Application of the Overclaiming Technique to Scholastic Assessment

Delroy L. Paulhus and Patrick J. Dubois

Educational and Psychological Measurement 2014 74: 975 originally published online 3 June 2014
DOI: 10.1177/0013164414536184

Article

Application of the Overclaiming Technique to Scholastic Assessment

Educational and Psychological Measurement
2014, Vol. 74(6) 975–990

DOI: 10.1177/0013164414536184
epm.sagepub.com

Delroy L. Paulhus[1] and Patrick J. Dubois[1]

Abstract

The overclaiming technique is a novel assessment procedure that uses signal detection analysis to generate indices of knowledge accuracy (OC-accuracy) and self-enhancement (OC-bias). The technique has previously shown robustness over varied knowledge domains as well as low reactivity across administration contexts. Here we compared the OC-accuracy index with multiple choice (MC) and short answer (SA) tests in assessing knowledge of introductory psychology topics in a sample of 108 undergraduates. Results indicated that OC-accuracy was (a) comparable to MC and SA in predicting overall course grades and (b) superior to SA tests in reliability achieved per unit administration time. By including the OC-bias index, the overclaiming method also adds a unique element to scholastic testing, namely, a measure of knowledge self-enhancement. The latter index was a *negative* predictor of overall course grade, suggesting a narcissistic self-destructiveness. Because the self-enhancement index adds no extra administration time to the knowledge measure, the overclaiming approach provides a more rich and efficient information source compared with traditional methods of scholastic assessment.

Keywords

knowledge, self-presentation, signal detection, test validity

[1]University of British Columbia, Vancouver, British Columbia, Canada

Corresponding Author:
Delroy L. Paulhus, Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4.
Email: dpaulhus@psych.ubc.ca

Introduction

Scholastic measures are often divided into two broad categories: Selected-Response (SR) and Constructed-Response (CR). SR measures such as multiple-choice items offer the convenience of objective scoring and automated evaluation. But they are vulnerable to guessing and strategic test-wiseness—skills boosting academic achievement but of dubious value in future occupations. CR measures (e.g., essay or short answer) can provide more ecologically valid assessments, but are less reliable and less convenient because onerous, subjective scoring procedures are necessary (see review by Downing, 2009).

In his review of the extensive literature comparing CR and SR items in scholastic assessment, Hogan (2013) concluded that they ''do not appear to be measuring different traits, abilities, or degrees of knowledge.''(p. 288) Despite that empirical evidence for convergence, many educators continue to view CR and SR methods as qualitatively different (Bleske-Rechek, Zeug, & Webb, 2007). Hogan did recommend further study on the possibility that the two approaches have unique individual difference confounds. Because they are the most common representatives of each category, our present focus is on multiple-choice and short-answer methods.

Surprisingly, one metric for comparing different assessment methods, namely, time efficiency, is singularly absent from the research literature. Time efficiency refers to psychometric performance achieved per unit administration time. As Parkes (2009) wrote, ''...reliability comes through length, and length comes at additional costs such as more testing time...'' (p.112). Others have alluded to this issue in touting the performance of brief measures that retain high reliability (Hopkins, Hakstian, & Hopkins, 1973; Jeyakumar, Warriner, Raval, & Ahmad, 2004).

Multiple-Choice Method

Although easily and objectively scored, MC tests do not measure students' ability to generate answers. They can also mask personality confounds based on test wiseness and willingness to guess. Because most MC formats provide only a small set of options to choose from,[1] guessing is a potentially rewarding strategy. So what to do about guessing? Rowley and Traub (1977) summarize the dilemma concisely:

> If one encourages students to answer all questions, whether they know the answer or not, a source of random variance is introduced (Lord, 1963) which decreases both reliability and validity; on the other hand if one attempts to discourage students from guessing, it is apparent that some students will comply to a greater extent than others, causing the test results to be contaminated by personality factors which the test was not intended to measure. (pp.16-17)

A variety of correction-for-guessing (CFG) techniques have been explored: They range from the simple and popular formula scoring (described in Lord, 1975), which ignores partial knowledge, to the sophisticated simulations of Espinosa and Gardeazabal (2010), which attempt to model partial knowledge. However, any

application of CFG introduces new confounds of how best to give instructions and how students react to novel, and more complicated, test-taking strategies (Baradaran, Ahanghari, & Semiari, 2009). Such stylistic differences are difficult to assess empirically.

Another stylistic difference—test wiseness—has also proven difficult to assess. According to Millman, Bishop, and Ebel (1965, p. 707), test wiseness is the ''subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score.'' In principle, it is independent of the examinee's knowledge of the subject matter being assessed. In his recent review, Cohen (2007) concluded that verbal reporting during the testing session is the only useful method for assessing test wiseness. As far as we know, there are no direct methods for measuring such stylistic differences without interfering with the test scores.

Short Answer Method

The use of short answer items has well-known pros and cons. Undoubtedly, the method bears a closer resemblance to everyday tasks in later life. However, the subjectivity of scoring handicaps interrater reliability. As with MC testing, higher reliabilities can be achieved by increasing the number of items and their content overlap. Trained scorers with a clear key can achieve high interrater reliabilities.

Small SA advantages have been found for female, high-ability, and low-anxiety students (Hogan, 2013). The most obvious confound is verbal ability. Verbally able students will be better able to communicate their knowledge of any subject. Unless the cause is obvious (e.g., ESL), educators are reluctant to correct scholastic achievement for natural abilities. To assess its role in our research, we included a validated measure of verbal ability.

Response Style Summary

In contrast to fundamental personality constructs such as the Big Five or self-esteem (Hair & Graziano, 2010; Poropat, 2009), response styles are habitual ways of responding during testing sessions (Cronbach, 1946). Dealing with response styles such as guessing and test wiseness in scholastic assessment has parallels with the venerable debate over self-report assessment methods, that is, how to separate response-style from valid content (Holden & Passey, 2010; Paulhus, 1991). Although correction methods continue to be addressed in some educational domains (e.g., Huibregtse, Admiraal, & Meara, 2002; Pellicer-Sanchez & Schmitt, 2012), response styles have not been investigated as constructs in their own right. To permit such analyses, we propose the application of the overclaiming technique, which does provide separate indices of knowledge and style. Hence, their independent contributions to scholastic success can then be evaluated.

Overclaiming Technique

The Overclaiming Technique (OCT) has its origins in personality and ability measurement (for a review, see Paulhus, 2011). Applying the method to scholastic measurement promises a number of advantages over both SR and CR measures. OCT assesses both knowledge and response style by asking respondents to rate their familiarity with a set of content-relevant items: Some of them exist (''reals'') and some of which do not (''foils''). The proportion of reals claimed (''hit rate'' = H) and the proportion of foils claimed (''false alarm rate'' = F) are analyzed with signal-detection formulas (Swets, 1964) to yield two indices: actual knowledge (OC-accuracy) and overclaiming (OC-bias).[2]

The use of foils to correct familiarity claims has a long history as far back as Raubenheimer (1925). Applications include assessment of vocabulary (Meara & Buxton, 1987) and literacy (Stanovich & Cunningham, 1992). In general, that research confirmed that correcting the accuracy scores improved predictions of relevant outcomes. However, only one vocabulary study considered the value of creating a separate measure of bias (Ziegler, Kemper, & Rammstedt, 2013).

The method of distinguishing and measuring both indices—the Overclaiming Technique—was first introduced by Paulhus and Bruce (1990). Since then OCT has undergone a thorough validation process and been applied successfully to a variety of measurement domains. The original overclaiming questionnaire was academic in content: It comprised 15 items in each of 10 categories (e.g., science, law, philosophy, history, literature, language). A series of studies demonstrated that the accuracy index predicted verbal IQ scores in the .40 to .60 range (Paulhus & Harms, 2004). The bias index correlated moderately (.25-.38) with trait self-enhancement measures such as narcissism and the Self-Deceptive Enhancement scale (Paulhus, Harms, Bruce, & Lysy, 2003). When the survey items concerned lay topics such as sports, music, and films, the bias link was more nuanced. Correlations with narcissism were significant only for topics that the respondent valued. Interestingly, the accuracy scores predicted IQ for virtually any of the lay topics (Lysy & Paulhus, 1998).

Several advantages of the OCT have already been demonstrated. For example, the validity of accuracy scores is sustained under fake-good conditions, even though bias scores increase substantially (Paulhus & Harms, 2004). The validity of the bias index, on the other hand, is sustained under warning conditions, where the presence of foils is made salient (Paulhus et al., 2003).

Some work has begun on clarifying the processes underlying overclaiming (Nathanson, Williams, & Paulhus, 2002). We wondered, for example, what would make individuals claim knowledge of nonexistent foils under anonymous circumstances? Preliminary evidence suggests both motivational and cognitive elements at work. Independent of narcissism scores, bias scores tend to correlate with a global memory bias (Nathanson et al., 2002).

A recent practical application is to the field of marketing surveys (Paulhus, 2011). In the traditional approach to indexing product familiarity, a survey with a list of product names is administered. But foils are rarely included. We developed a

marketing survey comprising 12 product categories (e.g., wine, cars, fashion designers, cosmetics brands). Results indicated that the validity of the accuracy index held up even when the bias index was inflated by instructions to fake good.

Validity evidence for the OCT emanates from laboratories outside our own. One recent application is to measurement of vocabulary (Ziegler et al., 2013). Others include the indirect measurement of both agentic and communal narcissism (Gebauer, Sedikides, Verplanken, & Maio, 2013). Given that previous measures addressed only agentic narcissism, Gebauer and colleagues reckoned that OCT could address communal axis simply by including such foils as ''the UN Act Against Childism.'' In the organizational behavior literature, the bias index has shown its worth in capturing faking in job applicants (Bing, Kluemper, Davison, Taylor, & Novicevic, 2011).

In short, the overclaiming technique has proven itself as an efficient and robust method for indexing self-enhancement. Because the apparent purpose is a survey of personal familiarities, the method minimizes reactivity. This property makes OCT robust across a variety of administration conditions.

In the present application to scholastic assessment, we hope to demonstrate two things: (a) the OC-accuracy index has psychometric characteristics comparable to traditional knowledge measures and (b) the OC-bias index captures a maladaptive individual difference variable representing knowledge exaggeration, a form of self-enhancement.

Overview

We compared the three knowledge assessment methods—multiple choice (MC), short answer (SA), overclaiming (OC-accuracy)—with respect to their reliability and validity for predicting final grades. As well as their absolute performance, we compared the methods with respect to time efficiency, that is, psychometric performance per unit administration time. After equating the three methods with respect to an administration time of 10 minutes, we compared their alpha reliabilities. Finally, we evaluated the association of all three knowledge measures as well as self-enhancement (OC-bias) with final grades.

Method

Participants and Procedure

Participants were 108 undergraduate students enrolled in one of two introductory psychology courses. They were recruited via the departmental subject pool and given an extra half mark for participating. The overall sample was 59% female with a mean age of 20.1 years. The huge majority (92%) were full-time students in bachelor programs. Their knowledge of psychology was tested in a single inventory with three sections counterbalanced for order: OC, MC, and SA. Students completed the test in

timed, supervised sessions, but at their own pace. The sessions were held roughly 2 weeks before the end of the course.

Materials

All psychology knowledge items were developed by the researchers with the assistance of various sources (e.g., study guides, textbooks). We ensured that the item topics (but not the exact questions) were identical across the three methods.

The overclaiming section consisted of 64 real items and 16 foils. The foils were created to appear plausible, though nonexistent. Similar to the format in Appendix A, participants rated their familiarity with each item on a scale ranging from 1 (*never heard of it*) to 5 (*extremely familiar*). A variety of signal detection formulas (all based on calculating H and F) yielded similar results. For simplicity, we only report results with the so-called common-sense indices (see Appendix B):

$$\text{OC-accuracy} = \text{H} - \text{F}$$
$$\text{OC-bias} = (\text{H} + \text{F})/2$$

The 20 MC items included five options each. For each of the 12 SA items, students were asked to write at least three sentences in their answer. Marks ranging from 0 to 3 points per item were assigned by two independent graduate student raters.

Verbal Ability. Participants also completed the 50-item UBC Word test (Nathanson & Paulhus, 2007). Items entailed a single-word stem (e.g., carnal) and four options (e.g., verbal, physical, artistic, soluble): Participants are asked to select the most appropriate synonym. Although they are not advised in advance, time is limited to 8 minutes. Test scores show have been validated with a high concurrent validity (disattenuated correlation = .66) against the verbal items on the Wonderlic Personnel Test (Nathanson & Paulhus, 2007).

Criterion Measure. We used participants' final grade in introductory psychology as the criterion for scholastic achievement. These overall grades included an aggregate of several exams, all of which used a combination of MC and SA formats. They contained no OCT items. There was no direct overlap in the items used in this study and the ones used for course evaluations.

Results

Table 1 provides the descriptive statistics for all study variables. We tested for the possibility of gender differences but found none. Therefore, we pooled across gender.

Table 2 displays the key statistics for each of the predictor variables. Note that alpha was highest for the Short Answer format—but so was the administration time. To ensure a fair comparison, the three reliability values were extrapolated to (an arbitrary length of) 10 minutes of testing time using the Spearman–Brown correction for

Table 1. Descriptive Statistics for All Study Variables.

Item format	No. of items	Mean	*SD*	Range possible
Short answer	12	12.9	5.7	0-25
Multiple-choice	20	12.4	2.5	0-20
OC-*Accuracy*	80	0.20	0.09	0-1.00
OC-*Bias*	80	0.23	0.10	0-1.00
Final exam	80	75.5	9.9	0-100

Note. $N = 108$. In principle, accuracies can range from −1.00 to +1.00, but no negative values were found in this sample.

Table 2. Mean Administration Times and Reliability Efficiencies for Three Knowledge Measures.

Item format	Alpha reliability	Administration time (minutes)	Reliability efficiency
Short answer	.72	24.2	.52
Multiple choice	.54	5.4	.68
Overclaiming	.48	3.8	.71

Note. Reliability efficiencies are raw alphas adjusted to a common administration time of 10 minutes using the Spearman–Brown correction for test length.

test length. This procedure yielded reliability efficiency estimates of .52, .68, and .71 for the SA, MC, and OC methods, respectively.

Convergent Validity

Correlations among the three knowledge indices were substantial: OC-accuracy correlated with the MC and SA scores at .55 and .57, respectively, and MC correlated with SA at .63 (all $p < .01$, two-tailed). None of these values differed significantly from each other (all Z scores < 1.10).

Predictive Validity

Table 3 presents the correlations of MC, SA, and OC-accuracy with final grades. The raw correlations were .42, .39, and .37, respectively (all p values $< .01$). None of these values differed significantly from each other. Nor did they change when verbal ability was partialed out. After equating the methods for administration time, the validity efficiency estimates were .57, .21, and .61, respectively. Although OC and MC values did not differ, both were both significantly higher than SA (both p values $< .01$). By contrast, OC-bias showed a significant *negative* association ($r = -.18$) with overall grades ($p < .05$).

Table 3. Correlations of Knowledge Tests With Verbal Ability and Final Grade.

		Predictive validity		
Item format	Verbal ability	Raw	Disattenuated	Time adjusted
Short answer	.53	.39**	.46	.21**
Multiple choice	.47	.42**	.57**	.57**
OC-*Accuracy*	.49	.37**	.53**	.61**
OC-*Bias*	.23	−.18*	−.26*	−.33

**$p < .01$. *$p < .05$ (both 2-tailed).

Table 4. Regression of Final Grade on Three Knowledge Measures and Self-Enhancement.

				Correlations	
Item format	Beta	*t*	Sig.	Raw	Partial
Short answer	0.20	1.65	.10	0.39	0.17
Multiple choice	0.25	2.14	.04	0.42	0.21
OC-*Accuracy*	0.16	1.48	.14	0.37	0.15
OC-*Bias*	−0.29	−3.24	<.01	−0.18	−0.31

Note. Significance values are based on 2-tailed tests. With all predictors combined, $R^2 = .29$.

Table 4 shows the results of regressing the final grades on all three knowledge measures plus OC-bias. The negative beta of OC-bias actually increases from .18 to .29, indicating a suppressor effect. Because the three knowledge members overlap, their betas appear lower than their values in a simple regression. To see the unique OC contributions, we ran a final regression with only those two variables as predictors. The results in Table 5 indicate strong betas for both OC-accuracy and OC-bias. Again the suppressor effect of adding bias to the equation is evident in the pattern of betas. In fact, the incremental increase in R^2 from .13 to .19 was significant, $F = 7.01$, $p < .01$. Note that the same suppressor pattern (with significant increase in R^2) was observed when OC-bias was added as a second predictor to simple regressions using short answer or multiple choice as the accuracy predictor. Hence, its contribution to scholastic assessment is not limited to use with OC-accuracy but has broader theoretical import.

Discussion

Assessing Knowledge

We have introduced the OCT as an alternative to standard methods of scholastic assessment. In contrast to other methods, OCT yields separate indices of knowledge

Table 5. Regression of Final Grade on OC-Accuracy and Bias.

				Correlations	
Item format	Beta	*t*	Sig.	Raw	Partial
OC-*Accuracy*	.41	4.42	<.001	.37	.41
OC-*Bias*	−.24	−2.65	<.01	−.18	−.26

Note. $N = 108$. $R^2 = .19$.

and self-enhancement, both of which performed well according to psychometric qualities of alpha reliability, time efficiency, and predictive validity. In a direct comparison predicting final grades, the knowledge accuracy score performed on par with multiple choice and better than short answer scores. These results held up after controlling for verbal ability, as measured by the UBC Word test. Hence, the predictive power of our three knowledge indices was not simply because they acted as proxies for intelligence.

We made our comparisons after equating the three methods for time efficiency, that is, psychometric performance per unit administration time. Educators seeking to assemble an efficient test should keep in mind the total test time rather than the number of test items. This difference is striking when considering the short response times to recognition items in OCT compared to MC items.

Self-Enhancement

With no added administration time, the OCT method also provided highly relevant stylistic information, namely, the OC-bias index of self-enhancement. The overconfidence one brings to ability assessment varies independently of actual ability and therefore adds unique information. As predicted, self-enhancement was a *negative* predictor of scholastic success. Individuals who exaggerated their knowledge ended up with poorer course grades. Thus, our results parallel the deleterious academic consequences found in previous work (Kim, Chiu, & Zhou, 2010; Robins & Beer, 2001). Students who assume superiority and feel entitled to special treatment may not put in the necessary effort and perform poorly: This phenomenon may be especially prevalent among (a) first-year students at competitive schools and (b) hard science students who expect that psychology is an easy subject.[3]

Regression analyses showed that the negative impact of self-enhancement was independent of all other predictors and added incremental variance. Finally, the fact that OC-bias acted as a suppressor variable for all three knowledge predictors, thereby releasing their full predictive power, is a persuasive argument for including self-enhancement measures in predicting scholastic success.

Current Limitations and Future Research

Item Coverage

Although the scholastic content in our tests was limited to psychology topics, we see no reason why OCT cannot be applied to other content domains. After all, it has demonstrated success on 15 nonacademic topics (Lysy & Paulhus, 1998). However, our current method for OCT item selection was less than systematic. Instead, it relies on the intuition and scholastic experience of the item creators. For that reason, we are currently exploring optimal methods of item generation and evaluation (Dubois & Paulhus, 2014).

Level Coverage

As a test of recognition memory, OCT may be limited to cognitive assessments below Bloom's application level. Although useful for evaluating recognition of famous mathematicians, it is hard to see how performance on mathematical procedures could be tested. Haladyna, Downing, and Rodriguez (2002) raised this point regarding multiple choice measures and OCT is equally vulnerable to that common criticism of SR formats.

Nonetheless, the overclaiming accuracy index was able to predict a broader criterion that included multiple formats of psychology knowledge. The effectiveness of such a low-level measure may reflect a ''thin-slice'' effect: That is, quick surface assessment of knowledge may suffice in some circumstances. In fact, OCT proved more efficient than other SR techniques. Its simplicity may be a boon, given that other item formats may involve multiple dimensions of complexity (Schwarz, 2007).

Until our results are replicated, however, it may be too early for OCT to be applied to high-stakes testing such as personnel selection. Within scholastic settings, it is also premature to apply this form of testing beyond informal assessments in lower level college courses. For example, it may be ideal for formative assessments such as pop quizzes: Pre-exam testing with multiple choice questions has already proven its worth (Glass & Sinha, 2013). The nonthreatening format and quick scoring give OCT an advantage over multiple choice for simple classroom exercises or self-assessments.

Criterion Variable

Our choice of overall grade score as our criterion variable had pros and cons. As a cumulative test, it seemed appropriate for capturing the broader construct of scholastic achievement. OC assessment took place shortly before the final exam and covered the same general topics.

However, a number of factors limited optimal prediction. In that sense, our statistical estimates are conservative. All methods were constrained by our use of grades in

one course. Full year GPAs (e.g., Noftle & Robins, 2007) or cumulative GPA (Hair & Graziano, 2010) are more reliable measures and should therefore yield stronger correlations in future research. The relative performance of our OC-accuracy measure was further constrained by the fact that the criterion involved only multiple-choice and short answer formats.

Self-Enhancement

At this point, the mechanisms underlying the deleterious effects of self-enhancement remain unclear. Our results cry out for more research on self-enhancement in scholastic settings. They certainly reinforce the notion that personality variables can play a role in scholastic achievement (Ackerman, Kanfer, & Beier, 2013; Poropat, 2009). But instead of viewing response styles as noise variables to be controlled, we directly assessed the stylistic tendency to overclaim knowledge and used it as an independent predictor. Note that further research is required to distinguish overclaiming from overconfidence (Stankov & Lee, 2008).

The in-class benefits of the self-enhancement index may be less obvious. Understanding student self-enhancement serves a broader theoretical purpose that warrants further investigation, that is, teaching students to calibrate their self-assessments (Halpern, 2003). Ideally, their confidence should match their actual knowledge. Consider that one traditional issue with short answer format is the variation in verbosity. Some students rattle on whereas others hold back to avoid making mistakes. Teaching the ability to gauge one's knowledge is a form of self-critical thinking that may prove valuable in future academic endeavors. Given our evidence about the deleterious effects of self-enhancement, knowledge of scores may also help instructors diagnose worrisome students—possible cheaters or the maladaptively overconfident.

Time Efficiency

This psychometric property goes beyond the more familiar reliability and validity. Given comparable levels of reliability and validity, two tests may still differ dramatically with respect to time efficiency. Admittedly, this property may be of more interest to researchers than classroom instructors. The latter care primarily about accurate evaluation of students' knowledge. Because the test time length is typically based on ensuring that slower students can finish, the overall test efficiency is irrelevant.

For researchers, however, time efficiency is often a critical advantage. Researchers typically seek to include as many measures as possible within a given time frame—all without compromising validity. Reducing the number of items also reduces subject exhaustion and/or alienation. The latter factors are known to undermine test validity (Furr & Bacharach, 2008).

Conclusions

This study is the first to support use of the overclaiming technique as a measure of scholastic achievement. Educators may now exploit its assets: robustness across contexts, low reactivity, ease of administration, and time efficiency. The ancillary information provided by the self-enhancement index comes at no extra cost in administration time. We encourage both researchers and educators to explore this promising method.

Appendix A

Signal Detection Theory

Signal detection theory (SDT; Swets, 1964) was developed to characterize the ability of a receiver system to accurately distinguish signals in a noisy background (e.g., planes on a radar screen). The approach distinguishes between the ability to recognize true signals (accuracy) and the tendency to claim recognition whether or not the signal is present (bias).

In our work, we apply SDT to claims of familiarity with item lists that contain both real items (*reals*) and nonexistent items (*foils*). Claiming to recognize a real item (a hit) can then be contrasted with claiming to recognize a foil (a false alarm). Conversely, the failure to claim a real item (a miss) can be compared to the failure to claim a foil (correct rejection).

Such data permit the calculation of H (proportion of hits) and F (proportion of false-alarms). Entering these two values into signal detection formulas produces measures of accuracy and bias: For each one, a broad variety of formulas is available (MacMillan & Creelman, 1991). In brief, bias indices assess the tendency to claim familiarity, whether or not the item is real, whereas accuracy indices assess discrimination of reals and foils. Popular measures of accuracy include d' and area under ROC curve. Common bias measures include beta and c.

Following Snodgrass and Corwin (1988), we prefer pairs of accuracy and bias measures that are simple to calculate and statistically independent. One convenient pair of measures is also the simplest: Accuracy = H − F and Bias = (H + F)/2. Paulhus and Harms (2004) called them *common-sense indices* whereas MacMillan and Creelman (1991) label them *difference scores* and *yes-rate*, respectively. Both are compatible with the High Threshold Model.

Signal detection theory seems particularly applicable to the scholastic context: Educators want to measure the strength of the ‘‘signal’’ of an individual’s knowledge, separate from ‘‘noise’’ factors, for example, guessing, overconfidence, test-wiseness, or other construct-irrelevant factors. Many manipulations in test delivery (difficult vs. easy; more or less time) tend to amplify or attenuate both signal and noise in parallel

rather than improving the signal-to-noise ratio (i.e., accuracy). Accuracy and bias tendencies vary across individuals but their contributions are automatically distinguished by use of signal detection formulas.

Appendix B

Sample Format

1. Sample Page from the Academic Overclaiming Questionnaire.

IF YOU ARE FAMILIAR WITH THE ITEM, PLEASE CHECK THE BOX.

Fine Arts	Respondent 1	Respondent 2
1. Mozart	√	√
2. A cappella	√	
3. The Pullman paintings*		
4. Art deco	√	√
5. Paul Gauguin	√	
6. Mona Lisa	√	√
7. La Neige Jaune*	√	
8. Mario Lanza	√	
9. Verdi	√	
10. Jan Vermeer	√	
11. Windermere Wild*	√	√
12. Grand Pooh Bah		
13. Botticelli	√	
14. Harpsichord	√	√
15. Dramatis personae	√	

Note. The three foils are marked with asterisks.

2. Sample Calculations of the Accuracy and Self-Enhancement Indices From Sample Responses.

	Respondent 1	Respondent 2
Hits (out of 12)	11	4
False alarms (out of 3)	2	1
Proportion of Hits (H)	(11/12) = .92	(4/12) = .33
Proportion of False Alarms (F)	(2/3) = .67	(1/3) = .33
Accuracy index: (H − F)	.25	.00
Self-enhancement index: (H + F)/2	.80	.33

Note. Alternatively, F can be used directly as an index of self-enhancement. If so, H must be partialed out.

Acknowledgments

We acknowledge assistance from Bryce Westlake and Craig Nathanson in collecting the data.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by a research grant to the senior author from the Social Sciences and Humanities Research Council of Canada: #410-1999-741.

Notes

1. Haladyna (1992) argued that use of more than three options is rarely beneficial.
2. For more details on signal detection, see Appendix A. An example of an item-set and scoring procedures are provided in Appendix B.
3. Note that this phenomenon is fraught with moderators, especially the underlying motive for self-enhancement (see Gramzow, Elliot, Asher, & McGregor, 2003).

References

Ackerman, P. L., Kanfer, R., & Beier, M. E. (2013). Trait complex, cognitive ability, domain knowledge predictors of baccalaureate success, STEM persistence, and gender differences. *Journal of Educational Psychology, 105*, 911-927.

Baradaran, A., Ahanghari, S., & Semiari, S. R. (2009). The impact of correction for guessing formula on MC and yes/no vocabulary tests' scores. *Journal of Applied Linguistics, 2*, 80-98.

Bing, M. N., Kluemper, D., Davison, H. K., Taylor, S., & Novicevic, M. (2011). Overclaiming as a measure of faking. *Organizational Behavior and Human Decision Processes, 116*, 148-162.

Bleske-Rechek, A., Zeug, N., & Webb, R. M. (2007). Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude. *Assessment & Evaluation in Higher Education, 32*, 89-105.

Cohen, A. D. (2007). The coming of age of research on test-taking strategies. *Language Assessment Quarterly, 3*, 307-331.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*, 475-494.

Downing, S. M. (2009). Written tests: Constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 149-184). New York, NY: Routledge.

Dubois, P., & Paulhus, D. L. (2014, February). *Optimal item selection for overclaiming questionnaires*. Paper presented at the meeting of the Society for Personality and Social Psychology, Austin, TX.

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology, 54*, 415-425.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

Gebauer, J. E., Sedikides, C., Verplanken, B., & Maio, G. R. (2013). Communal narcissism. *Journal of Personality and Social Psychology, 103*, 854-878.

Glass, A. L., & Sinha, N. (2013). Multiple choice questioning is an efficient instructional methodology that may be widely implemented in academic courses to improve exam performance. *Current Perspectives in Psychological Science, 22*, 471-477.

Gramzow, R. H., Elliot, A. J., Asher, E., & McGregor, H. A. (2003). Self-evaluation bias and academic performance: Some ways and some reasons why. *Journal of Research in Personality, 37*, 41-61.

Hair, E. C., & Graziano, W. G. (2010). Self-esteem, personality and achievement in high-school: A prospective longitudinal study in Texas. *Journal of Personality, 71*, 971-994.

Haladyna, T. M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5*, 73-88.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309-333.

Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking*. New York, NY: Psychology Press.

Hogan, T. (2013). Constructed-response approaches for classroom assessment. In J. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 275-293). Thousand Oaks, CA: Sage.

Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences, 49*, 446-450.

Hopkins, K. D., Hakstian, A. R., & Hopkins, B. R. (1973). Validity and reliability consequences of confidence weighting. *Educational and Psychological Measurement, 33*, 135-141.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing, 19*, 227-247.

Jeyakumar, S. L. E., Warriner, E. M., Raval, V. V., & Ahmad, S. A. (2004). Balancing the need for reliability and time efficiency: Short forms of the Wechsler Adult Intelligence Scale-III. *Educational and Psychological Measurement, 64*, 71-87.

Kim, Y, Chiu, C., & Zou, Z. (2010). Know thyself: Misperceptions of actual performance undermine achievement motivation, performance, and subjective well-being. *Journal of Personality and Social Psychology, 99*, 395-409.

Lord, F. M. (1963). Formula-scoring and validity. *Educational and Psychological Measurement, 23*, 663-672.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12*, 7-11.

Lysy, D., & Paulhus, D. L. (1998, August). *The Overclaiming Questionnaire: More construct validity*. Paper presented at the meeting of the American Psychological Association, San Francisco, CA.

MacMillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York, NY: Cambridge University Press.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*, 142-154.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707-726.

Nathanson, C., & Paulhus, D. L. (2007, June). *Validation of the UBC Word Test*. Paper presented at the meeting of the Canadian Psychological Association, Toronto, Canada.

Nathanson, C., Williams, K. M., & Paulhus, D.L. (2002, August). *Nature of overclaiming process*. Poster presented at the meeting of the American Psychological Association, Chicago.

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of SAT and GPA and SAT scores. *Journal of Personality and Social Psychology, 93*, 116-130.

Parkes, J. (2009). Reliability in classroom assessment. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 107-123). New York, NY: Routledge.

Paulhus, D. L. (1991). Measurement and control of response biases. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Paulhus, D. L. (2011). Overclaiming on personality questionnaires. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 151-164). New York, NY: Oxford University Press.

Paulhus, D. L., & Bruce, M. N. (1990, June). *Claiming more than we can know: The Overclaiming Questionnaire*. Paper presented at the meeting of the Canadian Psychological Association, Ottawa, Ontario, Canada.

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence, 32*, 297-314.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The overclaiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology, 84*, 890-904.

Pellicer-Sanchez, A., & Schmitt, N. (2012). Scoring yes-no vocabulary tests: Reaction time vs. non-word approaches. *Language Testing, 29*, 489-509.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*, 322-338.

Raubenheimer, A. S. (1925). An experimental study of some behavioral traits of the potentially delinquent boy. *Psychological Monographs, 159*, 1-107.

Robins, R. W., & Beer, J. S. (2001). Positive illusions about the self: Short-term benefits and long-term costs. *Journal of Personality and Social Psychology, 80*, 340-352.

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement, 14*, 15-22.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*, 277-287.

Snodgrass, J. G., & Corwin, J. (1988). The pragmatics of measuring recognition memory: Applications to amnesia and dementia. *Journal of Experimental Psychology: General, 117*, 34-50.

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Psychological Assessment, 100*, 961-976.

Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition, 20*, 51-68.

Swets, J. A. (1964). *Signal detection and recognition by human observers*. New York, NY: Wiley.

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The vocabulary and overclaiming test (VOC-T). *Journal of Individual Differences, 34*, 32-40.