients in offices located in five states of Australia. The test results represented the complete case data collected by the psychologists, with the exception of clients who were referred on to disabilities pension. In all cases, data were drawn from training and development sessions and had involved the supervised self-scoring of a series of psychological tests by clients as part of the training exercise. In addition to self-scored client ratings, hand-scored ratings undertaken by an occupational psychologist were also available. This unique opportunity was possible because both sets of ratings had been conducted and stored in separate locations. This dual scoring system was created as a result of a program of research initiated by the organization.

Little data on the clients could be gathered beyond that provided on the test sheets. Data were composed of both male (68.3%) and female (31.7%) clients, and the only criterion for the training sessions was that clients had been unemployed for a period of 6 months or more. Modal age of the clients was 29 years, with the majority completing Year 12 level education (45.3%). A number of sessions were directed specifically at white-collar employees, although a majority were run as generic sessions for all clients who were interested in participating in the national self-help training program being offered.

Psychometric testing sessions typically comprised 4 to 10 clients and were integrated within a range of training and self-analysis exercises that made up the 2- to 5-day self-help training programs. Typically, only one psychometric test was used in each exercise. Sessions were conducted in six Australian capital cities. A psychologist, who administered several psychometric tests usually over the course of 2 to 4 days, supervised each testing session. All participants were literate and were registered with a national employment service agency as unemployed clients seeking employment.

Instruments and Measures

Seven different survey instruments were selected for evaluation. Instruments were selected to provide a wide spread of the different types of measures currently in use and to allow for a wide range of complexity in scoring mechanisms to be represented. Of the test instruments selected, two were vocational interest surveys (VISA and RMIB), one a clinical measure of depression (Beck Depression Inventory II [BDI-II]), two popular measures of personality (Myers Briggs Typology Indicator [MBTI]–Form M) and values (Competing Values Managerial Skills Instrument [CVMSI]), and two measures of psychological aptitude (ACER Higher Test ML-MQ [ML-MQ] and Multifactor Leadership Questionnaire–5X Revised [MLQ-5Xr]).

BDI-II. Originally developed in 1961 and later revised in 1971 (Groth-Marnat, 1990), this instrument is composed of 21 items and was designed to measure characteristic attitudes and symptoms of depression (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). The instrument takes about 10 minutes to complete, and each question is scored from zero to three. The manual for the BDI-II discusses self-administration by clients (p. 7), and the subsequent discussion of scoring (p. 10) does not specifically address the issue of client scoring (Beck, Steer, & Brown, 1996). Scoring is done on a simple additive basis, and labels are offered for bands of scores. No mention of scoring error is made in the manual.

RMIB. Originally developed in the 1950s, the RMIB was designed as a practical aid for career counselors and others who needed a focus for a guidance interview. The RMIB is a comparative measure that examines respondents' interest across 12 fields of work. Designed for respondents to be "simple and quick" (p. ix), the manual indicated the instrument's suitability for individual or group administrations (p. 8) and, in the case of senior students or similar groups, to be self-scored under supervision (K. M. Miller, Tyler, & Rothwell, 1994, p. 11).

Respondents are asked to rank, in order of preference, nine sets each of 12 jobs representing the fields. The instrument takes approximately 15 minutes to complete, and scoring is done on the form itself, either by following the procedures laid out in the manual or as described by the test administrator. No mention of scoring error is made in the manual, although an arithmetical method to check computational accuracy is described (K. M. Miller et al., 1994).

MBTI–Form M. Based on the Jungian psychological typology, this instrument was designed as a personality inventory (Myers & McCaulley, 1985). Form M of this instrument is made up of 93 items that result in 16 personality types based on the scoring keys. In this study, the self-scorable version of the instrument was used. The instrument takes approximately 25 minutes to complete. No mention of scoring error is made in the manual.

VISA. The VISA was designed to assess vocational interests in Australian populations. Developed by applying factor analysis to the occupational interests of Australian high school students, the VISA was specifically designed to be self-scored by test takers (Pryor, 1995). The VISA comprises 64 work statements, each requiring respondents to rate how much they think they would like or dislike the work activity along a 7-point scale. The 64 items provide an assessment of vocational interest across eight interest dimensions; each scale is composed of eight items. Scoring is done by adding particular item scores and plotting on a profile form. No mention of scoring error is made in the manual.

MLQ-5Xr. The MLQ measures three categories of leadership factors: (a) transformational leadership, (b) transactional leadership, and (c) nontransactional leadership. The instrument is composed of 41 items that are rated on a 5-point Likert-type scale (1 = *not at all* to 5 = *frequently, if not always*) (Bass & Avolio, 1995). Test takers typically rate their own responses on the self-rating form. Scoring keys are provided for professional examiners, and conversion tables allow for the calculation of standardized scores. No mention of test-taker scoring or scoring error is made in the manual.

ML-MQ. This instrument was designed to assess general intellectual ability based on verbal and numerical test components. The test contains 68 multiple-choice test items and typically requires 35 minutes to complete (Australian Council of Educational Research, 1981). Suitable for group administration, scoring is facilitated by a list of correct responses provided in the manual and a simple additive score key. Conversion tables are provided in the manual to allow standardized scores and IQ equivalents to be calculated. No mention of test-taker scoring or scoring error is made in the manual.

CVMSI. This instrument has been designed to measure leadership value structures that feed into eight separate leadership roles and is provided to test users as part of a book, not as a separate instrument with a manual (Quinn, 1988). Scoring mechanisms for each of the eight leadership roles are provided in the book. The instrument is composed of 32 items, each of which is rated along a 7-point scale. The instrument takes approximately 10 minutes to complete. No mention of scoring error is made in the book.

Each of the eight psychologists involved in this investigation was asked to rank the test instruments being investigated in order of scoring procedure complexity against two criteria. First, psychologists were asked to rank instruments according to the volume of calculations required to arrive at the final profile. Second, psychologists were asked to rank the instruments in order of perceived ease of scoring based on scoring instructions and the associated layout of the test items and any accompanying calculation sheets. From averaging these responses, two measures of rank order of scoring complexity were calculated and used as the basis of further evaluation. The average rank orders of complexity are presented in Table 1.

The principal dependent variable for the present study was hand-scoring error rate. An error occurred when a difference between the final hand-scored values for a test differed from the values obtained when the scores were calculated using a computer-scoring algorithm and subsequent rechecking of hand scoring identified an error.

Procedure

For the purposes of this study, the raw data for all test instruments were entered twice and compared for accuracy using a computer algorithm. After obtaining concordance between the two sets of computer-entered data, these data were then used to computer score all test instruments. Scoring algorithms for each test were calculated and checked for accuracy, a procedure recommended by Allard et al. (1995) to enhance scoring accuracy. Records of client and psychologist scores derived from hand scoring were then checked for accuracy against computer scores generated from the electronic data and scoring algorithms. Where a discrepancy between the computer and hand-scored data was detected, the original hand scoring was reexamined so that any and all scoring errors were clearly identified and could be classified as either transcription or calculation errors. In this way, error rates were identified first by computer rescoring, which was followed

TABLE 1
Psychologists' Ranking of Each Instrument's Scoring Procedure Complexities

	Complexity (Number of Calculations Required)	*Complexity (Instructions and Layout)*
Psychologist rankings of complexity		
ML-MQ	1 (lowest)	2
BDI-II	2	1 (lowest)
RMIB	3	3
MBTI–Form M	4	4
VISA	5	6
CVMSI	6	7 (highest)
MLQ-5Xr	7 (highest)	5

NOTE: ML-MQ = ACER Higher Test ML-MQ; BDI-II = Beck Depression Inventory II; RMIB = Rothwell Miller Interest Blank; MBTI–Form M = Myers Briggs Typology Indicator–Form M; VISA = Vocational Interest Survey for Australian; CVMSI = Competing Values Managerial Skills Instrument; MLQ-5Xr = Multifactor Leadership Questionnaire–5X Revised.

by confirmation of the error by human inspection of the original scoring worksheets.

RESULTS

Each test instrument was examined and classified as either having been scored correctly (no errors) or scored incorrectly (one or more errors). Error totals are therefore the total numbers of administrations examined that contain one or more scoring errors. Error totals and percentage error rates arising from test-taker/client self-scoring and arising from scoring by a professional psychologist are presented in Table 2. Across all instruments, hand-scoring error rates were significantly greater than zero for both self-scoring, $t(1, 1452) = 27.81$, $p < .001$, and psychologist-scoring categories, $t(1, 1455) = 9.50$, $p < .001$. These average findings were true for each of six of the seven instruments investigated, with the exception being the ML-MQ when scored by professional psychologists. Therefore, Hypothesis 1 found substantial support.

Inspection of Table 2 indicates that for each instrument, self-scoring was associated with a higher error rate than scoring by a professional psychologist. Further analysis of error rates within specific tests identified greater than 40% of clients had made at least one error on the VISA, BDI-II, CVMSI, and MLQ-5Xr. Although significantly lower error rates resulted when tests were scored by professional psychologists, error rates still exceeded 5% on the RMIB, MBTI–Form M, VISA, and CVMSI. A series of *t* tests indicated that the observed difference between client and psychologist error rates was significant for all instruments, $t(1, 2123.06) = 20.71$, $p < .001$, providing support for the present study's second hypothesis (see Table 3).

As errors were detected by the algorithmic checking sequence, the original scoring sheets were reexamined, and all errors occurring on the sheet were identified by hand. In this way, the number of errors that were made on each score sheet was tallied. These data are summarized in Table 4. Furthermore, this checking sequence enabled errors to be identified and categorized according to how they were made (calculation or transcription error) and according to whether they could be considered serious or benign. Note that the percentages of errors presented in Table 4 are expressed as a proportion of the total number of errors rather than total number of administrations. Errors occurring as the result of a mistake when performing a calculation were coded as calculation errors. Errors resulting from a mistake in transcribing a score or subtotal were classified as transcription errors. Furthermore, as other authors (e.g., M. J. Miller, 1997) have distinguished "serious" errors as those that result in a changed scoring profile or STEN/STANINE score, cases where incorrect profiles occurred were also tallied. Table 5 presents a breakdown of error types that were observed for each instrument and scorer category.

The breakdown of error totals by error type (calculation or transcription) presented in Table 5 indicates that calculation errors are a notable source of errors for both test takers and psychologists alike. It is interesting that the BDI was associated with the lowest proportion of multiple errors for both groups and the CVMSI survey was associated with the greatest proportion of multiple errors for both groups. Perhaps importantly, Table 1 shows that the BDI's scoring procedure was considered by the participating psychologists to be of low complexity, whereas the CVMSI was considered to have one of the most complex scoring procedures of the instruments investigated.

When the analysis examined whether a scoring error resulted in an incorrect profile, looking at both the overall rate of serious errors and the breakdown of the errors themselves into serious or otherwise was undertaken. Overall percentages indicated that test-taker or client self-scoring resulted in 9.3% of all administrations being profiled incorrectly, and psychological scoring resulted in only 2.5% of administrations concluding with incorrect profiles. These serious error rates contrast with proportions derived from the error data alone. The breakdown of all administrations containing errors indicated that 26.7% of self-scoring errors and 42.3% of psychologist errors result in serious profile differences. Therefore, although psychologists make fewer serious errors than test takers, when they do make an error or series of errors, it is more likely to result in a changed profile.

TABLE 2
Hand-Scoring Error Totals and Error Rates for Test Takers Who Self-Scored and for Psychologist Scorers

Scorer	*Test Instrument*	*Total Test Administrations*	*Error Count*	*Percentage of Errors*	t[a]	df	p
Test taker	ML-MQ	198	41	20.71	7.17	197	.000
	BDI-II	108	46	42.59	8.91	107	.000
	RMIB	206	44	21.36	7.46	205	.000
	MBTI–Form M	315	76	24.13	9.99	314	.000
	VISA	137	91	66.42	16.40	136	.000
	CVMSI	210	97	46.19	13.39	209	.000
	MLQ-5Xr	157	64	40.76	10.36	156	.000
Total		1,453	505	34.76	27.81	1,452	.000
Psychologist	ML-MQ	179	2	1.12	1.42	178	.158
	BDI-II	105	5	4.76	2.28	104	.025
	RMIB	201	16	7.96	4.16	200	.000
	MBTI–Form M	328	19	5.79	4.48	327	.000
	VISA	152	10	6.58	3.26	151	.001
	CVMSI	198	18	9.09	4.44	197	.000
	MLQ-5Xr	168	6	3.57	2.49	167	.014
Total		1,446	85	5.88	9.50	1,455	.000

NOTE: ML-MQ = ACER Higher Test ML-MQ; BDI-II = Beck Depression Inventory II; RMIB = Rothwell Miller Interest Blank; MBTI–Form M = Myers Briggs Typology Indicator–Form M; VISA = Vocational Interest Survey for Australian; CVMSI = Competing Values Managerial Skills Instrument; MLQ-5Xr = Multifactor Leadership Questionnaire–5X Revised.
a. Tests null hypothesis that error rate would be zero.

TABLE 3
Comparison of Error Rates Arising From Self-Scoring and Psychologist Scoring for Each Instrument

Instrument	*Test Taker Error Rate (%)*	*Psychologist Error Rate (%)*	t	df	p
ML-MQ	20.71	1.12	6.55	226.05	.000
BDI-II	42.59	4.76	7.25	146.25	.000
RMIB	21.36	7.96	3.89	356.29	.000
MBTI–Form M	24.13	5.79	6.69	481.60	.000
VISA	66.42	6.58	13.23	200.75	.000
CVMSI	46.19	9.09	9.25	337.86	.000
MLQ-5Xr	40.76	3.57	8.88	197.07	.000
Total	34.76	5.88	20.71	2,123.03	.000

NOTE: ML-MQ = ACER Higher Test ML-MQ; BDI-II = Beck Depression Inventory II; RMIB = Rothwell Miller Interest Blank; MBTI–Form M = Myers Briggs Typology Indicator–Form M; VISA = Vocational Interest Survey for Australian; CVMSI = Competing Values Managerial Skills Instrument; MLQ-5Xr = Multifactor Leadership Questionnaire–5X Revised. Degrees of freedom are reported as decimals as homogeneity of variance was not assumed.

Using regression analysis, the prediction of error rates based on perceived complexity was conducted. Based on the average ranking of scoring complexity calculations summarized in Table 1, a significant prediction of scoring error was identified for self-scoring, $F(1, 1329) = 43.39$, $p < .001$, but not for psychologist scoring, $F(1, 1329) = 2.69$, $p > .05$. However, when scoring complexity was ranked according to complexity of layout, a significant prediction of error was identified for both client self-scoring, $F(1, 1329) = 51.73$, $p < .001$, and for psychologists, $F(1, 1329) = 4.91$, $p < .05$. The regression equations suggested that scoring errors by psychologists were less strongly associated with complexity of calculations than errors committed by the self-scoring client group.

DISCUSSION

In documenting error rates for both self-scorers and professional psychologists, the present study has supported its first hypothesis and the body of research that has concluded that the percentage of errors that can be expected when psychometric tests are scored by hand will be greater than zero (Allard et al., 1995; Bickham et al., 1998; M. J. Miller, 1997). Error rates for the self-scorers were all higher than 20%, a figure that warrants some concern. In supporting the second hypothesis, the data from the present study highlighted that psychologists are typically more accurate in scoring psychometric tests than self-rating procedures. However, it would be interesting to compare psychologist work environments to determine whether different work environments foster different results. It is well documented (Goddard, Patton, & Creed, 2000, 2001) that the psychologists in the present study would have been working within a framework of demanding workloads.

Discerning the type of error in the test scoring provided another valuable finding as different types of scoring

TABLE 4
Breakdown of Hand-Scoring Errors Into Single and Multiple Error Categories for Both Test-Taker Self-Scoring and Scoring by a Psychologist

	Self-Scoring (n = *1,453 Administrations)*			*Psychologist* (n = *1,446 Administrations)*		
Instrument	*Total Errors* (N)	*Single Errors (%)*	*Multiple Errors (%)*	*Total Errors* (N)	*Single Errors (%)*	*Multiple Errors (%)*
ML-MQ	41	22 (*n* = 9)	78 (*n* = 32)	2	100 (*n* = 2)	0 (*n* = 0)
BDI-II	46	89 (*n* = 41)	11 (*n* = 5)	5	100 (*n* = 5)	0 (*n* = 0)
RMIB	44	82 (*n* = 36)	18 (*n* = 8)	16	69 (*n* = 11)	31 (*n* = 5)
MBTI–Form M	76	30 (*n* = 23)	70 (*n* = 53)	19	47 (*n* = 9)	53 (*n* = 10)
VISA	91	32 (*n* = 29)	68 (*n* = 62)	10	40 (*n* = 4)	60 (*n* = 6)
CVMSI	97	19 (*n* = 18)	81 (*n* = 79)	18	33 (*n* = 6)	67 (*n* = 12)
MLQ-5Xr	64	22 (*n* = 14)	78 (*n* = 50)	6	33 (*n* = 2)	67 (*n* = 4)
Total errors	505	36 (*n* = 183)	64 (*n* = 322)	85	47 (*n* = 40)	53 (*n* = 45)

NOTE: ML-MQ = ACER Higher Test ML-MQ; BDI-II = Beck Depression Inventory II; RMIB = Rothwell Miller Interest Blank; MBTI–Form M = Myers Briggs Typology Indicator–Form M; VISA = Vocational Interest Survey for Australian; CVMSI = Competing Values Managerial Skills Instrument; MLQ-5Xr = Multifactor Leadership Questionnaire–5X Revised. Percentages are based on the total number of errors for clients and psychologists.

TABLE 5
Breakdown of Hand-Scoring Errors Into Error Type Categories for Both Test-Taker Self-Scoring and Scoring by a Psychologist

Scorer	*Instrument*	*Cases With Errors* (N)	*Calculation Errors*	*Transcription Errors*	*Serious—Cases With Incorrect Profile*
Test taker	ML-MQ	41	53	13	6
	BDI-II	46	51	9	16
	RMIB	44	59	2	4
	MBTI–Form M	76	83	10	7
	VISA	91	121	38	49
	CVMSI	97	149	11	12
	MLQ-5Xr	64	23	79	41
Total		505 (34.8%)			135 (9.3%)
Psychologist	ML-MQ	2	2	0	0
	BDI-II	5	5	0	2
	RMIB	16	29	1	2
	MBTI–Form M	19	14	14	5
	VISA	10	10	12	8
	CVMSI	18	17	16	14
	MLQ-5Xr	6	6	8	5
Total		85 (5.9%)			36 (2.5%)

NOTE: ML-MQ = ACER Higher Test ML-MQ; BDI-II = Beck Depression Inventory II; RMIB = Rothwell Miller Interest Blank; MBTI–Form M = Myers Briggs Typology Indicator–Form M; VISA = Vocational Interest Survey for Australian; CVMSI = Competing Values Managerial Skills Instrument; MLQ-5Xr = Multifactor Leadership Questionnaire–5X Revised. Percentages are based on the total number of cases considered for each scorer type, test-taker scorer (*n* = 1,453), and psychologist scorer (*n* = 1,446).

methods resulted in different types of errors and likely frequencies. There were more calculation errors than transcription errors. However, in cases with calculation errors where tests were self-scored, many cases resulted in more than one error per test. This rate was considerably less for psychologist scorers. Of greater concern is the proportion of errors that are likely to increase the proportion of incorrect profile analyses. In 9.3% of cases, self-scorers' data resulted in incorrect profiles, and in 2.5% of cases, psychologists' data resulted in incorrect profiles.

The third hypothesis in the current study was also mainly supported. It was possible to predict self-scorer error, but not psychologist error, based on ratings of complexity of calculation. In addition, error was also predicted for both self-scorers and psychologists according to ratings of complexity of layout. Although psychologists were prone to error due to the complexity of calculation, the increasing use of self-scoring methods in test administration practice warrants the examination of psychometric tests in layout, instructions, and calculation requirements. Calcu-

lation and transcription errors were evident for all seven psychometric measures used in the present study. We concur fully with Allard and Faust (2000) and McGrew et al. (1994), who emphasize the importance of reviewing the scoring mechanisms of test forms to reduce complexity and associated error rate.

In reviewing frequently used tests, a number of mechanisms are feasible to attempt to reduce the potential for calculation and transcription error. For example, transcription errors can be reduced by having clearer formats. In addition, test developers need to construct instruments that require less need for transcribing scores from one section to another. Such repeat actions increase potential for error, and although these may be only "silly mistakes," clearly they contribute to important error. Of more concern is the number of calculation errors evident in the current study. As these arise from calculation requirements (such as summing and dividing), any attempt to minimize the need for these in instruments to be used by self-scorers must enhance their effectiveness in providing usable data.

The present study is beset by a number of limitations. First, like much of the published research on this topic (e.g., Charter et al., 2000; Sullivan, 2000), data pertaining to professional scoring were gathered from a limited number (eight) of psychologists; therefore, any inference here must be cautionary. Second, no information is available on the procedures undertaken to instruct the participants in completing and scoring the tests and how much time was made available for each test. Third, little information was available on the different types of workshops within which testing occurred, time of testing, and outcomes of the workshops. It is possible that client error rates are influenced by factors such as clarity of instructions and nature of testing. Because clients were scoring tests that were of personal interest and of little value beyond workshops, one might speculate that client motivation for getting an accurate profile was not as open to the potential for faking good/bad that might otherwise be expected. The self-selection process into workshops is another factor that suggested that clients might be motivated to get an accurate score. Future research into the relationship between client motivation and error rates may identify a social cost associated with having clients score their own results.

Finally, given the consistency of the results from this study with results published elsewhere and the consistency of the error rates found across instruments investigated in the present study, there is now a strong argument to call for a systematic and comprehensive investigation of this phenomenon. Furthermore, as the current body of evidence suggests that human scoring errors are not unique to specific instruments or types of instruments, this topic demands routine attention in future test design and evaluation. The current dearth of research reporting probable hand-scoring error rates in technical manuals and in the professional literature generally is simply not acceptable, both for the professional practitioner and the consumer of psychological testing services.

REFERENCES

Allard, G., Butler, J., Shea, M. T., & Faust, D. (1995). Errors in hand scoring objective personality tests: The case of the Personality Diagnostic Questionnaire-Revised (PDQ-R). *Professional Psychology: Research and practice*, *26*(3), 304-308.

Allard, G., & Faust, D. (2000). Errors in scoring objective personality tests. *Assessment*, *7*(2), 119-131.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *1999 standards for educational and psychological testing*. Retrieved from http://www.apa.org/science/standards.html#overview

Australian Council of Educational Research. (1981). *ACER Higher Test manual Form ML-MQ*. Canberra, Australia: Author.

Australian Psychological Society Ltd. (1997). *Guidelines for the use of psychological tests*. Retrieved from http://www.psychsociety.com.au/about/testing.pdf

Bass, B. M., & Avolio, B. J. (1995). *Multifactor leadership questionnaire: Manual leader form, rater, and scoring key for MLQ (Form 5x-Short)*. Redwood City, CA: Mind Garden.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory* (2nd ed.). San Antonio, TX: Harcourt Brace.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561-571.

Bickham, P. J., Miller, M. J., O'Neal, H., & Clanton, R. (1998). Comparison of error rates on the 1990 and 1994 revised Self-Directed Search. *Perceptual and Motor Skills*, *86*, 1168-1170.

Canadian Psychological Association. (1996). *Guidelines for educational and psychological testing*. Retrieved from http://www.cpa.ca/guide9.html

Charter, R. A., Walden, D. K., & Padilla, S. P. (2000). Too many simple scoring errors: The Rey Figure as an example. *Journal of Clinical Psychology*, *56*(4), 571-574.

Christensen, K. C., Gelso, C. J., Williams, R. O., & Sedlacek, W. E. (1975). Variations in the administration of the Self-Directed Search, scoring accuracy and satisfaction with results. *Journal of Counseling Psychology*, *22*(1), 12-16.

Connor, R., & Woodall, F. E. (1983). The effects of experience and structured feedback on WISC-R error rates made by student-examiners. *Psychology in the Schools*, *20*, 376-379.

Cummings, R. W., & Maddux, C. D. (1987). Self-administration and scoring errors of learning disabled and non-learning disabled students on two forms of the Self-Directed Search. *Journal of Counseling Psychology*, *34*(1), 83-85.

Franklin, M. R., Stillman, P. L., Burpeau, M. Y., & Sabers, D. L. (1982). Examiner error in intelligence testing: Are you a source? *Psychology in the Schools*, *19*, 563-569.

Goddard, R. C., Patton, W., & Creed, P. (2000). Case manager burnout in the Australian job network. *Journal of Applied Health Behaviour*, *2*(2), 1-6.

Goddard, R. C., Patton, W., & Creed, P. (2001). Psychological distress in Australian case managers working with the unemployed. *Journal of Employment Counseling*, *38*, 50-61.

Groth-Marnat, G. (1990). *The handbook of psychological assessment* (2nd ed.). New York: John Wiley.

Holland, J. L., Powell, A. B., & Fritzsche, B. A. (1994). *The Self-Directed Search (SDS) professional user's guide 1994 edition*. Odessa, FL: Psychological Assessment Resources.

Levenson, R. L., Golden-Scaduto, C. J., Aiosa-Karpas, C. J., & Ward, A. W. (1988). Effects of examiners' education and sex on presence and type of clerical errors made on WISC-R protocols. *Psychological Reports*, *62*, 659-664.

Lin, H. X., & Salvendy, G. (1999). Instruction effect on human error reduction. *International Journal of Cognitive Ergonomics*, *3*(2), 115-129.

McGrew, K. S., Murphy, S. R., & Knutson, D. J. (1994). The development and investigation of a graphic scoring system for obtaining derived scores for the WJ-R and other tests. *Journal of Psychoeducational Assessment*, *72*, 33-41.

Miller, C. K., Chansky, N. M., & Gredler, G. R. (1970). Rater agreement on WISC protocols. *Psychology in the Schools*, *7*, 190-193.

Miller, K. M., Tyler, B., & Rothwell, J. W. (1994). *Rothwell Miller Interest Blank (Revised)*. London: Miller & Tyler.

Miller, M. J. (1997). Error rates on two forms of the Self-Directed Search and satisfaction with results. *Journal of Employment Counseling*, *34*, 98-103.

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

Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1988). *MBTI manual: A guide to the development and use of the Myers Briggs Type Indicator (3rd ed.)*. Palo Alto, CA: Consulting Psychologists Press.

Pryor, R. G. L. (1995). *Vocational Interest Survey for Australia (VISA): Professional manual*. Sydney, Australia: Psychological Corporation.

Quinn, R. E. (1988). *Beyond rational management: Mastering the paradoxes and competing demands of high performance*. San Francisco: Jossey-Bass.

Reason, J. T. (1990). *Human error*. Cambridge, MA: Cambridge University Press.

Sherrets, S., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC protocols. *Psychology in the Schools*, *16*(4), 495-496.

Slate, J. R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. *Psychology in the Schools*, *26*, 78-84.

Sullivan, K. (2000). Examiners' error on the Wechsler Memory Scale–Revised. *Psychological Reports*, *87*, 234-240.

Tracey, T. J., & Sedlacek, W. E. (1980). Comparison of error rates on the original Self-Directed Search and the 1977 revision. *Journal of Counseling Psychology*, *27*(3), 299-301.

Whitten, J., Slate, J. R., Shine, A. E., & Raggio, D. (1994). Examiner errors in administering and scoring the WPPSI-R. *Journal of Psychoeducational Assessment*, *12*, 49-54.

Woltz, D. J., & Gardner, M. K. (2000). Negative transfer errors in sequential cognitive skills: Strong-but-wrong sequence application. *Journal of Experimental Psychology*, *26*(3), 601-625.

Roland Simons is a research fellow in the Australian Centre in Strategic Management, Faculty of Business, at Queensland University of Technology. He lectures in the areas of organizational psychology and applied statistics and pursues applied research in the areas of organizational behavior, leadership, and financial performance. He also works extensively with Australian public and private sector organizations seeking assessment and change.

Richard Goddard is an associate lecturer in the School of Learning and Professional Studies, Faculty of Education, at Queensland University of Technology. He lectures in the areas of developmental psychology and counseling and pursues applied research in the areas of occupational stress, employee well-being, staff training, and career assessment. He has an extensive employment history within the Australian employment service industry where he was an occupational psychologist, staff educator, and regional manager for more than 12 years.

Wendy Patton is the head of the School of Learning and Professional Studies in the Faculty of Education at Queensland University of Technology (QUT). She initiated and developed the Career Guidance Area of Interest in the Master of Education at QUT and presently coordinates these units. She has an extensive publishing record in refereed journals and edited the *Australian Journal of Career Development* from 1997 to 1999. She has coedited and authored a number of theory and practice books in the career development area.