British Journal of Psychology (2018)


The British Psychological Society

www.wileyonlinelibrary.com

# A moderate financial incentive can increase effort, but not intelligence test performance in adult volunteers

Gilles E. Gignac*

School of Psychology, University of Western Australia, Crawley, Western Australia, Australia

A positive correlation between self-reported test-taking motivation and intelligence test performance has been reported. Additionally, some financial incentive experimental evidence suggests that intelligence test performance can be improved, based on the provision of financial incentives. However, only a small percentage of the experimental research has been conducted with adults. Furthermore, virtually none of the intelligence experimental research has measured the impact of financial incentives on test-taking motivation. Consequently, we conducted an experiment with 99 adult volunteers who completed a battery of intelligence tests under two conditions: no financial incentive and financial incentive (counterbalanced). We also measured self-reported test-taking importance and effort at time 1 and time 2. The financial incentive was observed to impact test-taking effort statistically significantly. By contrast, no statistically significant effects were observed for the intelligence test performance scores. Finally, the intelligence test scores were found to correlate positively with both test-taking importance ($r_c$ = .28) and effort ($r_c$ = .37), although only effort correlated uniquely with intelligence (partial $r_c$ = .26). In conjunction with other empirical research, it is concluded that a financial incentive can increase test-taking effort. However, the potential effects on intelligence test performance in adult volunteers seem limited.

Intelligence, as a process, may be considered the cognitive capacity to learn new things and solve novel problems (Reeve & Bonaccio, 2011). However, a positive correlation ($r \approx .30$) between test-taking motivation and intelligence test performance has been reported (Cole, Bergin, & Whittaker, 2008; Fervaha *et al.*, 2014). Furthermore, there is experimental research which suggests that intelligence test scores can be increased with a financial incentive (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). Consequently, the validity of intelligence test scores has been suggested to be compromised (Kirkwood, 2015). However, the vast majority of the experimental incentive research in the area of intelligence test performance has been conducted with non-adult samples. Additionally, a substantial amount of financial incentive and task performance research in the area of industrial psychology suggests that only effort and simple processing task performance can be increased with a financial incentive (Bonner, Hastie, Sprinkle, & Young, 2000). Arguably, the extent to which intelligence test scores

*Correspondence should be addressed to Gilles E. Gignac, School of Psychology, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia (email: gilles.gignac@uwa.edu.au).

DOI:10.1111/bjop.12288

can be manipulated with a financial incentive in a sample of adults remains relatively unexamined. Given the positive correlation between test-taking motivation and intelligence test scores, it is arguably important to determine whether test-taking motivation has a causal influence on intelligence test performance in adults. Consequently, the primary purpose of this investigation was to determine whether a financial incentive can be shown experimentally to increase both test-taking motivation and adult intelligence test performance in a sample of adult volunteers.

### *Previous experimental research*

A substantial amount of experimental research has been conducted on the effects of financial incentives on cognitive ability test performance (Bonner *et al.*, 2000; Duckworth *et al.*, 2011). Some financial incentive and cognitive ability performance experiments involve the administration of cognitive ability tests to two groups on two occasions. At time 1, the participants complete the cognitive ability testing under typical, non-incentivized conditions. At time 2, half of the participants are offered the opportunity to earn or win a reward, based on their performance. Other studies use a purely between-subjects design. Often, researchers use a tournament scheme of remuneration, where the participants in the experimental group are given an opportunity to win a financial reward, based on achieving a test performance score in the top range (e.g., top five scores). Researchers also tend to use a piece-rate remuneration scheme, where all participants receive a small financial reward for each item solved correctly within a defined period of time.

Based on a meta-analysis of 46 samples, Duckworth *et al.* (2011) reported an *N*-weighted mean effect size of $d = .64$, in favour of the group that received a financial or material (gift, toy) incentive to perform on the intelligence tests. An effect size of .64 corresponds to 9.6 IQ points ($15 \times .64$), which may be considered substantial, in practical terms. Based on such results, the validity of intelligence test scores obtained in low-stakes settings has been questioned (Kirkwood, 2015; Richardson & Norgate, 2015). Correspondingly, the intelligence validity coefficients reported in the literature have been contended to be inflated (Duckworth *et al.*, 2011). However, one of the limitations associated with the Duckworth *et al.* (2011) meta-analysis is that only two of the 46 samples were based on adults. Additionally, there is reason to believe that the effects of a financial incentive on adult intelligence test performance may be circumscribed to simple, processing speed tasks.

For example, Dickstein and Ayers (1973) administered the Information subtest from the WAIS to 32 undergraduate (adult) female participants. Based on the Information subscale scores, the participants were randomized into non-incentive and incentive intelligence test performance conditions. During a second testing session, all of the randomized participants were administered the performance subtests from the WAIS (PIQ) and the Advanced Progressive Matrices. The participants in the incentive condition were informed that those who obtained the top five intelligence test scores would receive \$1 (i.e., ≈\$6 in 2017 terms). The incentive group was reported to have achieved a statistically significant greater mean PIQ, in comparison with the non-incentive group ($d = .71$, $p = .032$).

However, at the individual test level, Dickstein and Ayers (1973) reported a statistically significant effect in favour of the incentivized group for the Object Assembly test only. By contrast, no statistically significant effects were observed for more cognitively demanding tasks (e.g., Advanced Progressive Matrices). As the WAIS Object Assembly test provides

substantial completion time bonus points (Wechsler, 1955), it may be suggested that the manipulation of test-taking motivation in adult volunteers via financial rewards can only potentially impact relatively simple tasks with a processing speed element, in comparison with more complex tasks (e.g., fluid reasoning, working memory).

More recently, Borghans, Meijers, and Ter Weel (2008) offered a sample of 180 adult volunteers various levels of financial remuneration for each intelligence test item answered correctly across four conditions: €.00, €.10, €.40, and €1.00 (piece-rate scheme). For each condition, the intelligence test consisted of 10 items derived from a mixture of intelligence test items (e.g., two items from Raven's; two anagram items). Borghans *et al.* (2008) reported that the testees engaged in a statistically significantly greater amount of effort (as measured by time spent attempting to complete the items), in accordance with the amount of money offered to complete the items, successfully. However, no statistically significant differences were observed with respect to the intelligence test scores across the incentive conditions. Thus, the results of Borghans *et al.* (2008) suggest that test-taking motivation can be increased with a financial incentive. However, intelligence test performance appears to be much more resistant to the effects of incentives, at least in adult volunteers.

### *Theoretical considerations*

Whether intelligence test performance can be enhanced experimentally in adults remains an open question. Theoretically, it is important to ask: By what mechanisms might financial incentives increase intelligence test performance? Within the area of industrial psychology, Wright and Kacmar (1995) developed a model of financial incentives and task performance, which suggested that the potential effects of financial incentives operate through enhancements of self-efficacy and attractiveness, which in turn impact the development of goals. Based on a sample of 80 undergraduate volunteers, Wright and Kacmar (1995) found that different types of financial schemes affected the goals the participants set for themselves, as well as the attractiveness of the task (solving anagrams). However, the financial incentive (75 cents per correct solution) was not observed to affect task performance. Bonner *et al.* (2000) conducted a quantitative review of the effects of financial incentives and performance in adult samples across a variety of laboratory administered cognitive tasks. With respect to the most intelligence-related tasks (e.g., anagrams, arithmetic problems), Bonner *et al.* (2000) found that only four of the 24 relevant studies reported a statistically significant effect of financial incentive on test performance. Consequently, Bonner *et al.* (2000) concluded that as the task becomes increasingly more cognitively complex, the effects of financial incentives diminish. By cognitive complexity, Bonner *et al.* acknowledged Wood's (1986) theoretical work, which includes the concept of component complexity (i.e., the number of subtasks required to perform the task). In the area of differential psychology, intelligence test complexity is considered to be correlated positively with general intelligence (Jensen, 1998). Correspondingly, a simple information processing speed task (e.g., one parameter reaction time task) is considered less complex cognitively than a working memory task (Larson, Merritt, & Williams, 1988). Furthermore, negative correlations between general intelligence loadings and the degree to which the tests are susceptible to practice and training effects have been reported (Jensen, 1998; te Nijenhuis, Jongeneel-Grimen, & Kirkegaard, 2014; te Nijenhuis, van Vianen, & van der Flier, 2007).

A limitation with Wright and Kacmar's (1995) model is that it omits more direct antecedents of test performance. For example, much theoretical work specifies that test

performance is due, at least in part, to test-taking effort. For example, in their model of human information processing task performance, effort was defined by Humphreys and Revelle (1984) as the motivational state of being engaged in, or trying hard at, the completion of a task. Additionally, Eklöf's (2010) model of test performance included test-taking effort as a direct effect. Finally, the Expectancy-Value Model of achievement motivation (Wigfield & Eccles, 2000) suggests that test-taking performance is a function of performance expectancies and the value attributed to a successful outcome. According to Shepperd and Taylor's (1999) interpretation, the expectancy dimension of the Expectancy-Value Model represents the perception that performance is contingent upon effort. Correspondingly, people's beliefs vary from low to high expectancy. Thus, a testee with high effort expectancy believes that an increase in effort will translate into an increase in performance. By contrast, a testee with low effort expectancy believes that increases in effort will not yield an enhancement in performance. The value dimension of the Expectancy-Value Model represents the degree of importance attributed to the successful completion of a task. Although to some degree distinct, the expectancy and value dimensions of achievement motivation are expected to be related positively (Shepperd & Taylor, 1999).

### *Self-report measurement of test-taking motivation*

Sundre (1999; Thelk, Sundre, Horst, & Finney, 2009) developed a self-report measure of test-taking motivation essentially consistent with the Expectancy-Value Model of academic achievement. Consequently, although similar in nature to other self-report measures of test-taking motivation (e.g., Arvey, Strickland, Drauden, & Martin, 1990; Penk, Pöhlmann, & Roppelt, 2014), the Student Opinion Scale (SOS) was explicitly designed to measure two dimensions of test-taking motivation: Importance and Effort. Half of the Student Opinion Scale's ten items pertain to the degree to which the testee values the outcome of his/her test performance (e.g., 'I am not concerned about the scores I receive on these tests', negatively keyed). The other half of the items pertain to the degree to which the testee applied effort in completing the test items (e.g., 'I gave my best effort on these tests').

Several investigations have found that the two subscales, Importance and Effort, from the Student Opinion Scale correlate with cognitive ability test performance (mostly academic achievement) in the order of .25 to .30 (e.g., Thelk *et al.*, 2009; Wise & DeMars, 2005), which is similar to the magnitude of effects reported for other self-report test-taking motivation scales (Arvey *et al.*, 1990; Chan, Schmitt, DeShon, Clause, & Delbridge, 1997). As financial incentives can increase a goal's attractiveness (importance) and expectancies for success (effort; Locke & Latham, 1990), it may be hypothesized that a financial incentive for intelligence test performance would increase scores on both the Importance and Effort subscales from the Student Opinion Scale.

Finally, we note that there is some evidence to suggest that the association between test-taking importance and test performance may be mediated completely by test-taking effort. Specifically, based on a sample 1,005 undergraduates who completed academic achievement testing and a series of self-report items pertinent to test-taking motivation, Cole *et al.* (2008) found that the association between an unspecified measure of test-taking importance and test performance was mediated completely by test-taking effort. Although the investigation by Cole *et al.* (2008) was sound, it may be considered a limitation that the two scales used to measure importance and effort had an unequal number of items. Additionally, it is also possible that the two scales had unequal

reliabilities (they were not reported), which, if so, would have affected the magnitude of the reported effects (Fan, 2003).

In the light of the above, it would be valuable to determine whether any of the effects of financial incentives on intelligence test performance are mediated by corresponding increases in test-taking Importance and Effort. Additionally, as a secondary objective, it would be useful to replicate the effects reported in Cole *et al.* (2008) with a self-report measure defined by an equal number of items designed to measure test-taking importance and effort, as well as analyses not affected by differences in subscale score reliability.

### *Summary*

To date, the adult experimental incentive and intelligence test performance research in adults is both limited in number and has yielded mixed results. Some research suggests that the effect may be isolated for processing speed tasks, rather than more fluid abilities (Dickstein & Ayers, 1973). Additionally, other research suggests that only test-taking effort can be enhanced through a financial incentive. In the light of the above, the primary purpose of this investigation was to test, experimentally, whether a financial incentive can influence both test-taking motivation and intelligence test performance in a sample of adult volunteers. As a secondary purpose, we sought to determine whether both test-taking importance and effort were related to intelligence test performance uniquely.

## Method

### *Sample*

The sample consisted of 99 first-year undergraduate university students (63.6% female). Such a sample size was associated with statistical power of .74 to detect a medium effect size (partial $\eta^2 = .06$; $\alpha = .05$). The participants were recruited from a first-year undergraduate psychology research pool within a large, English-speaking university in Australia. The students participated in the research voluntarily for a small amount of extra course credit. Mean age was 19.94 ($SD = 4.62$; age range: 17–47 years). Although information on ethnicity was not obtained from the participants, the university student body is known to be populated from a primarily White European background. Participants were required to speak English as a first language to participate.

### *Measures*

#### *Test-taking motivation*

Test-taking motivation was measured with the Student Opinion Scale (SOS; Sundre, 1999; Thelk *et al.*, 2009). The SOS consists of 10 items (5-point Likert scale) designed to measure two dimensions of test-taking motivation: Importance (five items) and Effort (five items). In this sample, the coefficient alphas were as follows: time 1 Importance, $\alpha = .67$; time 2 Importance, $\alpha = .71$; time 1 Effort, $\alpha = .80$; and time 2 Effort, $\alpha = .80$.

#### *Intelligence*

Intelligence was measured with eight tests. However, only five of the tests (processing speed and working memory) were associated with true alternate forms. The remaining three split-form tests (fluid, visuo-spatial, and crystallized intelligence) were administered

to help measure psychometric *g* at time 1 and time 2. None of the measures correlated significantly with age in this sample.

*Processing speed* was measured with the four subtests (Numbers, Letters, Numbers/Letters, and Letters/Numbers) within the Connections battery (Salthouse *et al.*, 2000). These tests are very similar to the well-known Trails A/Trails B tests (Reitan, 1958). Each of the four Connections tests has an alternate form. The Numbers and Letters subtests were considered simple processing speed, while the Numbers/Letters and Letters/Numbers subtests were considered more complex processing speed, as they required shifting between numbers and letters. It is not possible to estimate internal consistency reliability from this sort of test.

*Working memory capacity* was measured with alternate forms of a slightly adapted version of the Letter–Number Sequence subtest from the WAIS-IV (Wechsler, 2008). The cognitive processes underlying Letter–Number Sequencing were considered more complex than the Connections tests. In this sample, the coefficient alphas across the two conditions were as follows: time 1 LNS, $\alpha = .80$; and time 2 LNS, $\alpha = .83$.

*Visuo-spatial intelligence* was measured with Peters *et al.*'s (1995) redrawn version of the Vandenberg and Kuse (1978) mental rotation test. As the mental rotation test consists of 24 items, two short-forms were created based on the odd and even items. In this sample, the coefficient alphas across the two conditions were as follows: time 1 MR $\alpha = .71$; and time 2 MR, $\alpha = .76$.

*Fluid intelligence* was measured with the Advanced Progressive Matrices (APM; Raven, 1998). As the APM consists of 36 items, two short-forms were created based on the odd and even items. In this sample, the coefficient alphas across the two testing times were as follows: time 1 APM, $\alpha = .61$; and time 2 APM, $\alpha = .70$.

Finally, crystallized intelligence was measured with the Advanced Vocabulary Test (AVT; Gignac, Shankaralingam, Walker, & Kilpatrick, 2016). As the AVT is a 21-item multiple-choice test, two short-forms were created based on the odd and even items (one item was left out). In this sample, the coefficient alphas across the two testing times were as follows: time 1 AVT, $\alpha = .34$; and time 2 AVT, $\alpha = .49$. Although the reliabilities were low, the scores from the AVT nonetheless contributed meaningfully to measurement of general intelligence (see Supporting information).

Two principal component analyses of the respective eight intelligence tests administered at time 1 and time 2 were conducted to estimate time 1 and time 2 psychometric *g* scores (see Supporting information for more details). The internal consistency reliabilities of the component scores (theta; Armor, 1973) were .70 and .72 for the time 1 and time 2 testing times, respectively.

### Procedure

After reading an information sheet, the participants signed an online consent form and completed basic demographic questions. All of the testing was completed during a single testing session (approximately 50 min in total), and all participants were tested individually. Additionally, all participants completed the intelligence testing twice, based on the alternate (and split) forms of the intelligence tests described above. Furthermore, all of the participants were given the opportunity to enter a draw for three chances to win $75, based on achieving an overall cognitive ability score within the top 10% of the sample. Due to ethical considerations, a control group was not feasible (i.e., all participants deserved the opportunity to win money). Consequently, half of the participants were given the opportunity to enter the draw to win $75 at time 1 and the

other half were given the opportunity to enter the draw to win $75 at time 2. The participants who were given the opportunity to win $75 at time 2 were not told about the opportunity, until after they completed the time 1 intelligence testing. The order of intelligence test administration was the same across both conditions: AVT, Mental Rotation, APM, Letter–Number Sequencing, Numbers, Letters, Numbers/Letters, and Letters/Numbers. The participants volunteered to participate in the study over the course of a university semester. We considered the possibility that there may be important differences in the nature of students who participate early in the semester versus late in the semester. Consequently, to help control for any possible 'semester effects', the participants were allocated to the experimental conditions in an interleaved fashion (i.e., participant 1: time 1, money; participant 2: time 1, no money; participant 3: time 1, money).

### *Data analysis*

To test the difference between intelligence test score means across the no-incentive (no chance to win money) and incentive (chance to win money) conditions, a series of $2 \times 2$ mixed-design ANOVAs were performed on the four Connections subtests and the Letter–Number Sequencing test. Mixed-design ANOVAs were not conducted on the remaining intelligence tests, as they did not exist in alternative forms. Time 1 or time 2 was the within-subjects factor. Furthermore, whether the participant received the opportunity to win the money at time 1 or time 2 was the between-subjects factor. To estimate the association between test-taking motivation (Importance, Effort) and psychometric *g*, a series of Pearson correlations were performed. The correlations were disattenuated for imperfect reliability via the classic disattenuation formula (Nunnally & Bernstein, 1994). Furthermore, the disattenuated correlations were tested for statistical significance with a procedure described by Bobko and Rieck (1980).

## Results

First, no outliers were identified based on the outlier labelling rule with a 3.0 multiplier (Hoaglin & Iglewicz, 1987). Furthermore, the data were considered sufficiently normal (skew $< 2.0$) for the purposes of parametric statistical analyses (Schmider, Ziegler, Danay, Beyer, & Bühner, 2010).

### *Experimental results*

As shown in Table 1, a statistically significant interaction was observed for the test-taking motivation Effort subscale (partial $\eta^2 = .06, p = .018$). Furthermore, as shown in Figure 1 (panel B), the pattern of means was partially consistent with the hypothesis that the participants would report greater test-taking Effort in the opportunity to win money condition (see Table 2 for means and standard deviations). Specifically, the effect was essentially due to the observation of a reduction in test-taking Effort at time 2 by the group that received the opportunity to win the money at time 1, $t(47) = 3.72, p = .001, g = .39$. By contrast, the group that received the opportunity to win money at time 2 maintained the same level of test-taking Effort across the two testing conditions, $t$ (48) = .16, $p = .874$, $g = .02$. The statistically significant interaction implies that the two Hedge's *g* estimates were statistically significantly different from each other (Jaccard, 1998). There

were no statistically significant effects associated with the test-taking motivation Importance subscale. Thus, the experimental manipulation did not impact the participants' view of the importance of the testing across the two conditions (money vs. no money; see Table S1 for means and standard deviations).

In contrast to the experimental manipulation of test-taking Effort, no statistically significant effects were observed to be consistent with improved intelligence test performance across the money/no-money conditions (see Table 1). However, there was some statistically significant evidence of a practice effect for two of the four processing speed tasks (Numbers and Letters/Numbers). As shown in Figure 1, on average, the participants across both groups improved their performance on the Numbers (partial $\eta^2 = .12$, $p = .001$) and Letters/Numbers (partial $\eta^2 = .12$, $p = .001$) processing speed tasks from time 1 to time 2. By contrast, on average, the participants from both groups performed less well on the time 2 administration of the more complex Letter–Number Sequencing working memory task, in comparison with the time 1 administration (partial $\eta^2 = .11$, $p = .001$; see Figure 1; see also Table S1 for means and standard deviations).

### *Non-experimental results*

As shown in Table 3, test-taking Importance ($r_c = .28$, $p = .026$) and test-taking Effort ($r_c = .37$, $p = .004$) at time 1 correlated positively and statistically significantly with psychometric *g* at time 1. For the time 1 data, the correlation between Effort and psychometric *g*, controlling for the effects of Importance, was statistically significant (partial $r_c = .26$, $p = .009$). By contrast, the correlation between Importance and psychometric *g*, controlling for the effects of Effort, was non-significant (partial $r_c = .08$, $p = .421$). Finally, test-taking motivation (Importance and Effort) at time 2 failed to correlate significantly with psychometric *g* at time 1 or time 2, although the correlations were in the positive direction. However, test-taking Importance and Effort at time 1 correlated positively with psychometric *g* at time 2 ($r = .29, p = .019$; $r_c = .29, p = .018$).

## Discussion

The possibility of winning $75 was observed to increase self-reported test-taking Effort. However, a statistically significant increase in intelligence test performance was not

**Table 1.** Mixed-design ANOVAs: Main and interaction effects

| | Main effect | | | Interaction | | |
|---|---|---|---|---|---|---|
| Dependent variable | *F* | *p* | $\eta^2$ | *F* | *p* | $\eta^2$ |
| Test-taking Importance | 0.43 | .514 | <.01 | 1.76 | .187 | .02 |
| Test-taking effort | 5.82 | .018 | .06 | 6.54 | .012 | .06 |
| Numbers | 13.43 | <.001 | .12 | 3.81 | .054 | .04 |
| Letters | 2.93 | .090 | .03 | 0.94 | .334 | .01 |
| Numbers/letters | 3.48 | .065 | .04 | 0.79 | .779 | <.01 |
| Letters/numbers | 12.82 | .001 | .12 | 0.16 | .692 | <.01 |
| Letter–number sequencing | 11.84 | .001 | .11 | 0.01 | .997 | <.01 |

*Note.* Total $N = 99$; main effect and interaction effect *F*-values *df* = 1 and 97; $\eta^2$ = partial eta-squared; assumption tests of equality of covariance matrices were satisfied across all analyses via Box's *M*.

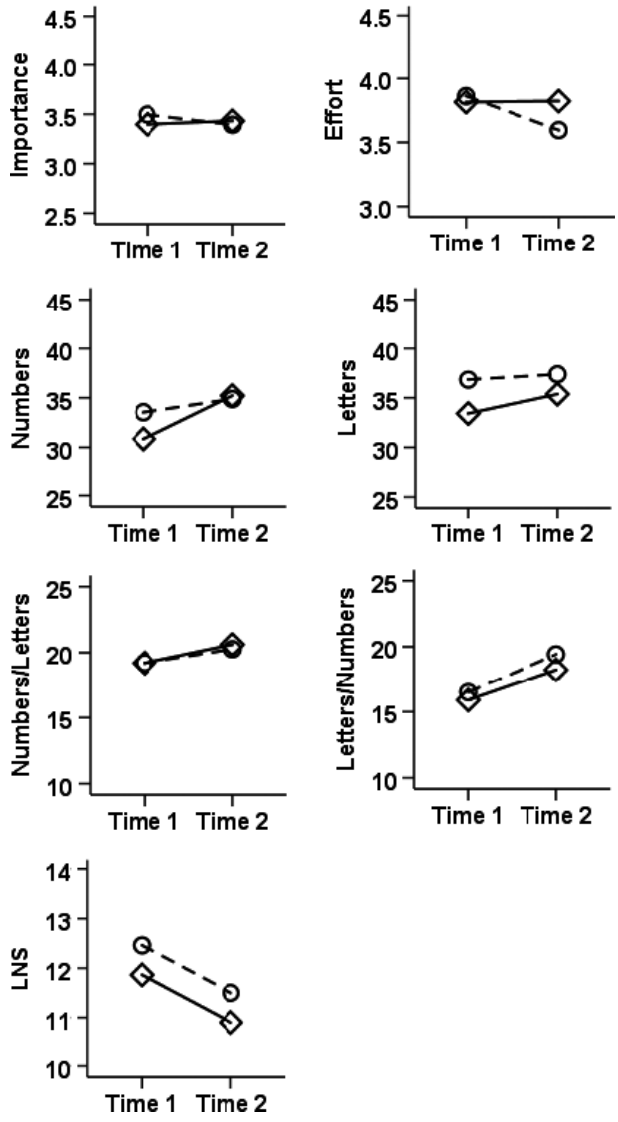


**Figure 1.** Time 1 and time 2 means associated with the test-taking motivation subscales (Importance and Effort) and the intelligence tests across the two experimental conditions: money at time 1 group (o- - -o) and money at time 2 group (◊—◊); LNS = Letter–Number Sequencing.

observed. Nevertheless, both test-taking Importance and Effort were found to correlate with intelligence test performance positively, although only Effort correlated with test-taking performance, uniquely.

### *Incentives and test-taking motivation*

The financial incentive in this investigation was observed to influence self-reported test-taking Effort. The magnitude of the effect may be considered moderate from a standardized effect size perspective (Richardson, 2011), although small in absolute terms

**Table 2.** Means and standard deviations associated with the time 1 and time 2 interaction effect analyses

| | Time 1 | | Time 2 | |
|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* |
| Time 1 money: No | | | | |
| Test-taking importance | 3.82 | 0.68 | 3.82 | 0.65 |
| Test-taking effort | 3.40 | 0.61 | 3.43 | 0.57 |
| Numbers | 30.84 | 9.18 | 35.22 | 6.51 |
| Letters | 33.41 | 8.62 | 35.37 | 9.96 |
| Numbers/letters | 19.20 | 6.59 | 20.61 | 6.82 |
| Letters/numbers | 15.90 | 7.31 | 18.24 | 8.16 |
| Letter–number sequencing | 11.86 | 2.91 | 10.90 | 3.47 |
| Time 1 money: Yes | | | | |
| Test-taking importance | 3.86 | 0.66 | 3.60 | 0.66 |
| Test-taking effort | 3.50 | 0.61 | 3.39 | 0.72 |
| Numbers | 33.54 | 8.00 | 34.88 | 7.47 |
| Letters | 36.83 | 8.21 | 37.38 | 7.34 |
| Numbers/letters | 19.21 | 6.54 | 20.25 | 6.68 |
| Letters/numbers | 16.50 | 6.95 | 19.42 | 8.44 |
| Letter–number sequencing | 12.46 | 3.09 | 11.50 | 3.05 |

*Note.* Participants who received the opportunity to win money at time 1, $N = 48$; participants who received the opportunity to win money at time 2, $N = 51$; see Table S1 for the main effect means.

**Table 3.** Descriptives and Pearson correlations between test-taking motivation and general intelligence test performance

| | 1. | 2. | 3. | 4. | 5. | 6. | *M* | *SD* | α |
|---|---|---|---|---|---|---|---|---|---|
| 1. Importance time 1 | – | **.59** | **.94** | **.45** | **.28** | **.29** | 3.44 | .61 | .67 |
| 2. Effort time 1 | **.43** | – | **.28** | **.83** | **.37** | **.29** | 3.84 | .67 | .80 |
| 3. Importance time 2 | **.65** | **.21** | – | **.46** | .16 | .18 | 3.41 | .64 | .71 |
| 4. Effort time 2 | **.33** | **.66** | **.35** | – | .07 | .11 | 3.71 | .66 | .80 |
| 5. *g* Time 1 | .19 | **.28** | .11 | .05 | – | **1.0** | .00 | 1.0 | .70 |
| 6. *g* Time 2 | .20 | **.22** | .13 | .08 | **.83** | – | .00 | 1.0 | .72 |

*Note.* $N = 99$; correlations in bold are statistically significant ($p < .05$); disattenuated correlations ($r_c$) are above the main diagonal; the disattenuated correlation between *g* Time 1 and *g* Time 2 was estimated in a latent variable model with several correlated residuals, to account for the disproportionately large amount of non-*g* shared variance between the corresponding subtests (e.g., Letter–Number Sequencing Time 1 and Time 2).

(see Figure 1, panel A). Borghans *et al.* (2008) also reported moderate effects on test-taking effort (i.e., time spent completing intelligence test items), based on a financial incentive (piece-rate scheme). Although a few more financial incentive investigations with intelligence tests would be useful, taken together with the broader area of financial incentives and laboratory-based task effort literature (Bonner & Sprinkle, 2002), it may be claimed with some confidence that financial incentives can, on average, increase the amount of effort testees apply to complete cognitive ability tasks.

By contrast, the experiment failed to produce a statistically significant effect on Importance, which suggests that, on average, the participants did not view their

performance on the intelligence tests as any more important, when given the opportunity to win the money. To our knowledge, this is the first financial incentive experiment to have examined the potential effects on self-reported test-taking importance. Of course, it is possible that a much more substantial financial incentive (e.g., $10,000) could have influenced test-taking importance statistically significantly. However, it may also be the case that a financial incentive, whether small or large, has a limited capacity to increase the internalization of motivation – a key characteristic associated with valuations and the successful completion of complex tasks (Weibel, Rost, & Osterloh, 2009).

In the light of the above, it may be suggested that alternative, more internally relevant, incentives may facilitate greater effects on test-taking effort and importance. For example, Liu, Rios, and Borden (2015) reported much more substantial experimental effects on Effort and Importance ($g$ = .75 and .89, respectively), based on a sample of 136 university seniors who completed a university-level academic achievement test. For the non-incentive group, the testees were informed, 'Your score on this test will have no effect on your grades or academic standing, but we do encourage you to try your best' (p. 84). By contrast, testees within the incentive group were informed that the scores on the test would be used for the purposes of comparing the quality of colleges in the United States. Consequently, the testees were strongly encouraged to do their best on the test for the sake of their college's national standing. Thus, it may be suggested that a non-financial incentive (e.g., pride) has the capacity to motivate testees substantially more than a financial incentive.

### *Incentives and intelligence test scores*

In contrast to the observation of a moderate and statistically significant effect on test-taking Effort, this investigation failed to observe a statistically significant effect of financial incentives on intelligence test performance.[1] The failure to observe a statistically significant increase in intelligence test performance in this study is consistent with previous financial incentive and cognitive ability testing studies with adults (Borghans *et al.*, 2008; Cole, Bergin, & Summers, 2016; O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005). Thus, the position of a causal effect of test-taking motivation on intelligence test performance in adults does not, yet, appear to be clearly tenable. If an effect exists in the population, the pattern of effect sizes reported in Table 1 suggests that it may be limited to simple processing tasks, which is consistent with our review of Dickstein and Ayers (1973), as well as Bonner *et al.* (2000). More empirical investigations with dedicated intelligence tests are encouraged, as a meta-analysis will be eventually feasible to help evaluate the influence of sampling effects.

The effect sizes observed in this financial incentive investigation are numerically smaller than the effects reported in Liu *et al.*'s (2015) non-financial incentive investigation (i.e., $g$ = .54 to .73). However, it is important to emphasize that only a small number of the adult volunteers in this investigation were substantially unmotivated to complete the intelligence tests. That is, even in the condition for which there was no financial incentive to increase intelligence test performance, only 13% of the testees scored a mean item response of <3.0 on the Effort subscale (theoretical range: 1.0–5.0). By contrast, Liu *et al.*'s (2015) control group was associated with an Effort subscale mean of 2.77, which

[1] *We note that in addition to the Connections subtests and Letter–Number Sequencing, no statistically significant experimental effects were observed for the split-form Advanced Vocabulary Test (*$F = 0.82$, *partial* $\eta^2 < .01$, $p = .367$*), Mental Rotation (*$F = 0.79$, *partial* $\eta^2 < .01$, $p = .375$*), or the Advanced Progressive Matrices (*$F = 0.34$, *partial* $\eta^2 < .01$, $p = .561$*).*

suggests that more than 50% of the testees scored <3.0 on the Effort subscale. Consequently, it may be suggested that Liu *et al.*'s (2015) sample was atypically unmotivated to complete their cognitive ability-type testing, in comparison with other investigations that have administered the Student Opinion Scale to university students (e.g., Thelk *et al.*, 2009; Effort subscale mean = 3.44). Importantly, once the testees who exhibited extremely low test-taking motivation were removed from the Liu *et al.* (2015) sample,[2] the incentive and non-incentive group academic achievement means were not found to differ statistically significantly. In the light of the above, it is suggested that the hypothesized causal effect of test-taking motivation on intelligence test performance may be curvilinear. Specifically, the relationship may be causal from very unmotivated to neutral motivation levels, and non-causal from neutral to very motivated levels. To evaluate such a hypothesis will likely require a large sample (>2,000) as few adult volunteers (<3%) score between 1.0 and 2.5 on the Student Opinion Scale.

Finally, we note that practice effects were apparent across several of the processing speed tasks. That is, on average, the participants tended to improve their performance, irrespective of the condition. There is a substantial literature that supports the observation of cognitive ability testing practice effects (Calamia, Markon, & Tranel, 2012). However, the vast majority of the research is based on pre- and post-test sessions that occur on different days. Based on a meta-analytic investigation, Driskell, Copper, and Moran (1994) found that the duration of mental practice was correlated negatively with test performance ($r \approx -.20$). In this investigation, the two intelligence testing sessions occurred within the same hour and lasted approximately 25 min each. Thus, the observation of a reduction in performance for both groups on the working memory alternate form tests may be considered consistent with the broader literature on the effects of practice and performance within the same session. That is, simpler cognitive capacities may evidence consistent practice effects, whereas more complex capacities (working memory) may evidence fatigue effects (Bovaird, 2002; Wise, 2006).

### *Non-experimental results*

Based on the results of this investigation and others (Chan *et al.*, 1997; Cole *et al.*, 2008; Thelk *et al.*, 2009), it may be stated with appreciable confidence that there is a positive association between test-taking motivation and intelligence test performance in adult volunteers of approximately $r = .30$. Furthermore, based on the partial correlations, it would appear that any possible influence of test-taking importance operates through test-taking effort, as would be expected theoretically (Wigfield & Eccles, 2000). The partial correlation results reported in this investigation coincide well with Cole *et al.* (2008), who reported that the association between self-reported test-taking importance (an unspecified 5-item survey) and academic achievement test scores was mediated completely by test effort (four subject-specific test items) in a sample of 1,005 university students. It is interesting to note that in this investigation, the association between Effort and intelligence test performance remained significant, controlling for the effects of Importance. Such a result implies that there may be stable, unique, trait-like individual differences in the tendency to apply effort at completing intelligence tests. Correspondingly, the distinction between state- and trait-level measurement of test-taking motivation

[2] *Testees were identified as extremely unmotivated with Wise and Ma's (2012) normative threshold method. The method involves omitting testees from the sample who consistently exhibited 10% or less of the average time taken by the whole sample to complete 10% or more of the test items.*

has been advanced (e.g., Penk & Richter, 2017), although not substantially investigated, to date.

Despite the above results, the failure of this investigation and others (Borghans *et al.*, 2008; Cole *et al.*, 2016; O'Neil *et al.*, 2005) to uncover a causal connection between test-taking effort and complex intelligence test performance raises further questions relevant to why the positive association exists, at least in adult volunteers. Of course, the existing experimental research could be regarded as flawed; consequently, it could be contended that the causal effect hypothesis has not yet been evaluated appropriately. Notwithstanding this possibility, it may be time to entertain alternative explanations.

First, the natural assumption that test-taking effort leads to intelligence may be incorrect, at least in adult volunteers. Instead, the effect may lead from intelligence to test-taking effort. It is useful to note a parallel between this contention and the research in the area of test-taking anxiety and intelligence test performance. Specifically, the results of several deficit and interference modelling investigations suggest that intelligence may influence the development of test-taking anxiety, rather than the other way round (Sommer & Arendasy, 2014, 2015; Wicherts & Scholten, 2010). Perhaps the deficit and interference modelling approach could be adapted to the area of test-taking motivation. If it is the case that the direction of causality leads mostly from intelligence to test-taking motivation, it would help explain why the experimental manipulations of test-taking motivation tend to be modest in size.

Second, it is also possible that a causal effect leading from test-taking motivation to intelligence test performance may reside only within the low-to-moderate spectrum of test-taking motivation. Consequently, as this investigation and others have relied upon adult volunteers (although perhaps not Liu *et al.*, 2015), the vast majority of which appear to have been at least neutral test-taking motivation, it may be that incentives would necessarily be ineffective at increasing intelligence test performance. Consequently, substantial experimental manipulations of test-taking motivation and intelligence test performance in adults may be implausible, in most low-stakes research settings, unless an appreciable percentage of the testees are for some reason acutely unmotivated to complete the intelligence testing. By contrast, in children, more substantial experimental effects may be possible, particularly in children with behavioural or learning difficulties (which formed a significant part of the samples included in the Duckworth *et al.* (2011) meta-analysis).

### *Limitations*

The sample size used in this study was reasonably large for a within-subjects design ($N = 99$). However, the power to detect a medium effect was not quite 80% (i.e., 74%). Given the trend of the means, it is possible that a larger sample size may have detected a test-taking motivation effect for the simplest processing speed task administered in this investigation (i.e., Numbers; partial $\eta^2 = .04$, $p = .054$). Thus, future research is encouraged with a diversity of alternate form intelligence tests, including simple processing speed tasks. Should the effects of test-taking motivation be largely restricted to simple processing speed tasks, which are only weakly related to general intelligence ($r^2 \approx .10$; Deary, Der, & Ford, 2001), then previous claims of the appreciable invalidity of intelligence test scores due to individual differences in test-taking motivation may need further qualifications.

Whether the prospect of winning $75 was sufficiently enticing for the typical testee in this study may be questioned, of course. Thus, the failure to observe a statistically

significant increase in intelligence test scores may be suggested to be due to an insufficient incentive. However, such a possibility does not seem tenable, based on comparisons with other investigations. For example, Dickstein and Ayers (1973) offered the testees with the top five total IQ scores $1 each. Based on the results of a purchasing power calculator (inflation-based), $1 in 1973 would be worth approximately $6 today. Also, the maximum a person could earn from Borghans *et al.*'s (2008) study was €30. Thus, the financial incentive offered in this study (three chances to win $75) was relatively large compared to previous investigations, one of which reported an experimental effect (i.e., Dickstein & Ayers, 1973), although only for a single subtest (i.e., Object Assembly).

Finally, it is somewhat concerning that the internal consistency reliabilities for some of the intelligence test scores were atypically low. In particular, the coefficient alphas for the Advanced Vocabulary Tests at time 1 and time 2 were <.50. However, the loadings of the AVT on the general intelligence component were reasonable (≈.50 to .70). Some work suggests that Pearson correlations on dichotomously scored data can underestimate internal consistency reliability (e.g., Sun *et al.*, 2007). However, corrections for such underestimation may not be straightforward (Revelle & Condon, in press). As the reliability of the general intelligence component scores was reasonable (≈.70 to .75), and the experimental analyses were focused on the processing speed and working memory tests, the reported results may be considered largely interpretable. However, it should, nonetheless, be noted that the relatively low reliabilities for some of the intelligence test scores was a limitation of this investigation.

### *Conclusion*

Clear evidence for the contention that intelligence test scores are, to some appreciable degree, invalid due to individual differences in test-taking motivation remains to be reported, at least for adult volunteer samples. Consequently, the substantial validity coefficients reported in the literature supporting the interpretation of IQ scores may not be as biased upwardly as some have suggested, based on analyses of non-adult samples (e.g., Duckworth *et al.*, 2011). As absence of evidence is not evidence of absence, we encourage more research to help understand precisely why test-taking motivation and intelligence test scores are correlated positively in both children and adults.

## Acknowledgement

Thanks to Michelle Ooi and Ka Ki Wong for data collection.

## References

Armor, D. J. (1973). Theta reliability and factor scaling. *Sociological Methodology*, *5*, 17–50. https://doi.org/10.2307/270831

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, *43*, 695–716. https://doi.org/10.1111/j.1744-6570.1990.tb00679.x

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement*, *4*(3), 385–398. https://doi.org/10.1177/014662168000400309

Bonner, S. E., Hastie, R., Sprinkle, G. B., & Young, S. M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting.

*Journal of Management Accounting Research*, *12*(1), 19–64. https://doi.org/10.2308/jmar.2000.12.1.19

Bonner, S. E., & Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: Theories, evidence, and a framework for research. *Accounting, Organizations and Society*, *27*(4), 303–345. https://doi.org/10.1016/S0361-3682(01)00052-6

Borghans, L., Meijers, H., & Ter Weel, B. (2008). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry*, *46*(1), 2–12. https://doi.org/10.1111/j.1465-7295.2007.00073.x

Bovaird, J. A. (2002). New applications in testing: Using response time to increase the construct validity of a latent trait estimate. *Dissertation Abstracts International*, *64*(02), 998B (UMI No. 3082643).

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, *82*(2), 300–310. https://doi.org/10.1037/0021-9010.82.2.300

Cole, J. S., Bergin, D. A., & Summers, J. (2016). A lottery improves performance on a low-stakes test for males but not females. *Assessment in Education: Principles, Policy & Practice*, 1–16. https://doi.org/10.1080/0969594x.2016.1224812

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*, 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Deary, I. J., Der, G., & Ford, G. (2001). Reaction times and intelligence differences: A population-based cohort study. *Intelligence*, *29*, 389–399. https://doi.org/10.1016/S0160-2896(01)00062-9

Dickstein, L. S., & Ayers, J. (1973). Effect of an incentive upon intelligence test performance. *Psychological Reports*, *33*(1), 127–130. https://doi.org/10.1037/h0048465

Driskell, J. E., Copper, C., & Moran, A. (1994). Does mental practice enhance performance? *Journal of Applied Psychology*, *79*(4), 481–492. https://doi.org/10.1037/0021-9010.79.4.481

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*, 7716–7720. https://doi.org/10.1073/pnas.1018601108

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement*, *63*, 915–930. https://doi.org/10.1177/0013164403251319

Fervaha, G., Zakzanis, K. K., Foussias, G., Graff-Guerrero, A., Agid, O., & Remington, G. (2014). Motivational deficits and cognitive test performance in schizophrenia. *JAMA psychiatry*, *71*, 1058–1065. https://doi.org/10.1001/jamapsychiatry.2014.1105

Gignac, G. E., Shankaralingam, M., Walker, K., & Kilpatrick, P. (2016). Short-term memory for faces relates to general intelligence moderately. *Intelligence*, *57*, 96–104. https://doi.org/10.1016/j.intell.2016.05.001

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *82*, 1147–1149. https://doi.org/10.1080/01621459.1987.10478551

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review*, *91*(2), 153–184. https://doi.org/10.1037/0033-295X.91.2.153

Jaccard, J. (1998). *Interaction effects in factorial analysis of variance*. Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412984508

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Kirkwood, M. W. (2015). A rationale for performance validity testing in child and adolescent assessment. In M. W. Kirkwood (Ed.), *Validity testing in child and adolescent assessment: Evaluating exaggeration, feigning, and noncredible effort* (pp. 3–31). New York, NY: Guilford Publications.

Larson, G. E., Merritt, C. R., & Williams, S. E. (1988). Information processing and intelligence: Some implications of task complexity. *Intelligence*, *12*(2), 131–147. https://doi.org/10.1016/0160-2896(88)90012-8

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment*, *20*(2), 79–94. https://doi.org/10.1080/10627197.2015.1028618

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. New York, NY: Prentice-Hall.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment*, *10*, 185–208. https://doi.org/10.1207/s15326977ea1003_3

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-Scale Assessments in Education*, *2*(1), 5. https://doi.org/10.1186/s40536-014-0005-4

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*(1), 55–79. https://doi.org/10.1007/s11092-016-9248-7

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test-different versions and factors that affect performance. *Brain and Cognition*, *28*(1), 39–58. https://doi.org/10.1006/brcg.1995.1032

Raven, J. (1998). *Manual for Raven's progressive matrices and mill hill vocabulary scales. Section 4, Advanced progressive matrices*. Oxford, UK: Oxford Psychologists Press.

Reeve, C. L., & Bonaccio, S. (2011). The nature and structure of "intelligence". In T. Chamorro-Premuzic, A. Furnham & S. von Stumm (Eds.), *Handbook of individual differences* (pp. 187–216). Oxford, UK: Wiley-Blackwell.

Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*(3), 271–276. https://doi.org/10.2466/pms.1958.8.3.271

Revelle, W., & Condon, D. M. (in press). Reliability. In P. Irwing, T. Booth & D. Hughes (Eds.), *The Wiley Blackwell handbook of psychometric testing*. West Sussex, UK: Blackwell Publishing Ltd.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measurements of effect size in educational research. *Educational Research Review*, *6*, 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Richardson, K., & Norgate, S. H. (2015). Does IQ really predict job performance? *Applied Developmental Science*, *19*, 153–169. https://doi.org/10.1080%2f10888691.2014.983635

Salthouse, T. A., Toth, J., Daniels, K., Parks, C., Pak, R., Wolbrette, M., & Hocking, K. J. (2000). Effects of aging on efficiency of task switching in a variant of the trail making test. *Neuropsychology*, *14* (1), 102–111. https://doi.org/10.1037/0894-4105.14.1.102

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, *6*, 147–151. https://doi.org/10.1027/1614-2241/a000016

Shepperd, J. A., & Taylor, K. M. (1999). Social loafing and expectancy-value theory. *Personality and Social Psychology Bulletin*, *25*, 1147–1158. https://doi.org/10.1177/01461672992512008

Sommer, M., & Arendasy, M. E. (2014). Comparing different explanations of the effect of test anxiety on respondents' test scores. *Intelligence*, *42*, 115–127. https://doi.org/10.1016/j.intell.2013.11.003

Sommer, M., & Arendasy, M. E. (2015). Further evidence for the deficit account of the test anxiety–test performance relationship from a high-stakes admission testing setting. *Intelligence*, *53*, 72–80. https://doi.org/10.1080/1045988X.2011.633285

Sun, W., Chou, C. P., Stacy, A. W., Ma, H., Unger, J., & Gallaher, P. (2007). SAS and SPSS macros to calculate standardized Cronbach's alpha using the upper bound of the phi coefficient for dichotomous items. *Behavior Research Methods*, *39*(1), 71–81. https://doi.org/10.3758/BF03192845

Sundre, D. L. (1999). *Does examinee motivation moderate the relationship between test consequences and test performance?* (Report No. TM029964). Harrisonburg, VA: James Madison University. (ERIC Document Reproduction Service No. ED432588).

te Nijenhuis, J., Jongeneel-Grimen, B., & Kirkegaard, E. O. (2014). Are Headstart gains on the g factor? A meta-analysis. *Intelligence*, *46*, 209–215. https://doi.org/10.1016/j.intell.2014.07.001

te Nijenhuis, J., van Vianen, A. E., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, *35*(3), 283–300. https://doi.org//10.1016/j.intell.2006.07.006

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *The Journal of General Education*, *58*, 129–151. https://doi.org/10.1353/jge.0.0047

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three- dimensional spatial visualization. *Perceptual and Motor Skills*, *47*, 599–604. https://doi.org/10.2466/pms.1978.47.2.599

Wechsler, D. (1955). *Wechsler adult intelligence scale – manual*. New York, NY: Psychological Corporation.

Wechsler, D. (2008). *Wechsler adult intelligence scale/Wechsler memory scale-fourth edition: Administration and scoring manual*. San Antonio, TX: Pearson Assessment.

Weibel, A., Rost, K., & Osterloh, M. (2009). Pay for performance in the public sector— Benefits and (hidden) costs. *Journal of Public Administration Research and Theory*, *20*, 387–412. https://doi.org/10.1093/jopart/mup009

Wicherts, J. M., & Scholten, A. Z. (2010). Test anxiety and the validity of cognitive tests: A confirmatory factor analysis perspective and some empirical findings. *Intelligence*, *38*(1), 169–178. https://doi.org/10.1016/j.intell.2009.09.008

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L. (2006). An investigation of the differential effort received by items on a low- stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method*. In annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wood, R. E. (1986). Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*, *37*(1), 60–82. https://doi.org/10.1016/0749-5978(86)90044-0

Wright, P. M., & Kacmar, K. M. (1995). Mediating roles of self-set goals, goal commitment, self-efficacy, and attractiveness in the incentive-performance relation. *Human Performance*, *8*, 263–296. https://doi.org/10.1207/s15327043hup0804_2

*Received 12 October 2017; revised version received 2 January 2018*

## Supporting Information

The following supporting information may be found in the online edition of the article:

**Data S1.** Supplementary material.