Self-Report Measures of Intelligence: Are They Useful as Proxy IQ Tests?

Delroy L. Paulhus
Daria C. Lysy
Michelle S. M. Yik
University of British Columbia

ABSTRACT Correlations between single-item self-reports of intelligence and IQ scores are rather low (.20–.25) in college samples. The literature suggested that self-reports could be improved by three strategies: (1) aggregation, (2) item weighting, and (3) use of indirect, rather than direct, questions. To evaluate these strategies, we compared the validity of aggregated and unaggregated versions of direct measures with four indirect measures (Gough's Intellectual efficiency scale, Hogan's Intellect composite scale, Sternberg's Behavior Check List, and Trapnell's Smart scale). All measures were administered to two large samples of undergraduates (*N*s = 310, 326), who also took an IQ test. Although results showed some success for both direct and indirect measures, the failure of their validities to exceed .30 impugns their utility as IQ proxies in competitive college samples. The content of the most valid items referred to global mental abilities or reading involvement. Aggregation benefited indirect more than direct measures, but prototype-weighting contributed little.

This research was supported by a grant to the senior author from the Social Sciences and Humanities Research Council of Canada. We wish to thank Harrison Gough, Robert Hogan, Robert Sternberg, Paul Trapnell, and several anonymous reviewers for comments on an earlier draft.

Correspondence should be addressed to Del Paulhus, Department of Psychology, University of British Columbia, Vancouver, Canada, V6T 1Z4. Electronic mail may be sent via the Internet to dpaulhus@cortex.psych.ubc.ca

Journal of Personality 66:4, August 1998.

Can people validly rate their own intelligence? Skeptics argue that such self-reports are hopelessly contaminated with a variety of distortions including self-deception, impression management, and reconstrual. Such defensive reactions may explain the low variance in self-ratings of intelligence: Rarely do people rate themselves as “below average” (McCrae, 1990). Even the most forthright and insightful individuals, skeptics warn, can never confirm the veracity of their self-assessments because the concept itself is so elusive in nature.

Despite its elusiveness, the concept of intelligence plays a central role in psychological research, particularly in such contexts as educational evaluation, personnel selection, and child development. To facilitate such research, considerable effort has been devoted to developing self-report alternatives to cumbersome IQ tests of intelligence. Such approaches have progressed well beyond a simple request to “rate how intelligent you are.” Three strategies, in particular, have been recommended. One is the use of indirect assessment to bypass the inevitable defensiveness of a direct request for a self-rating. A second, the aggregation strategy, favors multiple-item over single-item measures. A third strategy, prototype weighting, takes into account the differential importance of items within a measure. In this report we evaluate these three strategies by determining their ability to improve prediction of performance on IQ tests.

Use of IQ tests as the criterion for intelligence ratings has not yielded high validities,[1] particularly in college samples. The validities are somewhat higher for observer-ratings than for self-ratings. Values in the range of .25 to .50 have been found when the judgment is made by spouses (Bailey & Mettetal, 1977), by friends and strangers (Borkenau, 1993), by adolescent acquaintances (Bailey & Hatch, 1979), and by long-term discussion-group colleagues (Paulhus & Morgan, 1997). Except in the case of spouses, however, the achievement of these solid validities required aggregation across multiple observers.

Self-perceptions typically parallel other-perceptions but, to the extent that the trait being evaluated is highly evaluative (e.g., intelligence), the former are noticeably less valid (John & Robins, 1993). Studies using IQ test scores as the criterion have yielded single-item validities of .32 (Borkenau & Liebler, 1993) and .38 (Reynolds & Gifford, 1996) in

1. The term “validity” is used to mean correlation with a specific criterion. Its use does not imply that IQ is the sole criterion for measuring intelligence.

general population samples. But in college samples the validities never exceed .30: for example, .26 (DeNisi & Shaw, 1977), .25 (Paulhus & Morgan, 1997), and .26 (Reilly & Mulhearn, 1995). This modest level of validity is the starting point for the present report.

Improving Self-Report Measures of IQ

The literature contains a number of self-report instruments that show potential as proxy IQ scales, that is, economical substitutes for IQ tests. If valid, such scales have great practical advantages over their traditional counterparts: rather than running subjects one-by-one in a tightly supervised laboratory setting, researchers can administer such scales quickly to large groups of subjects. Moreover, self-report questionnaires are less threatening than IQ tests and therefore more likely to elicit cooperation.

Of course, such advantages are pointless unless validities can be improved over those values cited above. An ideal proxy scale would represent, in effect, a parallel measure showing a validity equal to the reliability of IQ tests, that is, upwards of .90. Given that standard IQ tests differ somewhat in emphasis, however, a more appropriate upper limit is the correlation between two well-validated IQ tests, that is, roughly, .80–.85 (Thorndike, 1982). Even that level of validity seems unlikely for proxy tests, given that they emphasize typical performance as opposed to the maximal performance tapped by IQ tests (Ackerman & Heggestad, 1997; Paulhus & Martin, 1987).

Potential proxy scales in the literature have relied on three strategies for improving the validity of self-report measures of intelligence. First is the reduction of evaluation-threat by using subtle, nonobvious questions. We use the term “indirect measures” to describe test formats that mask the purpose of the test. The second strategy involves aggregating a set of items to improve reliability. The third strategy is weighting the items according to their importance.

Indirect Measurement Strategy

Rather than referring directly to intelligence, items on indirect measures concern interests, behaviors, personality, and the like. In this report, we examined four such measures: Gough’s Intellectual efficiency (Ie) scale, Hogan’s Intellect composite scale, Sternberg’s Behavior Check List (BCL), and Trapnell’s Smart scale. All four have shown some validity in predicting criterion measures of intelligence. Although all have an indirect format, the rationale for each is rather different.

Gough's Intellectual efficiency (Ie) scale. In the first such effort, Gough (1953) developed a set of self-report items for use as a proxy measure of intelligence. He administered a pool of items assumed to tap aspects of personality associated with intelligence. Those 52 items correlating most highly with an IQ test in a sample of high school students were assembled and labeled the Intellectual efficiency (Ie) scale. In four cross-validation studies, Gough reported a mean validity of .47. In the four educated samples reported in the latest manual, however, the median validity is only .29 (Gough, 1996).

As for most tests derived from contrasted groups, the Ie items are rather heterogeneous; topics included self-confidence, neuroticism, and social skills as well as intellectual abilities and interests. The vast majority were subtle indicators, that is, they lacked face-validity as indicators of intelligence. As such they are less likely to trigger self-presentation.

Hogan's Intellect composite. Welsh (1975) developed the notion of "intellectance" to denote the "cognitive and interpersonal style that causes people to be perceived as bright." Hogan and Hogan (1992, p. 12) followed this peer-perception notion of intellect in assembling a set of items. A factor analysis revealed two factors. One was labeled Intellectance: "the degree to which a person is perceived as bright, creative, and interested in intellectual matters." The other factor was labeled School Success: "the degree to which a person seems to enjoy academic activities and to value educational achievement for its own sake."

Intellectance items refer to science ability, curiosity (about the world), thrill seeking, interest in intellectual games and generating ideas (ideational fluency), and interest in culture items, while School Success items concern education (being a good student), math ability, good memory, and enjoyment of reading. Observers tend to see high scorers on the Intellectance scale as "imaginative, inventive, and quick-witted, but easily bored and inattentive to detail" whereas low scorers tend to be "unimaginative, narrow, tolerant of boredom, and not needing much stimulation." In contrast, high scorers on the School Success scale are seen as "foresighted, thorough, and painstaking," whereas low scorers are seen as "touchy, restless, and impulsive" (Hogan & Hogan, 1992, p. 40).

Sternberg's Behavior Check List (BCL).[2] As part of his investigation into conceptions of intelligence, Sternberg (1988) developed the Behavioral Check List (BCL), a list of 41 behaviors that lay judges associated with intelligence (p. 238). Factor analyses indicated three clusters of items labeled Practical Problem Solving (PS), Verbal Ability (VA), and Social Competence (SC). In a community sample, correlations of the full-scale BCL with an IQ test were found to be .24 with unit-weighting of all items, and .52 when items were weighted according to diagnosticity. All these results were later replicated by Cornelius, Kenny, and Caspi (1989).

Sternberg recommended the BCL as a valuable supplementary measure of intelligence for a number of reasons. Compared to providing a global assessment of their ability, subjects should feel less-threatened by rating specific behaviors and, accordingly, be more accurate. A composite of a large set of these specific behaviors could then yield a maximally valid self-report. Finally, the BCL coverage was designed to extend beyond those aspects of intelligence measured by IQ tests (Sternberg, 1988).

Trapnell's Smart scale. The 4-item Smart scale assesses self-appraised intelligence indirectly via statements about the respondent's social reputation (Trapnell, 1994). The content of three of the items was based on the assumption that range restriction in self-ratings due to desirable responding can be reduced by the use of extreme qualifiers (e.g., "very," "extremely," "exceptionally") and by shifting the implied locus of evaluation from the self to others (e.g., "I'm considered to be . . ." in place of "I am . . ."). A fourth item assessed self-reported school grades, based on the assumption that grades provide an indirect but objective index of mental ability that can be recalled and self-reported fairly accurately. The Smart scale correlated .33 with an IQ test in a college sample (Trapnell & Scratchley, 1996).

Aggregation Strategy

Aggregation is a widely accepted strategy for improving the reliability (and therefore the predictive validity) of a measure by decreasing its error

2. Note that this instrument is not a checklist in the strict sense of requiring respondents to check off answers. Instead, the items are rated in a Likert format.

of measurement (for a strong case, see Epstein, 1983). Other things being equal, the addition of items similar to those already included should increase the validity of a self-report intelligence measure. In fact, the amount of improvement in reliability and validity can be estimated with available prophecy formulas (e.g., Gulliksen, 1967).

The reader may have noted, however, that the number of items in the indirect measures reviewed above varies dramatically (from 4 to 52). This natural variation should allow us to evaluate the utility of aggregation between-measures as well as within-measures.

Weighting Strategy

Typically, aggregation is performed simply by summing or averaging the available items: That is, equal (unit) weights are applied to all items. Although psychometricians usually recommend such unit weights (see Thorndike, 1982), weighting of items by their importance remains an appealing strategy.

One variant of importance weighting was exploited by Sternberg, Conway, Ketron, and Bernstein (1981) in order to improve the validity of the Behavior Check List (p. 49). Their technique, which we will label "the prototype weighting procedure," involved developing a set of weights corresponding to the diagnosticity of each item. A set of judges was asked to rate each BCL item in terms of "how characteristic it is of an ideally intelligent person" (p. 42). Instead of just adding up a respondent's responses to yield a total score, the responses were correlated with the corresponding diagnosticity ratings.

Note that a Pearson correlation is simply the mean of the item-by-item products of the two standardized variables. Therefore this procedure is equivalent to standardizing the 41 weights, multiplying them by the respondent's 41 standardized (within-subject) responses, and averaging the 41 products. The average then represents the subject's composite score.[3] Application of this weighting system by Sternberg et al. (1981) boosted the validity of the BCL up to the .50 range (values of .42–.46 were obtained by Cornelius, Kenny, & Caspi [1989]). Sternberg and

3. A similar approach has been used for some time with Q-sort data where a subject's self-sort is correlated with a particular criterion sort (e.g., Block, 1961, 1971). Again, a higher score indicates that the respondent assigned his/her highest ratings to items that were weighted the highest.

colleagues concluded that "a good estimate of IQ can be obtained, based on correspondence between a person's self-perceived pattern of behaviors and the pattern of behaviors in an ideal person" (p. 50).

The Present Study

The self-report measures examined in this report differ with respect to directness (direct vs. indirect) and aggregation (single-item vs. composites). Thus each falls into one of the four categories of a 2 × 2 table (see Table 1). The first category—single-item direct measures—is represented by the adjective "intelligent" (Sample 1) and "clear-thinking, intelligent" (Sample 2). The composite direct measure combined a set of four conceptually similar items. The four indirect measures were Gough's Ie, Hogan's Intellect, Sternberg's BCL, and Trapnell's Smart scale. To evaluate the fourth category of measures—single-item indirect—we calculated the average item validity for each indirect measure. Finally, we compared weighted and unweighted versions of the BCL.

Within the composite indirect category, we also had a specific interest in the comparative validity of the four measures. Although there is some evidence for the validity of each scale, they have never been pitted against one another in predicting a common criterion. A comparative validity study by nonpartisan researchers should provide much more convincing evidence than that offered by the authors of individual scales.

All measures were administered to two large and diverse samples of undergraduates. The criterion for validity of the self-reports was the Wonderlic Personnel Test, a 12-minute IQ test that compares favorably with longer IQ tests. Our analyses focused on comparing the validities of the four categories of measures via correlation and regression techniques. To put all these validities in perspective, we also estimated the performance of ideal proxy scales.[4]

METHOD

Subjects and Procedure

Data were collected from a total of 636 undergraduate students at the University of British Columbia. Sample 1 comprised 310 students (95 males; 208 females;

4. We did not include factor five measures such as Goldberg's Intellect Scale because they were targeted at personality, not intelligence.

Table 1
A Two-Factor Taxonomy of Self-Report Measures of Intelligence

	Aggregation Strategy	
	Single Item	Aggregated Items
Directness Strategy		
Direct	Global rating of intelligence	Global rating of intelligence plus similar items
Indirect	Average item from an indirect measure	Full indirect measures (BCL, Ie, Intellect, Smart)

7 did not specify their gender) enrolled in an introductory psychology course. Sample 2 comprised 326 students (87 males; 205 females; 34 did not specify) enrolled in a second year social-personality psychology course. Approximately 55% of the two samples were liberal arts majors, 20% science or engineering majors, and 15% business majors. All participated for extra marks.

For both samples, subjects were first asked to complete a self-report inventory in group sessions. It included all the direct self-ratings of intelligence. Later, a set of indirect measures of intelligence was distributed in a take-home package, which subjects were asked to complete privately and return for experimental credits. Finally, the IQ test was administered in a separate, supervised session.

Instruments

Direct measures. A number of intelligence-related items were included in the context of a larger personality inventory. They were selected a priori for their conceptual relevance to intelligence. In Sample 1 they included the following four items: “Is intelligent”; “Is ingenious, a deep thinker”; “Is smart”; and “Is not exceptionally gifted at academic things” (Reverse coded). In Sample 2 the direct items included: “Is clear-thinking, intelligent”; “Wants things to be simple and clear-cut”; “Is clever, sharp-witted”; and “Enjoys thinking about complicated problems.”[5] Subjects were asked to rate their agreement with these items on a scale ranging from “1” (“Disagree strongly”) to “5” (“Agree strongly”). For special consideration we identified the most face-valid item (“Is intelligent” in Sample 1; “Is clear-thinking, intelligent” in Sample 2).

5. Despite its conceptual relevance, we did not include the item “ingenious” in Sample 2 because of confusion the item caused. Apparently, some of our subjects thought the item meant “not a genius.”

To evaluate the utility of aggregating items, we combined the four items judged by three raters to be the most face-valid indicators of intelligence. Although the scale items differed somewhat in the two samples, the similarity of correlates (see Results section) suggests that the two direct composites measured a similar construct.

Indirect measures. Given our review of the four indirect measures in the Introduction, we will provide only a basic description here. The Intellectual efficiency (Ie) scale of the California Psychological Inventory (Gough, 1953) included 52 True-False statements. The content includes personality-related items, ranging in content from beliefs (e.g., “Success is a matter of will power”) and interests (e.g., “I like to read about history”) to bizarre items about experiences (e.g., “I have never seen a vision”).

Portions of two subscales of the Hogan Personality Inventory (Hogan & Hogan, 1992) were also included to represent the Intellect factor. All items are in True-False format and were developed from a peer-perception view of intellect. For reasons of convenience and space, we limited our selection to 15 items from the Intellectance subscale and 7 items from the School Success subscale. Examples are “I’m good at inventing games, stories, and rhymes” from the Intellectance subscale, and “As a child I was always reading” from the School Success subscale.[6]

The Sternberg BCL consists of short, specific, behavioral descriptions originally selected by lay judges as prototypical of intelligent people (Sternberg et al., 1981). We used the final 41-item version provided by Sternberg (1988, pp. 238–239). Our subjects rated from “1” (low) to “9” (high) the extent to which each item was an “accurate self-description.” The BCL includes three subscales: the 13-item Verbal Ability subscale (e.g., “Speaks clearly and articulately”), the 15-item Problem Solving subscale (e.g., “Makes good decisions”), and the 13-item Social Competence subscale (e.g., “Responds thoughtfully to others’ ideas”).

The 4-item Smart scale measures self-appraised intelligence via simple trait descriptive statements of high face validity (Trapnell, 1994). As with the BCL, subjects rated from “1” (low) to “9” (high) the extent to which each item was an “accurate self-description.” The items are: (1) “I’m considered exceptionally or unusually intelligent”; (2) “I’m considered a very ‘brainy,’ scholarly person”; (3) “I’m considered extremely ‘gifted’ or talented at academic things”; and (4) “My school grades have usually been near the top of every class.”

Objective measure (IQ test). The 12-minute Wonderlic Personnel Test was chosen to assess IQ (Wonderlic, 1992). It is a short-form test of general cognitive

6. Two of the Intellect–School Success items are identical to two of the Intellectual efficiency items: We only included them once in our inventory.

ability, that is, “the level at which an individual learns, understands instructions and solves problems” (Wonderlic, 1992, p. 5). Included are items sampled from verbal, quantitative, and analytic domains. Although a time limit is imposed, the Wonderlic behaves more like a power test than a speeded test[7] because the items are presented in ascending order (McKelvie, 1994).

The Wonderlic is very popular in applied settings because of its ease of administration and comprehensive norms combined with ample reliability and validity evidence. Expert reviews have been highly favorable (see Aiken, 1996; Hunter, 1989; Schmidt, 1985; Schoenfeldt, 1985).

The Wonderlic shows test-retest reliabilities ranging from .82 to .94 (Dodrill, 1983; Wonderlic, 1992), and alternate-form reliabilities ranging from .73 to .95 (Wonderlic, 1992). These findings are based on adult working populations, however. Because of restriction of range of ability, college samples should yield lower standard deviations, and therefore lower reliabilities. McKelvie (1989) reported a high internal consistency of .87 (odd-even split-half correlation) in a college sample. The fact that reliability is not increased by relaxing the time requirement (McKelvie, 1994) indicates that the time limit does not inflate the estimate.

In support of concurrent validity, the Wonderlic shows correlations above .80 with longer IQ tests such as the WAIS-R (Dodrill, 1981; Wonderlic, 1992). In fact, Dodrill (1981, p. 668) reported that the Wonderlic IQ scores were within 10 points of the WAIS Full Scale IQ scores in 90% of the cases. Of particular note for this report is the fact that correlations are high with measures of both verbal and quantitative abilities (Wonderlic, 1992). Previous studies in college populations have also shown useful predictive validity for college grades (McKelvie, 1994), performance tests (Kennedy, Baltzley, Turnage, & Jones, 1989), and supervisory rankings (Wonderlic, 1992).

RESULTS

Descriptive Statistics

Means, standard deviations, ranges, and reliability coefficient alphas are presented in Table 2. The values of these statistics in the two samples are virtually identical. Alpha values for the full scales and subscales are generally quite acceptable, ranging from .61 to .93 in Sample 1 and from .55 to .92 in Sample 2. The reliability of the single-item “intelligent”[8] was estimated from the mean intercorrelation of the four global items from the direct composite.

7. A true speeded test comprises all easy questions.
8. The item was “intelligent, clear-thinking” in Study 2.

Table 2
Descriptive Statistics

Scale	# Items	Rating Scale	Item Mean	Standard Deviation[a]	Range[a]	Alpha
Global Intelligence Ratings						
Composite Scale[b]	4	5-point	3.66	.67	3.75	.68
			3.24	.69	3.50	.65
Single-item[c]	1	5-point	4.02	.77	4.00	.46[d]
			3.94	.82	4.00	.39[d]
Gough Intellectual	52	True-False	.68	.11	.62	.72
efficiency (Ie) scale	.68	.10	.54	.67		
Trapnell Smart scale	4	9-point	5.45	1.54	8.00	.86
			5.09	1.65	8.00	.88
Hogan Intellect composite	22	True-False	.59	.19	.91	.73
			.58	.20	.91	.77
Intellectance	15	True-False	.62	.21	1.00	.72
			.62	.23	1.00	.76
School Success	7	True-False	.52	.27	1.00	.61
			.48	.26	1.00	.55
Sternberg Behavior Check	41	9-point	6.24	.82	4.63	.93
List (BCL)		6.40	.75	4.10	.92	
Verbal Ability	13	9-point	6.02	.96	5.15	.80
			6.12	1.03	5.38	.82
Practical Problem Solving	15	9-point	6.26	.92	5.87	.89
			6.42	.80	4.13	.88
Social Competence	13	9-point	6.45	.84	4.38	.77
			6.67	.72	4.31	.71

Note. Top row of each cell is from Sample 1 (N = 310); Bottom row is from Sample 2 (N = 326). [a]Standard deviations and ranges are calculated across subject means rather than across item means. [b]Includes "intelligent" and three conceptually similar items. [c]"intelligent" (Sample 1); "clear-thinking, intelligent" (Sample 2). [d]Mean intercorrelations among the four global items.

Not in the table are the statistics for the Wonderlic IQ test. Our Sample 1 and Sample 2 means (25.5, 26.3) were only slightly higher than the manual norms for college students (Wonderlic, 1992, p. 38). Our SDs (4.41, 4.72), however, were substantially lower than the manual norms of 5.73 for college students.[9] For comparison, note that the norms computed on a representative adult working population (p. 38) exhibited a substantially lower overall mean (21.6) and higher SD (7.1).

Although parallel forms is preferable for estimating the reliability of the Wonderlic, we did not have that information. Instead, reliabilities for the Wonderlic were estimated in two ways. First, we estimated internal consistency directly in our sample with the odd-even split half-reliability used by McKelvie (1989). Our values were .79 and .83 in Samples 1 and 2, respectively. A second calculation involved extrapolating from the appropriate reliability estimates (.90) taken on the broad norm sample (Wonderlic, 1992). Applying the correction formula from Gulliksen (1967, p. 124), to the reduction in standard deviation from 7.12 to 4.41 and 4.72, the alphas in our sample were estimated to be .74 and .77. Using either estimation formula, the reliabilities in our college sample were noticeably lower than in the general population, but certainly within the useful range for research instruments. We can expect that our validities, in turn, will be correspondingly lower than those calculated on the general population.

Range of Responses

As noted earlier, the strong tendency for respondents to claim high levels of intelligence tends to restrict the range of responses, skew the response distribution, and constrain correlations with other variables (McCrae, 1990; Thorndike, 1982). Note that the SDs shown in Table 2 were calculated on the subject means, rather than calculating the means of the item SDs. Given that the latter figures are more relevant to whether or not our subjects were using the entire range of our rating-scales, we proceeded to calculate those figures.

Recall that we measured the direct items on 1-to-5 rating scales: The exact distribution of responses was (0, .05, .25, .40, .30) across the two samples. The SD for the single direct item was only .80 and .85 in Samples 1 and 2, respectively. For the four items of the composite direct scale, the average SDs were still small: .93 (Sample 1) and .98 (Sample 2). Compare these values with the average SD of 1.12 for a set of

9. Estimated from the manual norms for mean and women weighted according to gender ratio in our samples.

personality items in the same test battery. In short, the direct items did show some restriction in range.

For the indirect measures, the means of the item standard deviations and ranges were not relevant for the two True-False scales (Gough's Ie and Hogan's Intellect composite) and were therefore not calculated. The other two indirect measures were administered in identical 9-point response format, but the variation of the Smart scale items was noticeably greater than the BCL items. The average standard deviations were 1.82 and 1.92 for Trapnell's Smart scale, and 1.59 and 1.53 for the BCL. The average range of the items of the Smart scale was fully 8.00 in both samples, higher than that for the BCL, 7.63 and 7.17.

Intercorrelations Among Predictors

The matrix of intercorrelations among the indirect scales and subscales is presented in Table 3. Note that the four indirect measures (not including subscales) intercorrelate positively but only modestly, with correlation coefficients ranging from .08 to .56 (Sample 1) and from .24 to .47 (Sample 2). Sternberg's subscales intercorrelated quite strongly, with correlation coefficients ranging from .67 to .77 (Sample 1) and from .65 to .68 (Sample 2), while Hogan's subscales intercorrelate modestly, with correlation coefficients of .25 (Sample 1) and .35 (Sample 2).

Table 3
Intercorrelations Among Indirect Scales and Subscales

Scale	1	2	3	4	5	6	7	8	9
1. Gough Ie	—	.08	.47 [a]	.42	.32 [a]	.43	.41	.36	.39
2. Trapnell Smart	.24	—	.23	.18	.21	.48	.47	.49	.33
3. Hogan Intellect	.27 [a]	.29	—	.90	.65	.56	.58	.51	.43
4. Intellectance	.27	.24	.92	—	.25	.51	.50	.49	.38
5. School Success	.24 [a]	.25	.68	.35	—	.36	.42	.27	.29
6. Sternberg BCL	.32	.45	.47	.44	.31	—	.88	.93	.89
7. Verbal Ability	.34	.43	.48	.42	.38	.89	—	.72	.67
8. Problem Solving	.29	.50	.43	.44	.22	.89	.67	—	.77
9. Social Competence	.19	.23	.30	.28	.20	.86	.65	.68	—

Note. Correlation coefficients in the upper right of the matrix are from Sample 1 ($N = 310$); coefficients in the lower left are from Sample 2 ($N = 326$). All correlations above .20 are significant, $p \leq .001$, two-tailed. [a]Two overlapping items from the Intellectual efficiency and Intellect-School Success scales were assigned to the latter scale for these calculations.

Predictive Validity

Table 4 contains the validities, that is, the correlations of all self-report intelligence measures with Wonderlic test scores. Our baseline validity is that of the single self-rated intelligence item: These values were .20 (Sample 1) and .23 (Sample 2). The corresponding validities for the composite direct measure were slightly higher: .24 (Sample 1) and .26 (Sample 2).

The ability of the four indirect measures to predict IQ test scores was examined in two ways. First we calculated and compared the validities of each predictor; then we performed regression analyses to determine which of the predictors made independent contributions.

Correlations. Table 4 indicates that all four indirect measures achieved significant validities in both samples. Of the indirect measures, Gough's Ie scale performed best, with validities of .20 (Sample 1) and .34 (Sample 2), followed by Trapnell's Smart scale, with validities of .24 (Sample 1) and .25 (Sample 2). Although not as successful overall, the other measures

Table 4
Correlations of Self-Report Measures With IQ Test Scores

Scale	Number of Items	Sample 1	Sample 2
Direct Measures			
Single-item[a]	1	.20***	.23***
Composite scale[b]	4	.24***	.26***
Indirect Measures			
Gough Ie	52	.20***	.34***
Trapnell Smart	4	.24***	.25***
Hogan Intellect	22	.15*	.22***
Intellectance	15	.08	.13*
School Success	7	.19**	.27***
Sternberg BCL	41	.20***	.13*
Verbal Ability	13	.24***	.18**
Practical Problem Solving	15	.17**	.10
Social Competence	13	.14*	.04

Note. $*p < .05$; $**p < .01$, $***p \leq .001$, two-tailed. Sample size ranges from 274 to 301 (Sample 1) and from 241 to 265 (Sample 2) due to the subject matching across the three sources of data (i.e., direct, indirect, and IQ measures). [a]"Intelligent" (Sample 1); "clear-thinking, intelligent" (Sample 2). [b]Refers to the direct, global intelligence ratings including "intelligent" and three conceptually similar items.

each offered a successful subscale: Hogan's School Success performed well at .19 (Sample 1) and .27 (Sample 2), and so did Sternberg's Verbal Ability subscale, at .24 (Sample 1) and .18 (Sample 2).

Recall that Sternberg (1988) found improved validity via a prototype-weighting procedure (see our Introduction). We followed this procedure by having five expert judges (research colleagues) rate each BCL item for diagnosticity of an ideally intelligent person. With strict adherence to Sternberg's method, however, we found no validity improvement. When we simply weighted without standardizing the BCL item scores, the correlations did improve slightly from .20 to .23 (Sample 1) and from .13 to .17 (Sample 2).

Regression analyses. To determine whether the indirect measures made independent contributions in predicting IQ, we conducted a regression analysis in each sample. When IQ score was regressed on all four indirect scales using simultaneous forced entry, significant weights appeared for Gough's Ie (.17, .29) and the Smart (.22, .19) across both samples. Hogan's Intellect composite was also significant in one sample. The resulting variance accounted for by the four indirect measures was 10% and 16% in the two samples.

A follow-up set of regression analyses was conducted to determine the predictive power of the Intellect and BCL subscales. When the three Sternberg subscales alone were simultaneously force-entered, they accounted for a total of 6 and 4% of the variance in our two samples. A similar forced-entry with the Intellect subscales accounted for a total of 7% of the variance in both samples. As might be expected, regression on the subscales accounted for more of the variance than that achieved by the composite Intellect or BCL scores.

A Two-Factor Organization of Strategies

Table 5 summarizes the key data for this report by displaying the mean validities of the four categories of measures of self-report intelligence. The performance of direct measures can easily be compared with those of indirect measures for both single items and the aggregated scales.

The entries in Table 5 (across the rows) are as follows: The single-item direct validities are the correlations of IQ with the single item "intelligent" (Study 1) or "clear-thinking, intelligent" (Study 2). The aggregated direct entries are the validities of the 4-item direct scales including

Table 5
Correlations of Four Types of Predictors With IQ Test Scores

	Aggregation Strategy			
	Single Item		Aggregated Items	
Directness Strategy	Sample 1	Sample 2	Sample 1	Sample 2
Direct	.20	.23	.24	.26
Indirect	.07	.07	.18	.24

Note. Single-item validities are correlations of IQ scores with the single item, "intelligent" (Study 1) or "clear-thinking, intelligent" (Study 2); the single-item indirect validities are mean item validities across all 119 items of the four indirect measures. The aggregated items direct validities are based on the 4-item direct composite measure. The aggregated items indirect validities are the mean full-scale validities across all four indirect measures.

"intelligent" and closely related items. The single-item indirect validities are mean item validities across all 119 items of the four indirect scales. Finally, the aggregated indirect validities are the mean full-scale validities across all four indirect measures.

In the case of the direct measures, aggregation boosts the validities from .20 and .23 to .24 and .26, respectively. This small improvement with aggregation was disappointing. Validity prophecy formulas,[10] for instance, would predict values of .26 and .30 for a measure comprising four items equivalent to our baseline single items. Apparently, the validities of our additional items did not parallel those of the original item (intelligent). Despite our best efforts to select conceptually similar items, aggregation provided only modest improvements in the validity of direct measures.

For the indirect scales, however, Table 5 reveals a dramatic effect for aggregation. Here the comparison is between the mean of the 119 item validities (.07 and .07) and the mean of the four full scale validities. Broken down scale by scale, aggregation raised the validities from .05 to .20, .20 to .24, .05 to .08, and from .11 to .20 for Ie, Smart, Intellect, and the BCL, respectively, in Sample 1, and from .07 to .34, .21 to .25, .09

10. The prediction formula when only one measure is lengthened is presented by Thorndike (1982, p. 153). An infinite number of equally good items would remove all unreliability to yield upper limits of .30 and .37 in the two samples.

to .13, and .06 to .13 in Sample 2. In short, all indirect scales benefited from aggregation.

A Closer Examination of the Indirect Measures Using an Empirical Approach

Table 6 presents the 10 best performing items from the indirect scales, as defined by consistently good validities. Every measure is represented in the top 10. The fact that the Ie scale contributes the largest number of representatives probably derives from the fact that it has the most items and therefore the greatest opportunity to capitalize on chance. It is noteworthy that the items with the highest validities are those related directly to mental ability: After all, the rationale behind the creation of these indirect measures was that indirect items should exceed the validity of the more blunt, direct items such as “intelligent” or “smart.”

All four Smart items performed well in both samples, and two made the top 10. Note that the top Intellect items are from the School Success subscale and that all concern reading. Also note that the top items of the BCL both come from the Verbal Ability subscale.

Table 6
The Top 10 Item Validities From Indirect Measures

Sample 1	Sample 2	Scale	Item Content
.27	.32	BCL	reads with high comprehension
.28	.26	BCL	has a good vocabulary
.27	.20	Intellect	As a child I was always reading
.21	.21	Ie/Intellect	I am quite a fast reader
.18	.23	Ie/Intellect	I read at least ten books a year
.20	.25	Smart	Is considered a very “brainy,” scholarly person
.21	.21	Ie	I was a slow learner in school
.20	.25	Smart	considered exceptionally or unusually intelligent
.20	.20	Ie	I seem to be at least as capable and smart as most others around me
.22	.17	Intellect	I would rather read than watch TV

Note. Values > .15 are significant at .01, while those > .20 are significant at .001, two-tailed. $N = (275, 265)$. “Ie” refers to Gough’s Intellectual efficiency scale; “Smart” refers to Trapnell’s scale; “Intellect” refers to Hogan’s scale; “BCL” refers to Sternberg’s Behavior Check List.

Summarizing the contents of Table 6, then, it appears that the top-performing items in the indirect scales were either (1) direct ability-related items, (2) indirect items about ability (i.e., Smart scale), or (3) items about reading behavior. If these 10 items are combined into a new “best items” composite scale, the correlation with IQ test is .34 in Sample 1 and .38 in Sample 2.[11]

A Closer Examination of the Indirect Measures Using a Theoretical Approach

Having discovered that the best performing items of the Sternberg, Hogan, and Gough scales were, in fact, the more direct and ability-related items, we decided to categorize the items of these indirect measures theoretically. Four a priori categories were considered: mental abilities, personality-related, behaviors, and interests. Two judges showed 95% agreement on classification.

For each category, the mean validities were calculated; they are presented here in Table 7. Mean item validities were .14 (Sample 1) and .12 (Sample 2) for the ability items. Means for next three categories (personality, interests, and behaviors) were in the .03 to .06 range—all substantially lower than the ability-related items.

Best of all, however, was the set of items addressing reading habits: “I read at least ten books a year” from the Ie and Intellect scales, “As a child

Table 7
Validities of Indirect Items: Means Within Content Category

Category of Items	# Items[a]	Sample 1	Sample 2
Ability-related	23	.14	.12
Personality-related	51	.05	.06
Interest-related	26	.03	.06
Behavior-related (Non-reading)	8	.05	.05
Reading	5	.19	.18

*Note. N*s = 275 (Sample 1) and 265 (Sample 2). Each entry is the mean of all item validities for each category. [a]Number of items refers to the number of item validities used in calculating the mean item validity.

11. These values are likely to be overestimates because of capitalization on chance. Unfortunately, cross-validation from one sample to the other is not feasible because the items were chosen on the basis of consistent performance across both samples.

I was always reading" from Hogan's Intellect composite scale, and "reads widely" and "Sets aside time for reading" from the BCL. In fact, these reading items showed exceptional mean validities of .19 (Sample 1) and .18 (Sample 2).

In sum, it appears that items directly related to mental ability and items about reading habits outperformed the other item content categories in predicting IQ test scores. The other categories of items show positive, but low, item validities. Nonetheless, they can be aggregated, as in the case of the Ie scale, to reach a reasonable level of validity.

DISCUSSION

We set out to evaluate whether IQ can be measured by proxy. That is, can the handy self-report format be used as a substitute for a cumbersome IQ test? Because the validity of a single self-rating of intelligence has not proved adequate, researchers have advocated a number of strategies for improving validity, namely, indirectness, aggregation, and prototype-weighting. Our results indicated that aggregated, direct measures were the most effective, but none could consistently exceed .30. Prototype weighting had minimal impact.

Performance of Direct Measures

We began by establishing the validity of our baseline, that is, a single face-valid self-rating of intelligence. In two large samples, the single-item showed validities of .20 and .23—values that are typical of previous studies. Some studies have reported higher validities but most of those were high-school or other samples with a wide range of talent. Competitive college samples, such as our own, suffer from a restricted range of ability, which limits potential validity values. In any case, our modest baseline values left plenty of room for improvement via aggregation.

The empirical benefits of aggregation were evaluated by pooling the single item with other intelligence-related items[12] in a composite direct measure. An improvement in validity was observed with the addition of the 3–4 items most synonymous/antonymous to "intelligent," namely, smart, clever, simple, and not gifted. The lack of improvement beyond

12. The items were selected by their conceptual similarity to "intelligent" (gifted, smart, clever, etc.).

four items suggests that further aggregation added more noise than valid variance. Our battery of items contained 13 items selected for relevance to mental ability. But few of these were able to capture that facet of self-perception linked to IQ.

Even the validities of the 4-item composites (.24, .26) do not match the values predicted by the validity prophecy formula (26, .31) based on projecting the validity of the single item "intelligent" (Thorndike, 1982, p. 152). Clearly that item plays a unique role, both conceptually and empirically.

Performance of Indirect Measures

Indirect measures promised to surpass the performance of direct measures by providing a less threatening, less evaluative assessment atmosphere. In terms of predicting IQ test scores, that promise was not fulfilled in our data. Given that the results for each indirect measure raised different issues, however, we will consider them one by one.

Gough Intellectual efficiency (Ie) scale. Recall that the Ie scale was constructed in a contrasted-groups fashion by selecting CPI items that correlated with an IQ test (Gough, 1953). Given that it was developed decades ago on California high-school students, its success in our contemporary Canadian college sample—that is, validities of .20 and .34—might be considered remarkable. Nonetheless, a full 52 items were required to achieve those full-scale validities because the mean item validity was low. Of course, True-False items are expected to show lower validities (but faster administration times) than corresponding Likert items.

Contrary to the original intent of the Ie, however, it was primarily the direct mental-ability items that correlated with IQ. The remaining items, concerned primarily with confidence and adjustment, were not as successful, although all items were originally selected because they correlated with IQ tests. Why the confidence and adjustment item validities did not replicate is difficult to say. It is understandable that distractability related to maladjustment could hamper performance on IQ tests independent of actual ability. And this handicap of maladjustment may be more true in high-school samples (where item-selection took place) than in the college samples where we chose to validate the items.

Perhaps the 4-decade gap in culture is somehow responsible. Even across a 10-year time span, Paulhus and Landolt (1994) found that the criteria for nominating intelligent people had changed noticeably whereas the criteria for the concept of intelligence had not changed. We suspect that, when criterion groups rather than rational methods are used to develop scales, items measuring temporary societal influences are more likely to intrude.

Sternberg's Behavior Check List. As a unit, the BCL showed only modest predictive efficacy—slightly higher if an item-weighting procedure was applied. One of the subscales, the Verbal Ability subscale, was effective. Sternberg et al. (1981) found the same pattern. Our detailed item analyses revealed that the high-validity items carrying the subscale were those concerning mental ability and items about reading habits.

Although the ability of the BCL to predict IQ was not impressive for a 41 Likert-item measure, we must call attention to its original purpose. Sternberg intended the BCL not as a *proxy* for IQ tests, but as a *supplementary measure*: It was designed to be administered along with an IQ test to tap components of intelligence that IQ tests were not capable of measuring (Sternberg, 1988, p. 239). From this perspective, high correlations with IQ tests should not be expected.

This supplementary role of the BCL is consistent with Sternberg's long-standing complaint that IQ tests measure only a limited part of lay conceptions of intelligence. Recently, we have followed up this notion in our work on "non-test intelligence" (Lysy & Paulhus, 1996). By partialing IQ and self-presentation out of self- and peer-ratings of intelligence, we formed a self-residual and a peer-residual to represent that part of intelligence that is "beyond IQ." We then correlated the residuals with a battery of personality and interest measures. The top correlates of the self-residual were self-rated conscientiousness and openness, self-esteem, the Intellectual efficiency scale, and the Smart scale, while the top correlates of the peer residual were peer-rated conscientiousness, openness, physical attractiveness, and athletic ability. The different correlates of self and peer suggest that "non-test intelligence" is largely a perceiver-dependent idiosyncrasy. There was, however, a small overlapping component indicating that self and others systematically misattribute intelligence to those who are conscientious and open. This component may be that facet of "true intelligence" that is not represented in IQ tests.

Hogan's Intellect composite. As a whole, the Intellect composite showed only a modest ability to predict IQ scores. Obscured in this overall figure, however, is the fact that the two subscales showed dramatically different validities. Recall that the Intellectance subscale was designed to capture an unconventional, creative conception of intellect whereas School Success was aimed at the more conventional goal-oriented conception of intellect. Our results support this distinction in that School Success was a distinctly better predictor of IQ with validities of .19 and .27 in our two samples. In fact, these are underestimates because we used only a 7-item version. The validity of a 21-item version, as predicted by the prophecy formula (Thorndike, 1982), would have been .22 and .32 in our two samples.

Trapnell's Smart scale. The newest instrument in the study, Trapnell's (1994) Smart scale, performed well. It was designed to reduce range-restriction in two ways: (1) by diminishing the desirability of claiming the item and (2) by shifting the implied locus of evaluation from self to others. As intended, the Smart scale did show a reduced range restriction: Subjects utilized almost the entire range of the 9-point scale—noticeably more than the range of Sternberg's BCH.

The Smart scale is certainly efficient, requiring only four items to match or even outperform the other indirect measures. It is now evident, however, that the success of the Smart scale did not derive from its indirect nature. Direct composites with four items of similar content (smart, clever, etc.) worked just as well as the Smart scale. Therefore its success was more likely a function of content rather than of Trapnell's strategic contextualizing of the items.

The Content of Predictors

This scale-by-scale analysis of successful items helped clarify the source of their success. Although the four inventories derived from four dramatically different domains, the successful items within each were almost entirely ability-related. Of course, the direct measures were designed to address ability directly. But in the case of indirect measures, it is certainly ironic that their most direct items work best. This finding has the added benefit of refuting a potential alternative explanation for our finding of lower validities for indirect measures than direct measures, namely, that the indirect measures were administered later in a separate test battery from the direct measures. But even those direct items included in the same battery as the indirect still outperformed the indirect items.

Out of all remaining item-content areas, the only one yielding consistently high validities was an interest in reading.[13] Why a lifelong enjoyment of reading is associated with achieving high scores on IQ tests is not clear. Many educational psychologists argue that reading behavior permanently boosts mental abilities (Rayner & Pollatsek, 1989) and is rightfully encouraged. Of course, other causal sequences are possible. High intelligence might make reading more enjoyable (Hogan & Hogan, 1992). Or third variables such as social class or openness to experience might nurture both (McCrae & Costa, 1985).

The Value of Aggregation and Weighting

Administration of the single item "intelligent" is certainly efficient given the practical costs of adding more items to a test battery. And, across all items administered in our studies, it was the most consistently valid. Addition of other direct items improved validities only up to five items. Beyond that, returns were marginal.[14] Apparently, the items linked to IQ scores have a limited semantic scope. The fact that the Ie scale benefited most from aggregation suggests that this strategy aids True-False more than Likert-item composites. It is understandable that dichotomous items, though potentially as valid as Likert items, require more aggregation because of lower item reliability.

Sternberg et al. (1981) reported substantial improvement in BCL validities via a correlation-with-prototype approach. As we showed in the Introduction, this approach is simply a form of weighting procedure, that is, counting certain items more than others in calculating total scale scores.[15] The traditional psychometric wisdom is that unit weighting of items selected from regression or factor analysis is preferable to any other weighting: That wisdom was not refuted by our data.

13. Interestingly, an instrument recently developed to predict school success contains the same two categories of predictors (Giddan, Jurs, Andberg, & Bunnell, 1996).
14. It seems that aggregation may pay off less in assessing intelligence than in assessing personality. However, our selection of intelligence items was not systematic enough to make such a strong claim here.
15. If the negatively keyed items have not yet been reversed, then the primary effect of weighting is simply to reverse these items.

A general-purpose inventory, such as Sternberg's BCL, however, may represent an interesting exception to that wisdom. Here, an eclectic set of items is to be used for a variety of purposes, in this case, all germane to intelligence (see Cornelius et al., 1989). For predicting IQ, our analyses showed that the Social Competence items should be given zero weights; in predicting some other criterion, a different set of items might be zero-weighted. In a sense, the weightings are used to "unload" the items that are irrelevant for the current purpose. Thus a heterogeneous set of items can be retained but weighted in different ways to predict different criteria.

According to this argument, the only instruments in our package with potential for improvement via weighting are the BCL and the Ie scale. Unfortunately, although we tried various weighting procedures, we achieved only minimal improvements. Nonetheless, in appropriate instruments, such weighting could prove useful.

Putting Our Results in Perspective

Are self-report measures useful as proxies for IQ tests in college samples? Our data suggest not. Given that the validity of an ideal proxy measure would be upwards of .55[16] in college samples, our validity cap of .30 is disappointing. We tried out the best available measures, as well as the most highly touted improvement strategies.

Limitations in the criterion? The criterion measure, the Wonderlic IQ test, does not appear to be at fault. Previous studies have shown sufficient construct validation in college populations (e.g., Kennedy et al., 1989; McKelvie, 1994; Paulhus & Morgan, 1997; Wonderlic, 1992). Rather than being inappropriate for measuring IQ in college samples, its lackluster performance here is directly attributable to its low standard deviation. It performed no better and no worse than any standard IQ test would have in this situation.

16. This estimate begins with the median correlation (.83) of the Wonderlic with other IQ tests in general populations (Wonderlic, 1992). Instead of 7.12, the general population SD, the mean SD of our two samples was only 4.6. Adjustment of the validity for this restriction in range (Cohen & Cohen, 1983) yields .55.

Contamination of self-reports? So why the poor correspondence between self-rated intelligence and IQ tests? As Sternberg (1998) has noted, correspondence is limited by the common tendency to base one's self-perceived intelligence on abilities different from those tapped by IQ tests. In addition, discordance is to be expected because of motivated as well as unmotivated ignorance (Paulhus, 1986). The motivated portion involves inflated self-perceptions due to narcissism or self-deception. Previous research shows that this component contributes even more than IQ scores—perhaps 20% of the reliable variance in self-perceived intelligence (Gabriel, Critelli, & Ee, 1994; Paulhus, Yik, & Lysy, 1996). This motivated component also includes idiosyncratic definitions of intelligence designed defensively to match the raters' own abilities and therefore ensure that they are intelligent (Dunning & Cohen, 1992). The unmotivated portion of ignorance may include a lack of interest, concern, or insight into such matters (Campbell & Lavallee, 1993).

Restriction of range? Finally, we must remind readers of the severe handicap placed on all the validities reported here. The restriction of range created by our use of college samples is likely to have diminished all validities as a function of the reduced variances (see Cohen & Cohen, 1983). Compared with the SD of 4.6 that we found for our IQ test, SDs of 7.1 are more typical of the general populations (Wonderlic, 1992). Adjusted for restriction of range, our baseline validities for the single item "intelligent" (.20–.23) would have reached .30–.35. Similarly, instead of our ceiling of .30 for aggregated instruments, we could have achieved values of .40–.45 in the normal population. The latter values appear strong enough to be useful in research, if not in diagnosing individuals.

Limitations of IQ tests? Self-reports of intelligence, we argue, should not be evaluated solely in terms of potential as proxies for IQ tests. Given that lay perceivers typically hold that there is more to intelligence than IQ, our participants may well have based their self-ratings on their creativity, their interpersonal sensitivity, their musical ability, or their self-insight—none of which are tapped by the Wonderlic. And we agree with the view of expert commentators such as Sternberg and Gardner that we must tie scientific conceptions of intelligence more closely to such lay conceptions.

Such arguments suggest an alternative criterion for evaluating self-reports of intelligence: the perceptions of knowledgeable peers. In support of this argument, we have elsewhere reported evidence that self-ratings predict peer-ratings of intelligence independent of IQ scores (Lysy & Paulhus, 1996). That is, some portion of observers' perceptions of intelligence is detectable by self and observers but not by IQ tests. Thus self-report measures of intelligence have validity beyond their use as proxy IQ measures. From this perspective, it would actually be surprising to find high correlations between IQ tests and perceptions of intelligence.

Some Promising Avenues

We see several potential avenues for clarifying the links between test performance and self-perceptions of intelligence. First is the development of new intelligence tests to encompass more of everyday conceptions of "intelligence." To the extent that test content corresponds to everyday conceptions, then associations should be higher. Wagner and Sternberg (1986) have pursued this avenue by developing objective measures of practical intelligence. Salovey and Mayer's (1990) "emotional intelligence" is another measure that shifts the conceptual borders of intelligence toward everyday conceptions.

A second avenue for future research is clarifying and perhaps improving the other side of the relationship, namely, the self-perceptions. What cues are people using to judge their intelligence? The lens model is proving profitable in specifying proximal cues, that is, objective behaviors that trigger attributions of intelligence (Reynolds & Gifford, 1996). We too are examining matches and mismatches between self- and peer-perceptions of intelligence and their correlates (Lysy & Paulhus, 1996). This research should help specify the missing content in current self-report measures.

In a third avenue of research, we have attempted to deal with the self-presentation typical of self-reported intelligence. The Overclaiming Questionnaire exploits a sophisticated methodology with great potential as a proxy IQ test (Paulhus, Bruce, & Lysy, 1996). Respondents are asked to rate their familiarity with a wide range of people, places, books, events, and so on. Because 20% of the items are fictitious, signal detection statistics can be used to separate accuracy from bias. In a series of college samples, the signal detection accuracy parameter (d') correlated .44–.50

with scores on an IQ test. Considering that these were college samples, the validities are quite promising.

Finally, we encourage further research on the indirect measures studied here. Their greatest potential asset has never been directly tested: they may actually outperform direct measures in ego-threatening administration conditions. Another issue worthy of study is whether the direct ability items work only when interspersed with a variety of other items.

CONCLUSIONS

The present paper constitutes the most comprehensive examination of self-reports of intelligence to date. We have organized the available measures into four categories of self-rated intelligence to investigate the effects of employing indirect versus direct measures, and the effects of aggregation, on predicting objectively scored intelligence. Administration of these measures to two large samples led us to a few key conclusions.

1. Both direct and indirect self-report measures of intelligence can reliably predict IQ scores. Because of the restricted range of abilities in competitive college samples, however, the validity limit appears to be .30.

2. Direct items about global mental abilities are more valid than indirect items. The one clear exception is the high validities of indirect items referring to enjoyment/frequency of reading.

3. Aggregation of global ability items is beneficial up to a point. With the exception of reading items, aggregation doesn't appear to help beyond 4–5 core items referring directly to close synonyms/antonyms of intelligence (e.g., smart, clever, simple, not gifted).

4. Prototype weighting is helpful only for excluding ineffective items in an inventory.

5. Among available measures, the most effective predictors of IQ scores were Gough's Intellectual efficiency and Trapnell's Smart scale. Equally effective were Hogan's School Success scale and Sternberg's Verbal Ability scale.

As a whole, our verdict is pessimistic about the utility of self-reports as proxy measures of IQ in college samples. Our verdict is more optimistic about their utility for assessing intelligence as a broader concept, particularly in the general population. Either way, researchers who require some proxy IQ test for their research should benefit from the guidelines we have provided here.

REFERENCES

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, **121**, 219–245.

Aiken, L. R. (1996). *Assessment of intellectual functioning* (2d ed.). New York: Plenum Press.

Bailey, R. C., & Hatch, V. (1979). Interpersonal perceptions of intelligence in late childhood and early adolescent friendships. *Journal of Genetic Psychology*, **135**, 109–114.

Bailey, R. C., & Mettetal, G. W. (1977). Perceived intelligence in married partners. *Social Behavior and Personality*, **5**, 137–141.

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Springfield, IL: Charles C. Thomas.

Block, J. (1971). *Lives through time*. Berkeley, CA: Bancroft Books.

Borkenau, P. (1993). To predict some of the people more of the time: Individual traits and the prediction of behavior. In K. H. Craik, R. Hogan, & R. N. Wolfe (Eds.), *Fifty years of personality psychology: Perspectives on individual differences* (pp. 237–249). New York: Plenum Press.

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology*, **65**, 546–553.

Campbell, J. D., & Lavallee, L. F. (1993). Who am I? The role of self-concept confusion in understanding the behavior of people with low self-esteem. In R. F. Baumeister (Ed.), *Self-esteem: The puzzle of low self-regard* (pp. 3–20). New York: Plenum Books.

Cohen, J., & Cohen, P. (1983). *Applied multivariate regression/correlation for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cornelius, S. W., Kenny, S., & Caspi, A. (1989). Academic and everyday intelligence in adulthood: Conceptions of self and ability tests. In J. D. Sinnott (Ed.), *Everyday problem-solving: Theory and applications* (pp. 191–210). New York: Praeger.

DeNisi, A. S., & Shaw, J. B. (1977). Investigation of the uses of self-reports of abilities. *Journal of Applied Psychology*, **62**, 641–644.

Dodrill, C. B. (1983). Long term reliability of the Wonderlic Personnel Test. *Journal of Consulting and Clinical Psychology*, **51**, 316–317.

Dunning, D., & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology*, **63**, 341–355.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, **51**, 360–392.

Gabriel, M. T., Critelli, J. W., & Ee, J. S. (1994). Narcissistic illusions in self-evaluations of intelligence and attractiveness. *Journal of Personality*, **62**, 144–155.

Giddan, N. S., Jurs, S. G., Andberg, M., & Bunnell, P. (1996). Noncognitive long-term prediction of college grades by the Academic Performance Scale. *Assessment*, **3**, 91–98.

Gough, H. G. (1953). A nonintellectual intelligence test. *Journal of Consulting Psychology*, **17**, 242–246.

Gough, H. G. (1996). *California Psychological Inventory* (3rd. ed.). Palo Alto, CA: Consulting Psychologists.

Gulliksen, H. (1967). *Theory of mental tests*. New York: Wiley.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.

Hunter, J. E. (1989). *The Wonderlic Personnel Test as a predictor of training success and job performance*. Technical report, Department of Psychology, Michigan State University.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domain, observability, evaluativeness, and the unique properties of the self. *Journal of Personality*, **61**, 521–551.

Kennedy, R. S., Baltzley, D. R., Turnage, J. J., & Jones, M. B. (1989). Factor analysis and predictive validity of microcomputer-based tests. *Perceptual & Motor Skills*, **69**, 1059–1074.

Lysy, D. C., & Paulhus, D. L. (1996, August). *Beyond IQ: The search for non-test intelligence*. Paper presented at the meeting of the American Psychological Association, Toronto.

McCrae, R. R. (1990). Traits and trait names: How well is openness represented in natural languages? *European Journal of Personality*, **4**, 119–129.

McCrae, R. R., & Costa, P. T. (1985). Updating Norman's "Adequate Taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology*, **49**, 710–721.

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

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

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaire* (pp. 143–165). New York: Springer-Verlag.

Paulhus, D. L., Bruce, M. N., & Lysy, D. C. (1996). *The Over-Claiming Questionnaire (OCQ)*. Unpublished instrument, University of British Columbia.

Paulhus, D. L., & Landolt, M. (1994, June). *Differential processes in accessing concepts of intelligence*. Paper presented at meeting of Canadian Psychological Association, Penticton, Canada.

Paulhus, D. L., & Martin, C. L. (1987). The structure of personality capabilities. *Journal of Personality and Social Psychology*, **52**, 354–365.

Paulhus, D. L., & Morgan, K. L. (1997). Determinants of perceived intelligence in leaderless groups: The dynamic effects of shyness and familiarity. *Journal of Personality and Social Psychology*, **72**, 99–107.

Paulhus, D. L., Yik, M.S.M, & Lysy, D. C. (1996, August). *Self- and peer-ratings of intelligence: Accuracy or self-presentation?* Paper presented at the meeting of the American Psychological Association, Toronto.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice-Hall.

Reilly, J. & Mulhern, G. (1995). Gender differences in self-estimated IQ: The need for care in interpreting group data. *Personality and Individual Differences,* **18**, 189–192.

Reynolds, D., & Gifford, R. (1996). *Measured and judged intelligence: A Brunswik lens analysis of verbal and nonverbal cues*. Manuscript in progress, University of Victoria.

Salovey, P., & Mayer, J. D. (1989–1990). Emotional intelligence. *Imagination, Cognition, and Personality,* **9**, 185–211.

Schmidt, F. L. (1985). Review of the Wonderlic Personnel Test. In J.V. Mitchell (Ed.), *Ninth mental measurements yearbook* (pp. 1755–1757). Lincoln, NE: Buros Institute of Mental Measurement.

Schoenfeldt, L. F. (1985). Review of Wonderlic Personnel Test. In J. V. Mitchell (Ed.), *Ninth mental measurements yearbook* (pp.1757–1758). Lincoln, NE: Buros Institute of Mental Measurement.

Sternberg, R. J. (1988). *The triarchic mind: A new theory of human intelligence*. New York: Penguin Books.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology*, **41**, 37–55.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton-Mifflin.

Trapnell, P. D. (1994). Openness versus intellect: A lexical left turn. *European Journal of Personality*, **8**, 273–290.

Trapnell, P. D., & Scratchley, L. (1996). *Predictors of intellectual performance*. Unpublished data, University of British Columbia.

Wagner, R. K, & Sternberg, R. J. (1986). Tacit knowledge and intelligence in the everyday world. In R. J. Sternberg & R. K Wagner (Eds.), *Practical intelligence: Nature and origins of competence in the everyday world* (pp. 51–83). New York: Cambridge University Press.

Welsh, G. S. (1975). *Creativity and intelligence: A personality approach*. Chapel Hill, NC: Institute for Research in Social Science.

Wonderlic, E. F. (1992). *Wonderlic Personnel Test and scholastic level exam user's manual*. Libertyville, IL: Wonderlic Personnel Test, Inc.