Intelligence 92 (2022) 101652

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

Effort impacts IQ test scores in a minor way: A multi-study investigation with healthy adult volunteers[☆]

Timothy C. Bates[a,*,1], Gilles E. Gignac[b,1]

[a] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, UK
[b] School of Psychological Science, University of Western Australia, Australia

ARTICLE INFO

Keywords:
Intelligence
IQ
Motivation
Incentives
Validity

ABSTRACT

Test motivation has been suggested to strongly influence low-stakes intelligence scores, with for instance, a recent meta-analysis of monetary incentive effects suggesting an average 9.6 IQ point impact ($d = 0.64$). Effects of such magnitude would have important implications for the predictive validity of intelligence tests. We report six studies ($N = 4208$) investigating the association and potential causal link of effort on cognitive performance. In three tests of the association of motivation with cognitive test scores we find a positive, but modest linear association of scores with reported effort ($N = 3007$: $r \sim 0.28$). In three randomized control tests of the effects of monetary incentive on test scores (total $N = 1201$), incentive effects were statistically non-significant in each study, showed no dose dependency, and jointly indicated an effect one quarter the size previously estimated ($d = 0.166$). These results suggest that, in neurotypical adults, individual differences in test motivation have, on average, a negligible influence on intelligence test performance. (≈ 2.5 IQ points). The association between test motivation and test performance likely partly reflects differences in ability, and subjective effort partly reflects outcome expectations.

1. Introduction

Researchers have argued that scores on cognitive ability tests largely reflect differences in test-taking motivation (e.g. Kirkwood, 2015). Consequently, some have suggested intelligence scores may underestimate student ability for practical purposes such as evaluating academic ability. Understanding whether cognitive ability scores are materially impacted by effort is therefore an important question.

Cross-sectional data have suggested associations between effort and test scores as high as $r = 0.50$ (Cole, Bergin, & Whittaker, 2008) although most are considerably smaller. It remains to be determined, however, to what extent such cross-sectional correlations reflect effort raising ability versus reversed causality with ability raising effort, or if, contrarily, subjective effort functions as an evaluation of the adequacy of performance for the goal. Experimental studies have been used to ostensibly support a causal role of effort on scores, with a meta-analysis concluding that small monetary incentives could improve test scores by 0.64 SDs (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). Despite the importance of the question, nearly all experimental studies have been small (e.g., per-cell Ns of under 16), and often based on psychiatric or forensic samples (e.g. Fervaha et al., 2014). Furthermore, a majority are decades old and questions have been raised about the largest effect sizes in the analysis (Breuning & Zella, 1978) as likely involving fraud by Breuning (without Zella's knowledge) and in need of retraction (Warne, 2022). Without these fraudulent data, the total N of all studies summarised by (Duckworth et al., 2011) falls to just 1523 subjects – surprisingly few given decades of study and the theoretical and practical importance of the question. Thus, the purpose of the present paper was to 1) Estimate the cross-sectional correlation of reported effort with cognitive ability scores in three large, neurotypical/community samples, and 2) Conduct a series of randomized control interventions manipulating effort via incentives. Before presenting these, we first summarise the especially relevant literature.

1.1. Effort and intellectual performance: correlational evidence

For professional tests, there is little question that subjects are motivated, with test procedures requiring the attention and participation of

[☆] Support for this project was provided in part by The John Templeton Foundation (US) (GF13666.154364).

[*] Corresponding author at: School of Philosophy, Psychology and Language Sciences, Psychology Building, 7 George Square, Edinburgh EH8 9JZ, UK.
E-mail address: tim.bates@ed.ac.uk (T.C. Bates).

[1] Both authors contributed equally to this project.

https://doi.org/10.1016/j.intell.2022.101652
Received 8 December 2021; Received in revised form 14 April 2022; Accepted 14 April 2022
Available online 25 April 2022

the test-taker, including specific procedures to elicit optimal responses to maintain morale (e.g., ordering items in terms of difficulty; (e.g. Wechsler, 2008). Correspondingly, for so-called “high-stakes” tests – such as employment hurdles or exams, no association of test-taking motivation is found (O'Neil, Sugrue, & Baker, 1995). What remains less clear is the role of effort in tests in low-stakes settings, e.g., purposes of no-consequence to the test-taker (Duckworth et al., 2011). The association between test-taking effort and intelligence test performance can be ascertained by administering one or more cognitive ability tests, followed by a measure of test-taking effort such as the Student Opinion Scale (SOS: Sundre & Thelk, 2007). The SOS consists of 10 items and measures two positively correlated dimensions: test-taking importance (5-items, e.g., ‘*Doing well on these tests was important to me*’) and test-taking effort (5-items, e.g., ‘*I engaged in good effort throughout these tests*’).

A recent meta-analysis of the association between self-reported test-taking effort and low-stakes test performance estimated the effect at $r = 0.33$ (Silm, Pedaste, & Täht, 2020), suggesting that approximately 90% of intelligence test variance is independent of test motivation. However, relatively few studies include participants where self-report test-taking effort was low, with much of the research to date based on university/school samples – people for whom test-taking motivation might be expected to be high as a trait (Gignac, Bartulovich, & Salleo, 2019). For example, Gignac et al. (2019) reported a mean of 3.82 on the SOS effort subscale (theoretical range 1 to 5) in a sample of undergraduate students ($N = 219$) who completed a battery of cognitive ability tests in an anonymous fashion with no opportunity to learn about their performance. Furthermore, less than 6% of the sample reported their test-taking effort at less than 3.0. Therefore, Gignac et al. (2019) suggested that a much larger sample, and ideally one more representative of the general community, would be required to estimate the precise nature and strength of the association between test-taking effort and intelligence test performance.

An additional limitation associated with the current literature is that investigators tend to administer multiple intelligence tests followed by the self-reported effort questionnaire. That is, it is known that test-taking effort decreases across a testing session and that there are individual differences in the degree to which test-taking effort reduces across a testing session (Gignac & Wong, 2020; Penk & Richter, 2016). Therefore, asking people to report their degree of test-taking effort based on their experience completing a collection of intelligence tests may be difficult for the typical test taker, suggesting less validity for the responses than would otherwise be the case had they completed only a single test of cognitive ability.

Considering the above, our first aim was to estimate with respectable precision the nature and strength of the association between self-reported effort and cognitive ability test performance across a range of cognitive ability tests under a low-stakes scenario. Sample sizes of 1000 would be employed, to ensure high statistical power and a wide range of responses. Furthermore, across each sample, the respondents would be required to complete a single, relatively short test of cognitive ability, followed by the requirement to complete the self-reported test-taking motivation questionnaire, to increase the validity of the test-effort responses.

1.2. *Effort and intellectual performance: experimental evidence*

Assuming a positive correlation between reported effort and test scores, the question of causality is raised. On the one hand, non-cognitive traits associated with motivation, such as openness to experience, conscientiousness, or self-control could increase test performance, independently of cognitive ability (Demange et al., 2021). On the other hand, the reversal of this causal assumption is also possible, such that greater ability is associated with higher reported effort, but with no effect on cognitive ability or task performance. In this case, effort reports would function as reflections of confidence or expectation of having completed the test successfully, rather than causing a change in ability (Gignac, 2018). Consider, for example, that the association between test-taking anxiety and test performance has been shown to be due to the effects of ability on test-taking anxiety, rather than the other way around (Sommer & Arendasy, 2015).

It is unclear, then, if effort has a causal effect on performance on cognitive ability items. Therefore, studies are needed to estimate the association between effort and test scores (the subject of our first set of studies), and most importantly, interventions manipulating effort are required to investigate any potential causal effect of motivation on cognitive ability performance. Testing this was our second aim and is developed in study 2 onward. First, we test the association of test scores with effort in three large samples.

2. Study 1

In study 1, we set out to obtain a large sample in which both ability and effort were measured. Gignac et al. (2019) suggested the effect of test-taking motivation and test performance may show threshold effect: that is, beyond a moderate level of test-taking motivation, the potential influence of test-taking motivation on test performance diminishes materially. Therefore, we also investigated whether test-taking effort was linearly associated with ability scores or if a quadratic component was present.

Three sub-studies were undertaken, each of 1000 subjects completing a self-report measure of test-taking motivation following one of three cognitive ability tests: (1) A sentence verification test linked to processing speed (Baddeley, 1968); (2) A paper folding task linked to spatial manipulation (Ekstrom, French, Harman, & Dermen, 1976); or (3) A vocabulary test (Warrington, McKenna, & Orpwood, 1998). These tasks were chosen to measure three primary domains of intelligence: processing speed, spatial reasoning, and crystallised intelligence, respectively. Hypotheses, methods, and *N* were pre-registered <https://aspredicted.org/KW8_7C1>, <https://aspredicted.org/YHS_5PJ>.

2.1. *Method*

2.1.1. *Samples*

Subjects in all studies were recruited from prolific academic, a crowd sourcing online platform to recruit human subjects for research purposes. For study 1a, we recruited 1001 adult subjects (age $M = 28.41$, $SD = 6.04$; range: 18 to 39 years, 499 male and 499 female, 2 did not answer this item). For study 1b, we recruited 1000 adult subjects (age $M = 34.49$, $SD = 11.75$; range: 18 to 76 years) from prolific academic (497 male and 503 female). The sample was predominantly white (White = 89.7%; Asian = 4.5%; Black = 1.8%; South-East Asian = 1.4%; Other = 2.6%). For study 1c, we recruited 1006 adult subjects (age $M = 24.31$, $SD = 4.79$; range: 18 to 39 years) from prolific academic (502 male and 504 female). The sample composition was: White = 41.5%; Asian = 0.9%; Black = 35.3%; South-East Asian = 0.4%; Native American = 0.9%; Other = 21.1%. Subjects were paid 50 pence for their participation. Ethics approval was obtained from the University of Edinburgh Psychology ethics committee.

2.1.2. *Materials*

In Study 1A, cognitive ability was assessed with Form B of the 64-item Baddeley (1968) 3-minute sentence verification test. Items require grammatical transformation to evaluate the truth of a simple sentence, e.g. “*A does not follow B: **AB***” (TRUE). Form B includes 32 items and subjects were given 90 s to complete as many items as possible. Coefficient ω in our sample was 0.90. Study 1B used the test of Single Word Comprehension (Warrington et al., 1998). This test consists of 52 target words, each presented with two potential response words arranged below them, and for each target must select the word which is the best synonym (e.g., MARQUEE: Tent; Palace). Half are concrete and half abstract. Based on item level data, we created a short form with 13 concrete items and 12 abstract items. Coefficient ωin our sample was

0.62. Study 1C used Form A (10-items) of the 20-item Visual Paper Folding test (Ekstrom et al., 1976). Dating back in form not only to the work of Thurstone, but at least as early as Binet (1905/1916), this scale consists of illustrations depicting a square sheet of paper being folded two or three times and a hole punched in it. The task is to select which of 5 graphical response options depicts how the holes would appear if the sheet was unfolded. Matched versions are provided as part of the *Kit of Factor-Referenced Cognitive Tests*. Each block consisted of 10 items with a 3-minute time limit. Coefficient ω in our sample was 0.68.

Across all three sub-studies, test-taking effort and importance was measured using the 10-item Student Opinion Scale (SOS: Sundre & Thelk, 2007). Five items assess test-taking effort, with a representative item being '*I gave my best effort on these tests*'. Subjects respond on a 1 to 5 Likert scale with anchors from 'Strongly Disagree' through "Neutral" to 'Strongly Agree'. The coefficient ωs in our samples were 0.72 for study 1a, 0.76 for study 1b, and 0.58 for study 1c, respectively.

2.1.3. *Procedure*

Testing took place online using the Prolific academic and Qualtrics platforms. After providing informed consent, subjects completed the paper folding test followed by the SOS. Testing took around 5 min.

2.2. *Results*

All analyses were conducted with *SPSS (Version 27)*. As can be seen in Table 2, the distribution of ability and effort scores were sufficiently normal across all three samples (e.g., skew < |2.0|; (Schmider, Ziegler, Danay, Beyer, & Bühner, 2010; Zuo et al., 2011). Furthermore, only one outlier was identified across all three studies, based on the outlier labelling rule with an inter-quartile range multiplier of 3.0 (Hoaglin & Iglewicz, 1987). Specifically, an effort value of 1.0 was identified as an outlying value in study 1a. However, winsorizing the value to the next smallest value that was not an outlier (i.e., 2.0) changed the results to only the third decimal place, consequently, we proceeded with the analyses.

Descriptive statistics for each of the three studies are shown in Table 1, along with the results of regression models testing the association of cognitive score with SOS effort.

As can be seen in Table 1, the correlations between self-reported test-taking effort and cognitive ability test performance were consistent across the three different tests, with estimates in a tight range of $r = 0.26$ and $r = 0.29$ (all $p < .001$). Disattenuated for imperfect reliability in the cognitive ability and test effort scores, the correlations were $r' = 0.34$ (sentence verification), $r' = 0.39$ (synonyms), and 0.43 (paper folding). Thus, test-taking effort and cognitive ability test performance shared between 12 and 18% of their true score variance.

There was no evidence for any quadratic relationship, based on a series of hierarchical multiple regressions with the effort linear term entered at step 1 and the effort quadratic term entered at step 2. Specifically, all of the R^2 changes were statistically non-significant (see Table 1), as were the quadratic effect standardized beta-weights: Sentence verification: $\beta = -0.21$, $t(999) = -0.81$, $p = .419$; Synonyms: $\beta = 0.18$, $t(998) = 0.49$, $p = .625$; Paper folding: $\beta = 0.44$, $t(1004) = 1.66$, $p = .098$. Furthermore, as can be seen in Fig. 1, the association was linear across the whole spectrum of effort/ability for all three abilities, based on a LOESS curve fitting analysis (Cleveland, 1979). (Figs. 2 and 3.)

2.3. *Discussion*

The main finding from study 1 was the observation of a consistent, positive monotonic association of reported test-taking effort with obtained test performance scores across all three cognitive ability dimensions: processing speed (sentence verification), visual-crystallised intelligence (synonyms) and visual-perceptual intelligence (paper folding). Furthermore, the magnitude of the correlations was consistent with the meta-analytically estimated correlation of $r = 0.33$ (Silm et al., 2020). Thus, our results, based on the largest samples to-date, confirm the previously published literature in the area, but demonstrating this in single, large pre-registered studies.

The effect of effort, accounting for around 10% of variance, was 40% smaller than that reported by Cole et al. (2008) and less than half the meta-analytic effect reported by Duckworth et al. (2011) as the incremental effect on scores based on a monetary incentive. A unique feature to this investigation was that subjects self-reported their test-taking effort after the completion of a single, relatively short cognitive ability test. By contrast, previous investigations typically had subjects complete a battery of tests over many minutes and then had them report their test-taking effort (e.g., Gignac, 2018; Gignac et al., 2019; Merritt et al., 2019). Thus, in our investigation, effort should therefore be maximally informed by the cognitive task.

We also failed to support the suggestion by Gignac et al. (2019) that the association between effort and test performance may be consistent with a threshold effect: that is, beyond approximately the midpoint of effort, the association between effort and test performance diminishes substantially. On the basis of three large samples and LOESS regression analyses, we failed to identify any hint of a threshold effect. We return this finding and interpretation in the General Discussion.

This association of effort and performance found in study 1 is compatible with a potentially modest to moderate effect of effort on cognitive ability. Equally, however, it is compatible with a modest to moderate effect of ability on effort (or feeling of effort) – for instance, subjects with higher expectation of performing well exert more effort. Related to this, subjective effort may be a post-hoc evaluation of effective performance, i.e., feelings of effective effort reflect and underlying computation of expected performance. This is of course speculation. One prediction from such a model would be that subjective effort should decrease with increasing task difficulty. The results were compatible with this, with subjects reporting less effort on the (objectively more difficult) paper folding test (see Table 1) and they thus expected to have done less well. Next, in study 2, we turn to a randomized control trial targeting the question of direction of causation, testing if an incentive manipulation can cause an increase in performance.

3. Study 2

Study 1 established a positive association between effort and test scores equal to approximately $r \approx 0.30$. However, causality remains unclear. In study 2, therefore, we turned to a manipulation of effort. If effort is causing the association with scores, and motivation in low-stakes research settings is a major factor in scores, then, as Duckworth

Table 1
Association of effort with test performance score: studies 1A, 1B and 1C.

		Ability			Effort			Linear			Nonlinear		
Study/Measure	*N*	*M* (*SD*)	Skew	Kurtosis	*M* (*SD*)	Skew	Kurtosis	*r*	95% CI	*t*	ΔR^2	*F*	*p*
1A: Baddeley task	1001	17.23 (6.72)	0.10	−0.54	4.11 (0.60)	−0.74	0.87	0.29	[0.22, 0.34]	9.50	0.001	0.65	0.419
1B: Synonyms	1000	17.95 (3.11)	−0.23	−0.26	4.42 (0.52)	−0.81	0.44	0.27	[0.21, 0.33]	8.91	0.001	0.24	0.625
1C: Paper folding	1006	4.33 (2.43)	0.35	−0.67	3.83 (0.59)	−0.36	0.26	0.27	[0.22, 0.33]	8.99	0.003	2.74	0.098

Note. All correlations statistically significant, $p < .001$; Baddeley task = sentence verification; nonlinear = quadratic function in hierarchical multiple regression; $\Delta R^2 =$ hierarchical multiple regression change in R^2 associated with effort quadratic term predicting ability beyond the linear term.

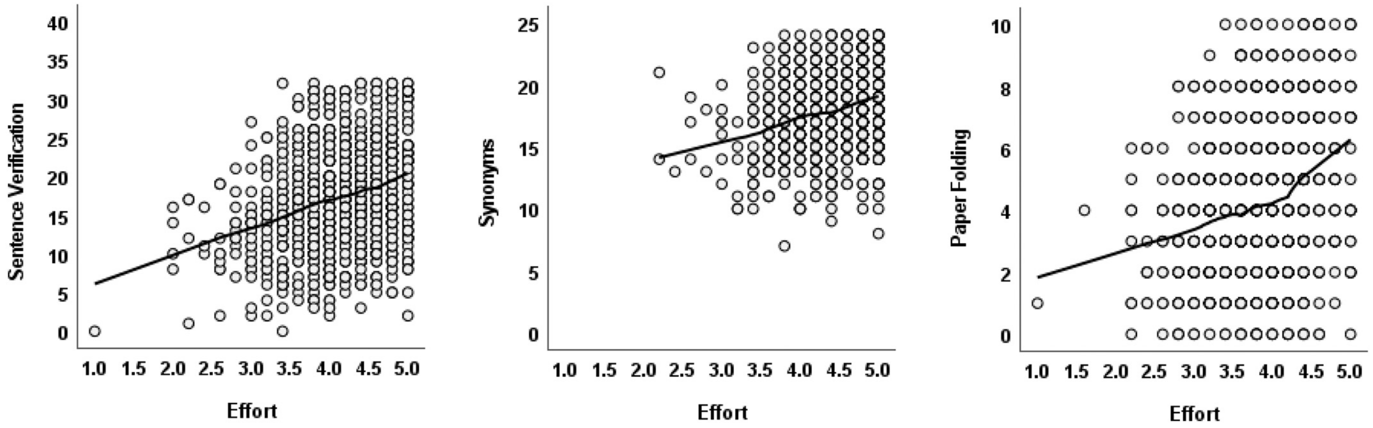

Fig. 1. Scatter plots depicting the association between effort and test performance across all three cognitive abilities (LOESS regression line of best fit; Epanechnikov: Span = 50%).

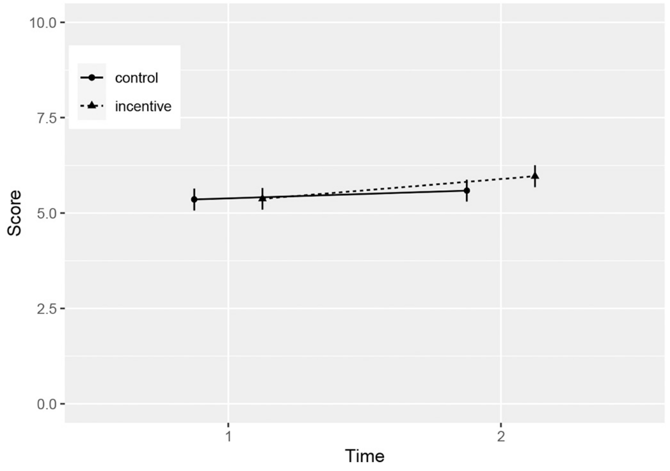

Fig. 2. Plot of Ability (Paper Folding) Means and 95% Confidence Intervals Across Conditions (time 1 = non-incentive; time 2 = incentive) and Across Groups (Study 2).

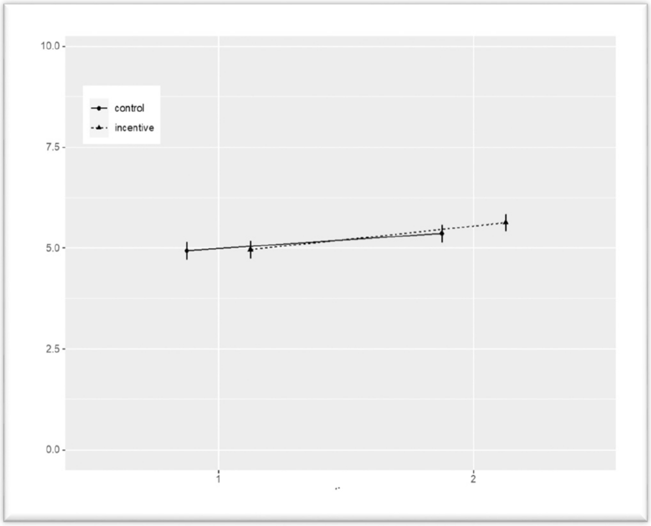

Fig. 3. Plot of ability (paper folding) means and 95% confidence intervals across conditions (time 1 = non-incentive; time 2 = incentive) and across groups (Study 3).

et al. (2011, p. 7717) note "…*incentives should substantially improve their performance*". We next introduce background research on such experimental trials motivating the present study.

Duckworth et al. (2011) found a statistically significant effect, concentrated among individuals with lower performance in non-incentive conditions, making it perhaps even more consequential. However, there are several limitations associated with the studies included in the Duckworth et al. (2011) meta-analysis. First, around half the studies had no baseline measure for comparison. Additionally, half of the studies had *N* of 16 or less in the incentive group, with the largest reported *N* = 105. Only two of the studies were based on neurotypical adults. Some studies confounded incentive with information in the form of trial-by-trial feedback. Three of the samples (Breuning & Zella, 1978) were collected by researchers subsequently convicted of research fraud (see Haynes, 1988) rendering the veracity of these data in doubt. Furthermore, meta-analyses themselves are hindered by choice of study, publication bias affecting study availability, and other issues (Flather, Farkouh, Pogue, & Yusuf, 1997; Ioannidis & Lau, 1999). All of these research characteristics are now known to be indicative of an inflated effect size in a meta-analysis (Warne, 2021). Finally, we note that, surprisingly, given the intense interest in non-cognitive effects on educational and life outcomes (Garcia, 2014; Kautz, Heckman, Diris, Ter Weel, & Borghans, 2014) and the role which effort may play in ability test scores, the total *N* of all studies in the meta-analysis was equal to only 2008.

In a relatively recent experimental investigation, Gignac (2018) had first year university student volunteers (*N* = 99) complete a battery of five cognitive ability tests under two conditions: non-incentive and incentive. The incentive consisted of three chances to win $75 for achieving an overall cognitive ability test score in the top 10%. Half of the subjects were given the chance to win the money at time 1 and the other half were given the chance of to win the money at time 2: thus, the subjects served as their own control. Although a correlation of *r* = 0.28 was found between test-taking effort and overall cognitive ability test performance, Gignac (2018) failed to observe any effect of incentive on performance. Gignac (2018) interpreted the results to suggest that test-taking effort may not have a causal effect on test performance, at least for healthy, adult volunteers. In another recent study, Merritt et al. (2019) reported similar null effects in a between-groups design, based on a sample 81 undergraduate volunteers (*N* = 42 incentive group) and the possibility of winning $20 for achieving a cognitive ability test score in the top one third of the sample.

Although these results are useful, larger studies, using incentives that are available to all subjects for improving their score and in a repeated measures design with controls for the improvement expected with repeated practice are necessary to understand the effect of incentive on performance. Incentivising subjects by basing the reward upon each subjects' own performance obviates the concern that many subjects may view the possibility of obtaining a score in the top 10% or 30% as unachievable. If subjects can earn a financial reward by increasing their own performance at time 2, a larger percentage of subjects may interpret the possibility as achievable, and, correspondingly, expend more effort.

Another limitation associated with Gignac (2018) and Merritt et al. (2019), and the whole field more generally, is statistical power. Ideally, a larger sample size would be employed to achieve power greater than 0.80 to detect even effects much smaller than those proposed in Duckworth et al. (2011), perhaps $d = 0.20$ or even smaller. Non university samples are also desirable. Thus, in study 2, we sought to test experimentally the potential causal effect of test-taking motivation on cognitive ability test performance, by employing an experimental manipulation strategy that may be expected to increase test-taking motivation more substantially than previous investigations, in a relatively large, unselected sample of adult volunteers.

In devising our randomized intervention targeting effort with a material incentive, we chose a mixed repeated measures and between-subjects design, allowing control of practice effects. Specifically, subjects were randomly assigned to an incentive or control condition. Both groups completed two matched blocks of visual-spatial ability tests (paper folding), however, the motivation group was given the opportunity to receive a financial award for increasing their performance on the second occasion by 10% or more.

We hypothesized:

1. Test performance will be higher at time 2 than time 1 (practice effect).
2. The incentive group will increase more from time 1 to time 2 compared to the control group (Time * condition interaction).

Based on the meta-analytic effect size of 0.64, and using the Superpower package (Lakens & Caldwell, 2021), we estimated power at 95% with N per condition = 50. We wished to comfortably exceed this, and therefore ran 400 subjects (N per condition 200, power = 80% for an effect of 0.2). The study was pre-registered: <https://aspredicted.org/8GC_8P8>

3.1. Method

3.1.1. Sample

We recruited 400 adult subjects (age $M = 29.75$, $SD = 5.90$; range: 18 to 40 years) from prolific academic (202 male and 198 female). The sample was predominantly white (White = 92.5%; Asian = 3.0%; Black = 1.3%; South East Asian = 0.5%; Other = 2.8%). Subjects were paid £1.20 for their participation (excluding bonus).

3.1.2. Materials

Intelligence was measured using Form A and Form B of the Visual Paper Folding test (Ekstrom et al., 1976). Each form includes 10 items (3-minute time limit) and the forms are calibrated as approximately equally difficult. For further details, see Study 1. For the total sample, coefficient ω was estimated: Form A = 0.688; Form B = 0.627. Effort was measured using the 10-item Sundre and Thelk (2007) Student Opinion Scale (see Study 1 for further details). Internal consistency reliability was estimated at $\omega = 0.81$ and 0.80 for the pre- and post-effort conditions. The reliabilities for each experimental condition are reported in Table 2.

3.1.3. Procedure

Subjects completed the testing online. After consenting, subjects completed Form A of the paper folding test, followed by the completion of the effort questionnaire. Next, half the subjects at random were selected for the motivation condition, receiving the message, "*In this second part of the study, you will be given the opportunity to earn an additional £2 if you can improve your test performance by 10%. The test is very similar to the first test you completed. You will also have the same amount of time to complete the test. We will score both tests and if your score is 10% higher on the second test, you will be awarded a bonus £2 pounds*." Subject then saw an item asking if incentive was nothing, £2, or 10p, and could not proceed before choosing the correct answer. Next, the subjects completed Form B of the paper folding test followed by the effort questionnaire.

3.2. Results

All analyses were conducted with *SPSS* (version 27). As can be seen in Table 2, the distribution of ability and effort scores were sufficiently normal (e.g., skew < |2.0|; (Schmider et al., 2010; Zuo et al., 2011). Furthermore, no outliers were identified, based on the outlier labelling rule with an inter-quartile range multiplier of 3.0 (Hoaglin & Iglewicz, 1987). Prior to conducting the 2 × 2 mixed-design ANOVA, the assumption of equality of covariance matrices across groups was tested and satisfied, Box's M test: $F(3, 28{,}603{,}250.33) = 2.37$, $p = .501$.

The mixed-design ANOVA identified a statistically significant main effect of time, $F(1, 398) = 19.66$, $p < .001$, $\eta^2_{\text{partial}} = 0.047$, suggesting that, with the data collapsed across both groups, performance on the paper folding test increased from time 1 ($M = 5.37$; $SD = 2.20$) to time 2 ($M = 5.78$; $SD = 1.92$), i.e., the presence of at least a practice effect. By contrast, the group main effect was not significant statistically, $F(1, 398) = 1.14$, $p = .286$, $\eta^2_{\text{partial}} = 0.003$. Finally, the time × group interaction was not statistically significant, $F(1, 398) = 3.78$, $p = .053$, $\eta^2_{\text{partial}} = 0.009$ (See Fig 2), suggesting that the degree of improvement in ability scores for the incentive group was not statistically significantly greater than the control group.

The correlation between the time 1 and time 2 test performance was $r = 0.60$ (95%CI: 0.53, 0.66). Furthermore, the magnitude of the time 1 and time 2 test performance difference in the means in standardized format (Hedges' g) for the control and incentive groups were $g = -0.11$ ($t = -1.86$, $p = .065$) and $g = -0.28$ ($t = -4.30$, $p < .001$), respectively. Disattenuated for imperfect reliability (Bobko, Roth, & Bobko, 2001), the Hedges' g estimates were $g' = -0.17$ and $g' = -0.42$, respectively. Thus, the numerical difference in the magnitude of the disattenuated

Table 2
Mean, SD, Skew and Kurtosis for ability and effort measures by Group and Time (Study 2).

	Time 1					Time 2				
	M	*SD*	Skew	Kurtosis	*ω*	*M*	*SD*	Skew	Kurtosis	*ω*
	Control group (*N* = 199)									
Ability	5.36	2.15	−0.03	−0.42	0.68	5.59	1.92	−0.11	−0.41	0.61
Effort	4.17	0.58	−0.38	−0.46	0.81	4.13	0.61	−0.43	−0.20	0.79
	Incentive group (*N* = 201)									
Ability	5.37	2.26	−0.06	−0.96	0.70	5.97	1.90	−0.59	−0.08	0.64
Effort	4.06	0.61	−0.14	−0.74	0.70	4.20	0.58	−0.18	−0.92	0.81

Note. Ability = performance on the paper folding test; Effort = test-taking effort; ω = coefficient Omega reliability.

effects was equal to $\Delta g' = 0.25$, suggesting an IQ point difference of 3.75, although, to repeat, the magnitude of the interaction effect was not significant statistically ($p = .053$).

Similar to test performance, with the data collapsed across both groups, mean effort was found to increase from time 1 ($M = 4.11$; $SD = 0.60$) to time 2 ($M = 4.17$; $SD = 0.60$), $F(1, 398) = 6.38$, $p = .012$, $\eta^2_{\text{partial}} = 0.016$. However, the magnitude of the change was statistically significantly larger for the incentive group (Hedges' $g = -0.23$, $t = -4.56$ $p < .001$, $g' = 0.31$) than the control group (Hedges' $g = 0.06$, $t = 1.21$, $p = .228$, $g' = 0.08$), based on the time × group interaction, $F(1, 398) = 17.36$, $p < .001$, $\eta^2_{\text{partial}} = 0.042$. As can be seen in Table 2, the data were also sufficiently normal (skew < |2.0|) and no outliers were identified. Finally, the assumption of equality of covariance matrices was satisfied, Box's M test: $F(3, 28{,}603{,}250.33) = 0.93$, $p = .424$.

3.3. Discussion

Although study 2 was well powered based on the Duckworth et al. (2011) meta-analytic effect size estimate, we found only a numerically small and statistically non-significant interaction effect of incentive on test scores. This is compatible with the hypothesis that the bulk of the association of effort with test scores is due to effects of ability on effort or feelings of effort, rather than causal effects of effort on test scores, as suggested by Gignac (2018). However, a post-hoc power analysis suggested the interaction effect analysis had power of just 0.492, and considering the key interaction effect p-value of 0.053 reported in study 2, we sought to re-test the hypothesis that test performance could be improved (beyond a practice effect) with the same financial incentive (£2) but with a larger sample size, selected after re-evaluating our power for a possible real but much smaller effect.

4. Study 3

Using the Superpower package (Caldwell, Lakens, & Parlett-Pelleriti, 2021) to model a mixed effects 2b*2w ANOVA with the observed 0.6 correlation between test performance found in study 2 indicated that $N = 400$ subjects per group would yield 88% power to detect a reduced effect size of $d = 0.2$ improvement in performance due to incentive (i.e., 3 IQ points greater performance improvement than the control group). We therefore conducted a replication of study 2 with double the sample size.

4.1. Methods

4.1.1. Sample

We recruited 801 adult subjects (age $M = 36.11$, $SD = 12.89$; range: 18 to 76 years) from prolific academic (402 male and 399 female). The sample was predominantly white (White = 89.0%; Asian = 4.6%; Black = 2.2%; South-East Asian = 1.0%; Other = 3.1%) . Subjects were paid £1.20 for their participation (excluding bonus).

4.1.2. Materials

The materials were identical to those used in study 2. For the total sample, paper folding Form A coefficient $\omega = 0.72$; Form B coefficient $\omega = 0.70$; SOS-Effort pre-test coefficient $\omega = 0.80$; SOS-Effort post-test coefficient $\omega = 0.82$. The reliabilities for each experimental condition are reported in Table 3.

4.1.3. Procedure

The procedure was identical to that used in study 2. Subjects who participated in study 1 or study 2 were excluded from participating in study 3.

4.2. Results

As can be seen in Table 3, the distribution of ability and effort scores were sufficiently normal (e.g., skew < |2.0|; (Schmider et al., 2010; Zuo et al., 2011). Furthermore, no outliers were identified, based on the outlier labelling rule with an inter-quartile range multiplier of 3.0 (Hoaglin & Iglewicz, 1987). Prior to conducting the 2 × 2 mixed-design ANOVA, the assumption of equality of covariance matrices across groups was tested and satisfied, Box's M test: $F(3, 115{,}478{,}999.00) = 0.22$, $p = .883$.

The mixed-design ANOVA identified a statistically significant main effect of time, $F(1, 799) = 72.28$, $p < .001$, $\eta^2_{\text{partial}} = 0.083$, suggesting that, with the data collapsed across both groups, performance on the paper folding test increased from time 1 ($M = 4.95$; $SD = 2.36$) to time 2 ($M = 5.50$; $SD = 2.09$), i.e., the presence of at least a practice effect. By contrast, the group main effect was not significant statistically, $F(1, 799) = 1.03$, $p = .310$, $\eta^2_{\text{partial}} = 0.001$. Finally, as per study 2, the time × group interaction was not statistically significant, $F(1, 799) = 3.45$, $p = .064$, $\eta^2_{\text{partial}} = 0.004$ (See Fig 3), suggesting that the degree of improvement in ability scores for the incentive group was not statistically significantly greater than the control group.

The correlation between the time 1 and time 2 test performance was $r = 0.67$ (95%CI: 0.63, 0.70). Furthermore, the magnitude of the time 1 and time 2 test performance means in standardized format (Hedges' g) for the control and incentive groups were $g = -0.19$ ($t = 4.67$, $p < .001$) and $g = -0.31$ ($t = -7.38$, $p < .001$), respectively. Disattenuated for imperfect reliability, the Hedges' g estimates were $g' = -0.27$ and $g' = -0.44$, respectively. Thus, the numerical difference in the magnitude of the disattenuated effects was equal to $\Delta g' = 0.17$, suggesting an IQ point difference of 2.55, although, to repeat, the magnitude of the interaction effect was not significant statistically ($p = .064$).

Similar to test performance, with the data collapsed across both groups, mean effort was found to increase from time 1 ($M = 4.17$; $SD = 0.61$) to time 2 ($M = 4.24$; $SD = 0.62$), $F(1, 799) = 23.14$, $p < .001$, $\eta^2_{\text{partial}} = 0.028$. However, the magnitude of the change was statistically significantly larger for the incentive group (Hedges' $g = -0.26$, $t = -7.41$, $p < .001$ $g' = -0.34$) than the control group (Hedges' $g = 0.03$, $t = 0.91$, $p = .362$, $g' = 0.04$), $F(1, 799) = 36.62$, $p < .001$, $\eta^2_{\text{partial}} = 0.044$. As can be seen in Table 3, the data were also sufficiently normal (skew < |2.0|) and no outliers were identified. Finally, although the assumption

Table 3
Mean, SD, Skew and Kurtosis for Ability and Effort Measures by Group and Time (Study 3).

	Time 1					Time 2				
	M	*SD*	skew	kurtosis	*ω*	*M*	*SD*	skew	kurtosis	*ω*
	Control group (*N* = 398)									
Ability	4.93	2.39	0.04	−0.83	0.73	5.36	2.12	−0.37	−0.62	0.70
Effort	4.21	0.60	−0.90	0.98	0.80	4.20	0.66	−1.07	1.96	0.81
	Incentive group (*N* = 403)									
Ability	4.96	2.33	0.09	−0.61	0.72	5.63	2.04	−0.44	−0.15	0.70
Effort	4.13	0.62	−0.53	0.04	0.80	4.29	0.58	−0.62	0.02	0.74

Note. Ability = performance on the paper folding test; Effort = test-taking effort; ω = coefficient ω reliability.

of equality of covariance matrices was not satisfied, Box's M test: F(3, 115,478,998.996) = 6.47, p < .001, the conclusions were considered accurate, as the corresponding effort variances/covariance were different by less than 30%, and the sample sizes were essentially equal (Keselman, Algina, & Kowalchuk, 2001).

4.3. Discussion

Consistent with Study 2, we again failed to observe a statistically significant effect to suggest that test performance could be increased with a financial incentive, even though the power to detect an effect was estimated at 88%. With the larger sample size used in Study 3, the effect size point-estimate was found to be numerically smaller than the (statistically non-significant) effect estimated in Study 2 (1.8 IQ points vs. 2.55): a reminder that an appreciable increase in sample size does not necessarily turn a marginally statistically significant effect (p = .054) into a significant effect.

In contrast to test performance, and consistent with Study 2, effort was found to be statistically significant larger at time 2 in the incentive group, in comparison to the control group, however the effort means were 4.2 vs. 4.3 on the 5-point Likert scale (Hedges' g = −0.15). Considering these results, for Study 4 we turned our attention to attempting to increase effort more substantially, by offering a five times greater financial incentive (£10) to improve test performance and combining data across studies 2 and 3 to maximise power.

5. Study 4

In study 4, we wished to further test the potential effect of incentive on test performance by running subjects in a high incentive condition. If incentives play a major part in cognitive performance (Bonner & Sprinkle, 2002), then their effects should be dose-responsive. That is, the effect of incentive on cognitive performance should be greater for greater reward. To test this, we offered incentives of £5 for any improvement over time 1 performance and £10 for improvement of 2 items or more over time 1 performance. A second goal of this study was to combine data across all of our experimental samples, enlarging the N at the control and base (£2) incentive conditions, thus, allowing a powerful test of the hypothesized monotonic effect of incentive on cognitive performance. Because studies 2–4 all drew on the same subject pool, with the same testing procedure and materials, this combining of data across collection times was possible.

5.1. Method

5.1.1. Sample

We recruited an additional 150 adult subjects (age M = 28.83, SD = 6.24; range: 18 to 39 years) from prolific academic (75 male and 75 female). The sample was predominantly white (White = 85.3%; Asian = 7.3%; Black = 3.3%; South-East Asian = 0.7%; Other = 3.1%). Subjects were paid £1.20 for their participation (excluding bonus).

5.1.2. Materials

The materials were identical to those used in study 2 and 3. For the total sample, paper folding Form A coefficient ω = 0.72; Form B coefficient ω = 0.69; SOS-Effort pre-test coefficient ω = 0.83; SOS-Effort post-test coefficient ω = 0.73. The reliabilities for each experimental condition are reported in Table 4.

5.1.3. Procedure

The procedure was identical to that used for the experimental groups in study 2 and 3 except subjects received the following instruction about the possibility of earning bonus money for improving their performance: "*This next half block contains the same number of items, but this time, we will pay you a £5 bonus! if you improve your score by 1 item and £10 if you improve your score by 2 or more items on the next block compared the block you just completed.*"

5.2. Results

As can be seen in Table 4, the distribution of ability and effort scores were sufficiently normal for parametric analyses (e.g., skew < |2.0|; (Schmider et al., 2010). Furthermore, no outliers were identified, based on the outlier labelling rule with an inter-quartile range multiplier of 3.0 (Hoaglin & Iglewicz, 1987). Prior to conducting the 2 × 3 mixed-design ANOVA, the assumption of equality of covariance matrices across groups was tested and satisfied, Box's M test: F(6, 1,483,250.01) = 0.28, p = .946.

The mixed-design ANOVA identified a statistically significant main effect of time, F(1, 1348) = 73.20, p < .001, $\eta^2_{\text{partial}} = 0.052$, suggesting that, with the data collapsed across both groups, performance on the paper folding test increased from time 1 (M = 5.08; SD = 2.32) to time 2 (M = 5.60; SD = 2.04), i.e., the presence of at least a practice effect. By contrast, the group main effect was not significant statistically, F(2, 1348) = 1.02, p = .362, $\eta^2_{\text{partial}} = 0.002$. Finally, the time × group interaction was statistically significant, F(2, 1348) = 3.55, p = .029, $\eta^2_{\text{partial}} = 0.005$, suggesting that the degree of improvement in ability scores for at least one of the incentive groups was statistically significantly greater than the control group (see Table 4).

A follow-up 2 × 2 mixed-design ANOVA isolating the high-incentive group (£10) and the control group failed to yield a statistically significant interaction, F(1, 745) = 1.57, p = .211, $\eta^2_{\text{partial}} = 0.002$. Thus, the improvement in test performance associated with the high-incentive group (Hedges' g = −0.26, t(149) = −3.36, p < .001, g' = −0.37) was not, statistically significantly greater than the practice effect associated with the control group (Hedges' g = −0.17, t(596) = −4.90, p < .001, g'

Table 4
Mean, SD, Skew and Kurtosis for Ability and Effort Measures by Group and Time (Study 4).

	Time 1					Time 2				
	M	*SD*	skew	kurtosis	*ω*	*M*	*SD*	skew	kurtosis	*ω*
	Control Group (*N* = 597)									
Ability	5.08	2.32	−0.01	−0.71	0.72	5.44	2.06	−0.31	−0.52	0.68
Effort	4.20	0.59	−0.73	0.49	0.80	4.17	0.64	−0.88	1.33	0.82
	Incentive Group - £2 (*N* = 604)									
Ability	5.10	2.31	0.04	−0.73	0.71	5.74	2.00	−0.49	−0.13	0.68
Effort	4.11	0.62	−0.40	−0.25	0.81	4.26	0.58	−0.47	−0.36	0.81
	Incentive Group – £10 (*N* = 150)									
Ability	5.07	2.38	0.09	−0.77	0.72	5.64	2.05	−0.30	−0.06	0.69
Effort	4.13	0.67	−1.11	2.47	0.83	4.28	0.55	−0.34	−0.74	0.73

Note. Ability = performance on the paper folding test; Effort = test-taking effort.

$= -0.24$). By comparison, the 2 × 2 mixed-design ANOVA isolating the moderate-incentive group (£2) and the control group yielded a statistically significant interaction, $F(1, 1199) = 6.98$, $p = .008$, $\eta^2_{\text{partial}} = 0.006$; thus, combining the control groups from study 2 and study 3 increased the statistical power associated with the key interaction analysis carried out individually in studies 2 and 3. The magnitude of the difference in test performance improvement between the moderate-incentive group (Hedges' $g = -0.30$, $t(603) = -8.49$, $p < .001$, $g' = -0.43$) and the control group (Hedges' $g = -0.17$, $t(596) = -4.90$, $p < .001$, $g' = -0.24$) was equal to $\Delta g = -0.13$ or 1.95 IQ points. Disattenuated for imperfect reliability, the difference in Hedges' g' values corresponded to $\Delta g' = -0.19$ or 2.85 IQ points.

Finally, the 2 × 2 mixed-design ANOVA isolating the moderate-incentive group (£2) and the high-incentive group (£10) failed to yield a statistically significant interaction effect, suggesting a failure to identify a dose dependent effect (moderate incentive group Hedges' $g = -0.30$, $g' = -0.43$; high-incentive group Hedges' $g = -0.26$, $g' = -0.37$, $t(149) = -3.66$, $p < .001$).

Turning our attention to effort, a 2 × 2 mixed design ANOVA isolating the high-incentive group and the control group identified a statistically significant interaction, $F(1, 745) = 22.49$, $p < .001$, $\eta^2_{\text{partial}} = 0.029$, supporting the hypothesis that the £10 pound incentive increased test-taking effort (Hedges' $g = -0.25$, $t(149) = -3.98$, $p < .001$), in comparison to the control group (Hedges' $g = 0.04$, $t(596) = 1.45$, $p = .149$). By contrast, isolating the high-incentive group and the moderate-incentive group, the 2 × 2 mixed-design ANOVA failed to reach statistical significance, $F(1, 752) = 0.01$, $p = .920$, $\eta^2_{\text{partial}} < 0.001$, suggesting that the magnitude of the increase in effort from time 1 to time 2 was not statistically significantly different across the two incentive groups (high-incentive group Hedges' $g = -0.25$, $t = -3.98$, $p < .001$; moderate-incentive group Hedges' $g = -0.25$, $t = -8.69$, $p < .001$).

5.3. Discussion

By combining samples across studies (control $N = 597$; inventive $N = 604$), the hypothesis that test performance would be impacted positively by motivation was supported. In partial η^2 terms, the magnitude of the effect (≈ 0.005) is unequivocally small (Richardson, 2011). Stated alternatively, the control versus incentive group difference at time 2 corresponded to a standardized effect (Hedges' g) of approximately 0.13, a value that corresponds in conventional IQ points to an effect of 1.95 points (using z-score transform). Ultimately, on average, the incentivised groups managed to increase their performance by only ¼ of one question (5.69–5.44 = 0.25).

The £10 bonus incentive increased effort. Despite the size of the bonus, in standardized effect size terms, the magnitude of the effect (partial $\eta^2 = 0.029$; Hedges' $g \approx 0.15$) is modest relative to the effects effort has been expected to show. Gignac (2018) found a similar numerical mean difference in the hypothesized direction (Hedges' $g \approx 0.15$) but this was non-significant. In absolute terms the SOS-Effort mean change from 4.13 to 4.28 may not be considered large (3.6% increase). No dose-response effect was observed with respect to effort: those offered the opportunity to earn an additional £10 reported, on average, reported the same level of effort as those offered the opportunity to earn an additional £2 (4.28 vs. 4.26).

6. General discussion

Our results agree well with recent empirical investigations into the experimental effects of motivation on cognitive ability test performance. For example, Gignac (2018) failed to observe statistically significant effects ($N = 99$), and on average, across the five processing speed tasks that were analysed, the reported effect sizes were not large (partial $\eta^2 \approx 0.016$). Similarly, Merritt et al. (2019) failed to observe a statistically significant effect of incentive on test performance in a between-subjects design ($N = 42$ and $N = 39$), and the average effect size across the four cognitive ability tests (visual memory, verbal memory, visual motor speed and reaction time) corresponded to $g \approx 0.23$ (in favour of the incentive group). Taken together, our results, and the most recent experimental research in the area, suggest that the causal effect of motivation on test performance is likely small, at least in healthy adult volunteers.

We acknowledge that some recent experimental investigations have reported statistically significant and substantial effects of motivation on test performance. For example, in a sample of university students who completed a battery of academic achievement tests, Liu, Rios, and Borden (2015) reported incentive effects of approximately $g = 0.55$ to 0.75. However, once the subjects who spent less than 10% of average time to complete 10% or more of the test items were removed from the sample ("rapid responders"), no statistically significant incentive effects were found. Arguably, the rapid responders removed from the Liu et al. (2015) sample were not only non-motivated, but they were also not even paying much attention to the test items. Correspondingly, we note that the Duckworth et al. (2011) meta-analysis included many neurotypical non-normal samples (e.g., behavioural and learning difficulties) which may have included subjects who barely attended to the test items. The results of our investigation suggest that if a person attends to the test items and applies some motivation to complete them, they will likely manifest a test performance essentially commensurate with their true intelligence level. Once a modicum of effort is applied, the prospect of increasing motivation to improve test performance is diminished greatly. Thus, it may be concluded that intelligence test scores, in most scenarios, may be interpreted validly, when the test manual instructions are followed. Such a conclusion is in clear contrast to the conclusion by Duckworth et al. (2011).

6.1. Correlation and direction of causality

Across all three samples and cognitive ability tests (sentence verification, vocabulary, visual-spatial reasoning), the magnitude of the association between effort and test performance was approximately 0.30, suggesting that higher levels of motivation are associated better levels of test performance. Our results are in close accord with existing literature, including a recent meta-analysis in the area ($r = 0.33$; Silm et al., 2020). A unique contribution of our investigation are the large sample sizes (Ns ≈ 1000) and the fact that the subjects completed only a single test, prior to self-reported effort. With the large sample sizes, we were also in the unique position to examine the nature of the association, and we found the association to be clearly linear. Gignac (2018) and (Gignac et al., 2019) suggested that the association might be non-linear, such that a more pronounced effect of motivation may be observed at the lower-end of the effort spectrum. However, we failed to detect any suggestion of such a non-linear effect. The presence of an entirely linear effect may be suggested to be indirectly supportive of the notion that reported effort largely reflects perceived ability, as the observation of a threshold effect would be inconsistent with the effort as perceived ability hypothesis.

As is well-known, the observation of a correlation is a necessary but not sufficient condition for causality. The failure to observe concomitant increases in test effort and test performance, when test effort is manipulated, suggests the absence of a causal effect between test motivation and test performance. Consequently, the positive linear assocation between effort and performance may be considered either spurious or the direction of causation reversed – flowing from ability to motivation. Several investigations have shown that the correlation between test-taking anxiety and test performance likely flows from ability to test-anxiety, not the other way around (Sommer & Arendasy, 2015; Sommer, Arendasy, Punter, Feldhammer-Kahr, & Rieder, 2019). Thus, if the direction of causation flows from ability to test motivation, it would help explain why effort is so difficult to shift via incentive manipulation.

6.2. Limitations & future research

We acknowledge that the evidence for the causal direction between effort and ability remains equivocal, as our evidence rests upon the absence of evidence (absence of experimental incentive effect). Ideally, positive evidence would be provided. Indirect positive evidence may be obtained by conducting an experiment, whereby half the subjects are given a relatively easy version of the paper folding task (10 easiest items) and the other half are given a relatively more difficult version (10 most difficult items). It is hypothesized that those given the relatively easier version of the paper folding task would then, on average, self-report greater levels of test-taking effort. Partial support for such a hypothesis is apparent in Table 1 of this investigation. Specifically, it can be seen that there is a perfect correspondence between the difficulty of the test (synonyms mean 73.4% correct; sentence verification mean 53.8% correct; paper folding mean 43.3%) and the mean level of reported effort (synonyms mean effort 4.42; sentence verification mean 4.11; paper folding mean 3.83).

Another limitation is that we measured effort with only one method: self-report. Self-report measures have limitations, including that they rely upon the introspection ability of the responder (Paulhus & Vazire, 2007). Additionally, the degree of convergent validity between self-report measures of effort and behavioural measures of effort is only moderate ($r \sim 0.25$; Wise & Kong, 2005). However, behavioural measures of effort have their own limitations (Gignac & Wong, 2020). Perhaps, ideally, effort would be measured with a multi-method approach, a strategy we encourage for future research. In light of our use of a single (and imperfect) method to measure effort, the degree to which our financial incentives increased effort was likely underestimated across studies 2 to 4.

We also acknowledge that we manipulated effort with only one incentive: money. Financial incentives have been shown to increase effort (Bonner & Sprinkle, 2002), however, it is only one type of incentive. Other types of incentives might show larger effect sizes than those reported in this investigation. Consequently, our results are accurate to the degree that alternative incentives do not yield appreciably larger effects. Finally, situations with higher workload, more fatigue, etc., effort may be more (or less) important, although the existing studies with longer testing sessions (e.g. Gignac, 2018) failed to show such larger effects.

7. Conclusion

There is almost undoubtedly a positive correlation between reported effort and IQ-type test scores, and the magnitude is likely $r \approx 0.30$. However, the accumulating experimental evidence with neurotypical adult volunteers suggests that the correlation does not reflect a substantive causal effect, at least not one that leads from effort to test performance.

Data and analysis scripts

Data and analysis scripts are available at: https://osf.io/5uesw/?view_only=705617acaf734286844b1521ed87afdc

References

Baddeley, A. D. (1968). A 3 min reasoning test based on grammatical transformation. *Psychonomic Science, 10*, 341–342.

Binet, A. (1905/1916). New methods for the diagnosis of the intellectual level of subnormals. L'Année Psychology, 12, 191-244. (E. S. Kite, Trans.). In *The development of intelligence in children*. Vineland, NJ: Publications of the Training School at Vineland.

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of d for range restriction and unreliability. *Organizational Research Methods, 4*(1), 46–61.

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*(4–5), 303–345. https://doi.org/10.1016/s0361-3682(01)00052-6

Breuning, S. E., & Zella, W. F. (1978). Effects of individualized incentives on norm-referenced IQ test performance of high school students in special education classes. *Journal of School Psychology, 16*(3), 220–226. https://doi.org/10.1016/0022-4405(78)90004-3

Caldwell, A. R., Lakens, D., & Parlett-Pelleriti, C. M. (2021). Power analysis with superpower. Retrieved from https://aaroncaldwell.us/SuperpowerBook/.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829–836.

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*(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Demange, P. A., Malanchini, M., Mallard, T. T., Biroli, P., Cox, S. R., Grotzinger, A. D., … Nivard, M. G. (2021). Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction. *Nature Genetics, 53*(1), 35–44. https://doi.org/10.1038/s41588-020-00754-2

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 of the United States of America, 108*(19), 7716–7720. https://doi.org/10.1073/pnas.1018601108

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests*. Princeton, New Jersey: ETS Research and Development.

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*(9), 1058–1065. https://doi.org/10.1001/jamapsychiatry.2014.1105

Flather, M. D., Farkouh, M. E., Pogue, J. M., & Yusuf, S. (1997). Strengths and limitations of meta-analysis: Larger studies may be more reliable. *Controlled Clinical Trials, 18*(6), 568–579. https://doi.org/10.1016/s0197-2456(97)00024-x

Garcia, E. (2014). *The need to address noncognitive skills in the education policy agenda* (pp. 1–28).

Gignac, G. E. (2018). A moderate financial incentive can increase effort, but not intelligence test performance in adult volunteers. *British Journal of Psychology, 109*(3), 500–516. https://doi.org/10.1111/bjop.12288

Gignac, G. E., Bartulovich, A., & Salleo, E. (2019). Maximum effort may not be required for valid intelligence test score interpretations. *Intelligence, 75*, 73–84. https://doi.org/10.1016/j.intell.2019.04.007

Gignac, G. E., & Wong, K. K. (2020). A psychometric examination of the anagram persistence task: More than two unsolvable anagrams may not be better. *Assessment, 27*(6), 1198–1212. https://doi.org/10.1177/1073191118789260

Haynes, D. (1988). The integrity of research published by Stephen E. Breuning. *Bulletin of the Medical Library Association, 76*(3), 272.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association, 82*(400), 1147–1149. https://doi.org/10.1080/01621459.1987.10478551

Ioannidis, J. P. A., & Lau, J. (1999). Pooling research results: Benefits and limitations of meta-analysis. *The Joint Commission Journal on Quality Improvement, 25*(9), 462–469. https://doi.org/10.1016/s1070-3241(16)30460-6

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). *Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success* (Vol. 110). OECD Publishing.

Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology, 54*, 1–20.

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.

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920951503

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

Merritt, V. C., Rabinowitz, A. R., Guty, E., Meyer, J. E., Greenberg, L. S., & Arnett, P. A. (2019). Financial incentives influence ImPACT validity indices but not cognitive composite scores. *Journal of Clinical and Experimental Neuropsychology, 41*(3), 312–319. https://doi.org/10.1080/13803395.2018.1551519

O'Neil, J. H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the national assessment of educational progress mathematics performance. *Educational Assessment, 3*(2), 135–157. https://doi.org/10.1207/s15326977ea0302_2

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology*. New York: Guilford.

Penk, C., & Richter, D. (2016). 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

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? *Methodology, 6*(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review, 31*. https://doi.org/10.1016/j.edurev.2020.100335

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/10.1016/j.intell.2015.08.007

Sommer, M., Arendasy, M. E., Punter, J. F., Feldhammer-Kahr, M., & Rieder, A. (2019). Do individual differences in test-takers' appraisal of admission testing compromise measurement fairness? *Intelligence, 73*, 16–29. https://doi.org/10.1016/j.intell.2019.01.006

Sundre, D. L., & Thelk, A. D. (2007). *The student opinion scale (SOS): A measure of examinee motivation: Test manual*.

Warne, R. T. (2021). No strong evidence of stereotype threat in females: A reassessment of the meta-analysis. *Journal of Advanced Academics.*. https://doi.org/10.1177/1932202x211061517

Warne, R. T. (2022). More articles by Stephen Breuning that need retraction. Retrieved from https://russellwarne.com/2022/02/22/more-articles-by-stephen-breuning-that-need-retraction/.

Warrington, E. K., McKenna, P., & Orpwood, L. (1998). Single word comprehension: A concrete and abstract word synonym test. *Neuropsychological Rehabilitation, 8*(2), 143–154. https://doi.org/10.1080/713755564

Wechsler, D. (2008). *Wechsler adult intelligence scale: Technical and interpretive manual* (4 ed.). San Antonio, Tx: Psychological Corporation.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Zuo, L., Zhang, C. K., Wang, F., Li, C. S., Zhao, H., Lu, L., … Luo, X. (2011). A novel, functional and replicable risk gene region for alcohol dependence identified by genome-wide association study. *PLoS One, 6*(11), Article e26726. https://doi.org/10.1371/journal.pone.0026726