LETTERS
https://doi.org/10.1038/s41562-018-0509-y
nature human behaviour

Intelligent problem-solvers externalize cognitive operations

Bruno R. Bocanegra[1,2*], Fenna H. Poletiek[2,3], Bouchra Ftitache[4] and Andy Clark[5]

Humans are nature's most intelligent and prolific users of external props and aids (such as written texts, slide-rules and software packages). Here we introduce a method for investigating how people make active use of their task environment during problem-solving and apply this approach to the non-verbal Raven Advanced Progressive Matrices test for fluid intelligence. We designed a click-and-drag version of the Raven test in which participants could create different external spatial configurations while solving the puzzles. In our first study, we observed that the click-and-drag test was better than the conventional static test at predicting academic achievement of university students. This pattern of results was partially replicated in a novel sample. Importantly, environment-altering actions were clustered in between periods of apparent inactivity, suggesting that problem-solvers were delicately balancing the execution of internal and external cognitive operations. We observed a systematic relationship between this critical phasic temporal signature and improved test performance. Our approach is widely applicable and offers an opportunity to quantitatively assess a powerful, although understudied, feature of human intelligence: our ability to use external objects, props and aids to solve complex problems.

Intelligence shows consistent and strong associations with important life outcomes such as academic and occupational achievement, social mobility and health[1,2]. Over the past decades, great advances have been made by investigating intelligence in terms of the encoding, maintenance and manipulation of internal mental representations, most notably in working memory[3–15]. However, real-world problems regularly exceed the capacity of working memory and require people to offload memory and intermediate processing onto the environment. Whether it is a scientist composing and rearranging equations and diagrams on a blackboard or a hunter-gatherer planning a hunting strategy by positioning and re-positioning place-holder objects in the sand, many theorists have argued that understanding the full breadth of human intellectual performance depends on extending our focus to encompass the storage and manipulation of external information[16–21].

Humans routinely use their environment when solving problems that require complex inferences[22–25]. For example, a police investigator may use an evidence board to solve a criminal case. After an initial look, she generates a first interpretation of the evidence. This interpretation may trigger her to reconfigure the evidence board according to this initial hypothesis. Subsequent inspection of this new configuration may then lead her—even in the absence of new evidence—to a novel interpretation and another reconfiguration of the board and so on[22]. Another example is a scientist trying to write a paper. She begins by looking over some old notes and original sources. While reading, she comes up with a preliminary outline for the paper, which is externalized using highlights, notes and textual operations. The reconfigured task environment then triggers a more refined conceptual structure and the cycle repeats[25]. In both cases, problem-solvers externalize (partial) solutions to the problem and reflect on them. The environment is used as an external working memory which unburdens internal processing resources and allows increasingly complex inferences to be made. We are so accustomed to these cognitively potent loops into the world that we may not realize how strange they are. Existing artificial intelligence programs never proceed by printing out intermediate results to repeatedly re-inspect them. Yet we humans have developed an adaptive form of fluid intelligence that relies heavily on this trick.

Although external cognitive operations have recently been investigated in perception, attention, memory, numerical and spatial cognition[26–33], so far, they remain relatively unexplored in fluid intelligence[34]. To address this, we designed a click-and-drag version of one of the most common and popular intelligence quotient (IQ) tests across the lifespan: the non-verbal Raven Advanced Progressive Matrices test for fluid intelligence[26] (Fig. 1b). In this complex problem-solving task, participants compare and contrast figures within a spatial array to infer a missing figure (Fig. 1a). The high complexity of the array precludes participants from solving items in a single glance. Instead, they have to actively inspect different (subsets of) figures, each of which will highlight different emergent perceptual patterns[35]. Our objective was to examine the externalization of cognitive operations by measuring participants' active manipulation of the layout of items while attempting to solve them.

To verify that performance in this click-and-drag Raven test would reflect general cognitive ability[1], we first assessed the test's ability to predict academic achievement, compared to the conventional static Raven test. In Experiment 1a, we tested a sample of 211 university students. Planned contrasts indicated a medium-to-large positive correlation between Raven accuracy and academic achievement in the click-and-drag test (correlation coefficient $r(101) = 0.46$, $P < 0.001$, 95% confidence interval (CI) = [0.29, 0.60]), and a small-to-medium positive correlation in the static test ($r(106) = 0.20$, $P = 0.038$, 95% CI = [0.01, 0.37]). The correlation was stronger in the click-and-drag test compared to the static test when analysed by Fisher's r-to-z transformation (difference in correlation coefficients $r_{diff} = 0.26$, r-to-z transformation of the difference in correlation coefficients $z = 2.11$, $P = 0.035$, 95% CI = [0.02, 0.51]). In addition, a regression analysis indicated a significant interaction between Raven-type and Raven accuracy on academic achievement ($t(209) = 2.08$, $P = 0.038$, unstandardized regression

[1]Department of Psychology, Educational, and Child Studies, Erasmus University Rotterdam, Rotterdam, the Netherlands. [2]Institute of Psychology, Leiden University, Leiden, the Netherlands. [3]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands. [4]Institute for Mental Health Care GGZ Rivierduinen, Leiden, the Netherlands. [5]School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, UK. *e-mail: bocanegra@essb.eur.nl

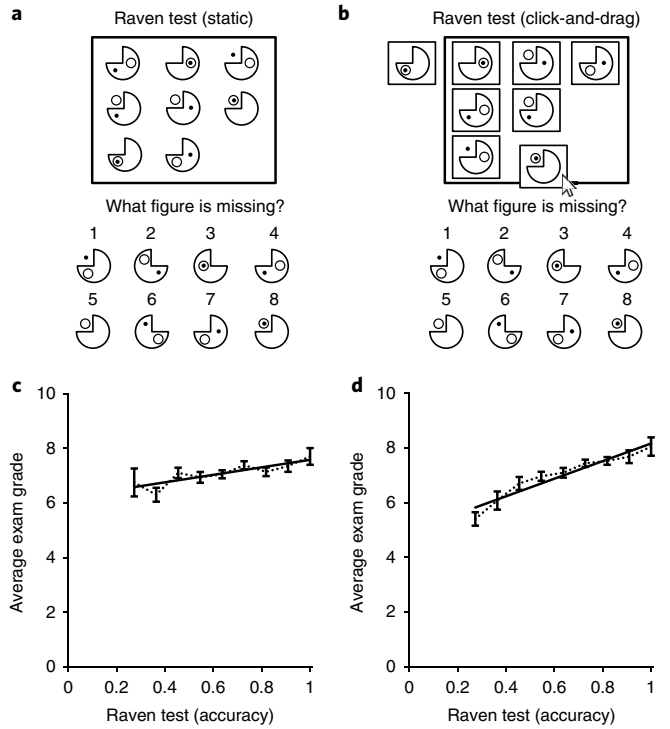

Fig. 1 | Predicting academic achievement using the conventional and the adapted click-and-drag Raven Advanced Progressive Matrices test in Experiments 1a and 1b ($n = 495$). a, Conventional IQ test item in the style of the Raven Advanced Progressive Matrices. **b**, Adapted click-and-drag Raven IQ test item. **c**,**d**, Average exam grades for performance levels (accuracy) in Experiments 1a and 1b for the static Raven test ($n = 251$) (**c**) and the click-and-drag Raven test ($n = 244$) (**d**). Error bars represent the mean ± s.e.m.

coefficient $b = 0.16$, standard error of unstandardized regression coefficient $SE_b = 0.08$, standardized regression coefficient $\beta = 0.14$, 95% CI = [0.01, 0.31]), indicating that the click-and-drag Raven was a stronger predictor of academic achievement ($t(101) = 5.15$, $P < 0.001$, $b = 2.88$, $SE_b = 0.56$, $\beta = 0.46$, 95% CI = [1.77, 3.99]), compared to the static Raven ($t(106) = 2.10$, $P = 0.038$, $b = 1.64$, $SE_b = 0.78$, $\beta = 0.20$, 95% CI = [0.09, 3.18]). In Experiment 1b, we performed a replication of the two Raven conditions in a sample of 284 students from a new cohort: we observed a medium-to-large positive correlation in the click-and-drag test ($r(139) = 0.37$, $P < 0.001$, 95% CI = [0.22, 0.50]) and a non-significant small-to-medium positive correlation in the static test ($r(141) = 0.16$, $P = 0.052$, 95% CI = [−0.001, 0.32]). Although the correlation was numerically larger in the click-and-drag test compared to the static test, the contrast between the correlations failed to reach a conventional level of significance when analysed by Fisher's r-to-z transformation, ($r_{diff} = 0.21$, $z = 1.92$, $P = 0.054$, 95% CI = [−0.003, 0.44]). However, a regression analysis indicated a significant interaction between Raven-type and Raven accuracy on academic achievement ($t(283) = 2.35$, $P = 0.019$, $b = 0.12$, $SE_b = 0.05$, $\beta = 0.14$, 95% CI = [0.02, 0.23]), indicating that the click-and-drag Raven was a stronger predictor of academic achievement ($t(139) = 4.76$, $P < 0.001$, $b = 2.37$, $SE_b = 0.50$, $\beta = 0.37$, 95% CI = [1.39, 3.35]), as compared to the static Raven task ($t(141) = 1.96$, $P = 0.052$, $b = 0.84$, $SE_b = 0.43$, $\beta = 0.16$, 95% CI = [−0.008, 1.69]). Given that the P value of the difference between the Fisher r-to-z transformed correlations did not reach conventional levels of significance but the P value of the interaction effect between Raven-type and Raven accuracy did reach conventional levels of significance, we consider Experiment 1b to have partially replicated the pattern of results observed in Experiment 1a. Pooling the two experiments for increased power, we observed a larger correlation in the click-and-drag test ($r(242) = 0.43$, $P < 0.001$, 95% CI = [0.32, 0.53], Fig. 1d), compared to the static test, ($r(249) = 0.18$, $P = 0.004$, 95% CI = [0.06, 0.30], Fig. 1c). The correlation was stronger in the click-and-drag test compared to the static test when analysed by Fisher's r-to-z transformation ($r_{diff} = 0.25$, $z = 3.08$, $P = 0.002$, 95% CI = [0.10, 0.43]). Finally, a regression analysis indicated a significant interaction between Raven-type and Raven accuracy on academic achievement ($t(494) = 3.27$, $P = 0.001$, $b = 0.16$, $SE_b = 0.05$, $\beta = 0.15$, 95% CI = [0.07, 0.26]), indicating that the more naturalistic click-and-drag Raven was a stronger predictor of academic achievement ($t(242) = 7.37$, $P < 0.001$, $b = 2.77$, $SE_b = 0.38$, $\beta = 0.43$, 95% CI = [2.03, 3.51]), compared to the static Raven task ($t(249) = 2.87$, $P = 0.004$, $b = 1.16$, $SE_b = 0.40$, $\beta = 0.18$, 95% CI = [0.36, 1.95], (see Supplementary Information for additional analyses).

Experiments 1a and 1b suggest that the click-and-drag version of the Raven might be tapping into an additional behavioural aspect of intelligence that is not currently measured in the conventional static Raven. One possibility is that participants in the click-and-drag Raven are using their task environment to externalize cognitive operations that would otherwise be performed internally in working memory. To investigate this, we tested a new sample of 70 participants in Experiment 2, with the aim of measuring in detail the extent to which participants in the click-and-drag test were making active use of the task environment during problem-solving. To do this, we focused on the temporal distribution of executed actions during the entire task. Our rationale was that, if cognitive operations are being externalized, changes made to the external layout should guide how figures are being compared and contrasted immediately after that change. For example, a participant may initially hypothesize a relationship between the figures. This may trigger actions, which change the layout, which itself triggers a new hypothesis and more subsequent actions. If there is periodic coupling between action-induced changes in the environment and environment-induced triggers of action, actions should cluster together in between periods of inactivity. However, if actions are performed independently of the changes they produce in the environment, actions should be uncorrelated and evenly distributed over time.

To illustrate how to quantify the externalization of cognitive operations, we simulated action sequences for an idealized dual-mode and single-mode problem-solver (the number of discrete temporal intervals was $T = 3 \times 10^5$ for each simulated problem-solver; see Supplementary Information). A dual-mode problem-solver uses a queuing procedure to go back-and-forth between an external mode where cognitive operations are externalized on the screen and an internal mode where cognitive operations are performed internally (see Fig. 2a). The idea is that a dual-mode problem-solver is switching between externally projecting the outcome of previously generated internal evaluations and internally evaluating the outcome of previously executed external actions. On the other hand, a single-mode problem-solver executes a single type of cognitive operation in the absence of competitive queuing (Fig. 2b). In other words, a single-mode problem-solver does not perform external projections of generated ideas nor internal evaluations of executed actions. As a consequence, there is no interaction between the two modes and therefore no clear distinction between them. Importantly, single-mode versus dual-mode problem-solving is not an all-or-nothing dichotomy but rather a gradual distinction. A dual-mode problem-solver simulates a strong coupling between internal and external operations in the sense that the outcome of the external operations provide the input to the internal operations and vice versa, whereas a single-mode problem-solver simulates the situation when internal and external operations are decoupled. Because external operations are executed independently of internal operations (and vice

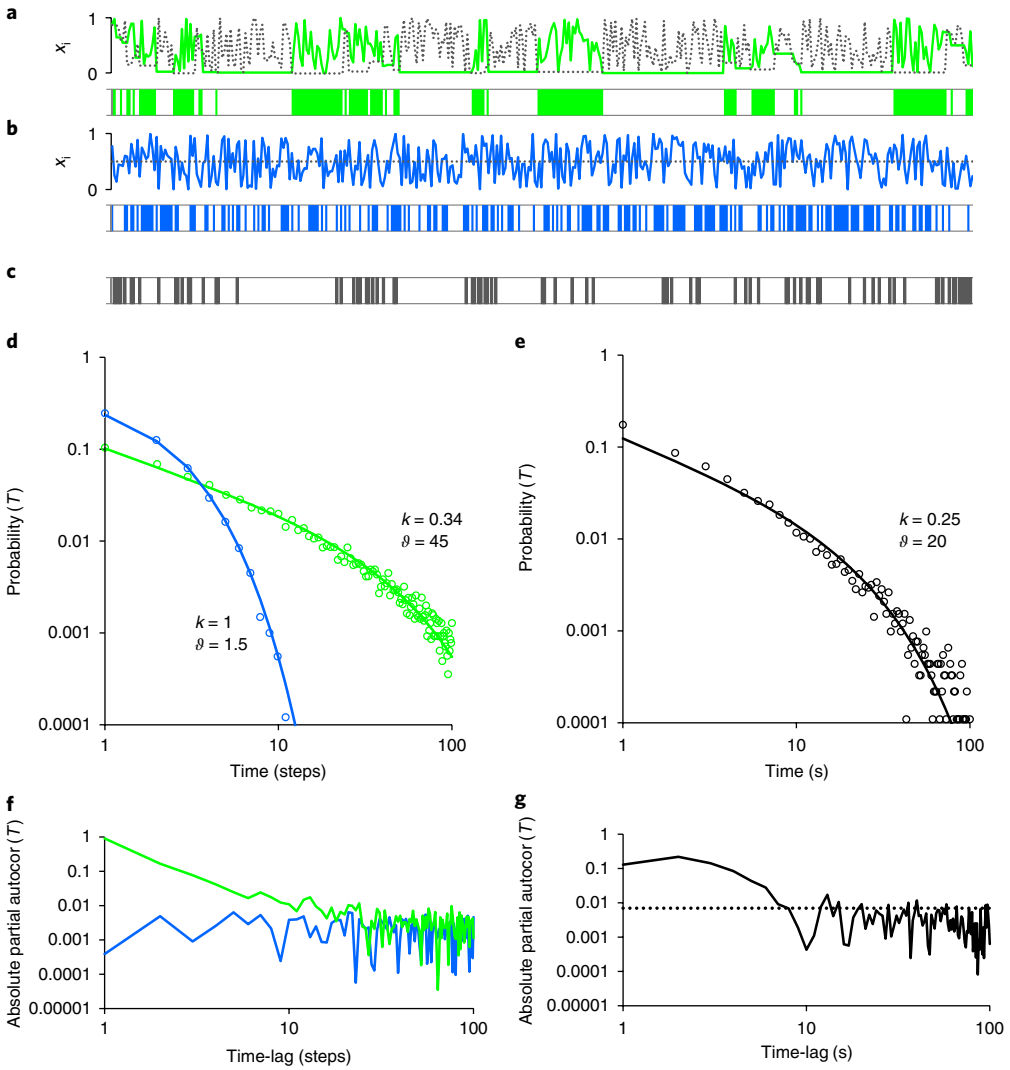

Fig. 2 | Simulated data for the dual-mode (green) and single-mode model (blue) and empirical data for experimental participants (black) in Experiment 2 ($n = 70$). **a**, Time course of the dual-mode priority parameters $x_i \in [0,1]$ for external operations (solid green line) and internal operations (dashed grey line) and the resulting action intervals (green bars) and rest intervals (white bars). **b**, Time course of the single-mode action parameter $x_i \in [0,1]$ (solid blue line) and the action threshold value (dashed grey line) and the resulting action intervals (blue bars) and rest intervals (white bars). **c**, sample of action intervals (dark grey bars) and rest intervals (white bars) from participants' experimental data. This sample was selected visually to represent the typical degree of temporal clustering observed in our dataset. **d**,**e**, Probability distribution of rest intervals (open circles) and gamma distribution functions (solid lines) for the dual-mode model (green) and single-mode model (blue, $T = 3\times10^5$ simulated intervals per model) (**d**) and the experimental data (black, $n = 70$, $T = 7.1\times10^4$ intervals in total) (**e**). **f**,**g**, Partial autocorrelation function (absolute coefficients) for the dual-mode model (green) (**f**) and single-mode model (blue) and the experimental participants (black, dashed line indicates the upper-bound of the 95% CI for uncorrelated temporal intervals) (**g**).

versa), they cannot be regarded as separate processing modes, which is functionally equivalent to a single-mode of processing (see Supplementary Information and Supplementary Fig. 8 for additional simulations).

As demonstrated previously[36], balancing the execution of two distinct processing modes should result in a heavy tailed probability distribution of temporal intervals between consecutive actions that approximates $P(T) \approx T^{-1}$, whereas executing a single processing mode should show an exponential distribution $P(T) \approx e^{-T}$. These distributions are markedly different: the latter distribution decays rapidly, indicating that actions are executed at fairly regular intervals, whereas the former distribution decays slowly, allowing for clusters of actions that are separated by longer intervals[36]. To differentiate these temporal signatures we fit two-parameter gamma distribution functions with shape parameter k and scale parameter θ to the distribution of rest intervals between actions:

$$P(t) = \frac{1}{\Gamma(k)\theta^k} t^{k-1}\, e^{-\frac{t}{\theta}} \text{ with mean } \mu = k\theta \qquad (1)$$

Please note in equation (1) that when the shape parameter is equal to one ($k = 1$) and the scale parameter is equal to the mean ($\theta = \mu$), the distribution will be exponential $P(t) = \frac{1}{\theta} e^{-\frac{1}{\theta}t}$, indicating that actions are uncorrelated. However, when the shape parameter is smaller than one ($k < 1$) and the scale parameter is larger than the mean ($\theta > \mu$), the gamma distribution will show a heavier tail and approximate $P(t) \approx kt^{k-1}$, indicating correlated actions. As can be seen in Fig. 2d, a simulated single-mode problem-solver (blue) produces an exponential distribution ($k = 1.0$, $\theta = 1.5$, mean number of consecutive rest intervals $\bar{x} = 1.51$), whereas a simulated dual-mode problem-solver (green) produces a heavy tailed distribution ($k = 0.34$, $\theta = 54$, $\bar{x} = 18.26$), indicating that the balancing of external and internal cognitive operations results in periods of action that

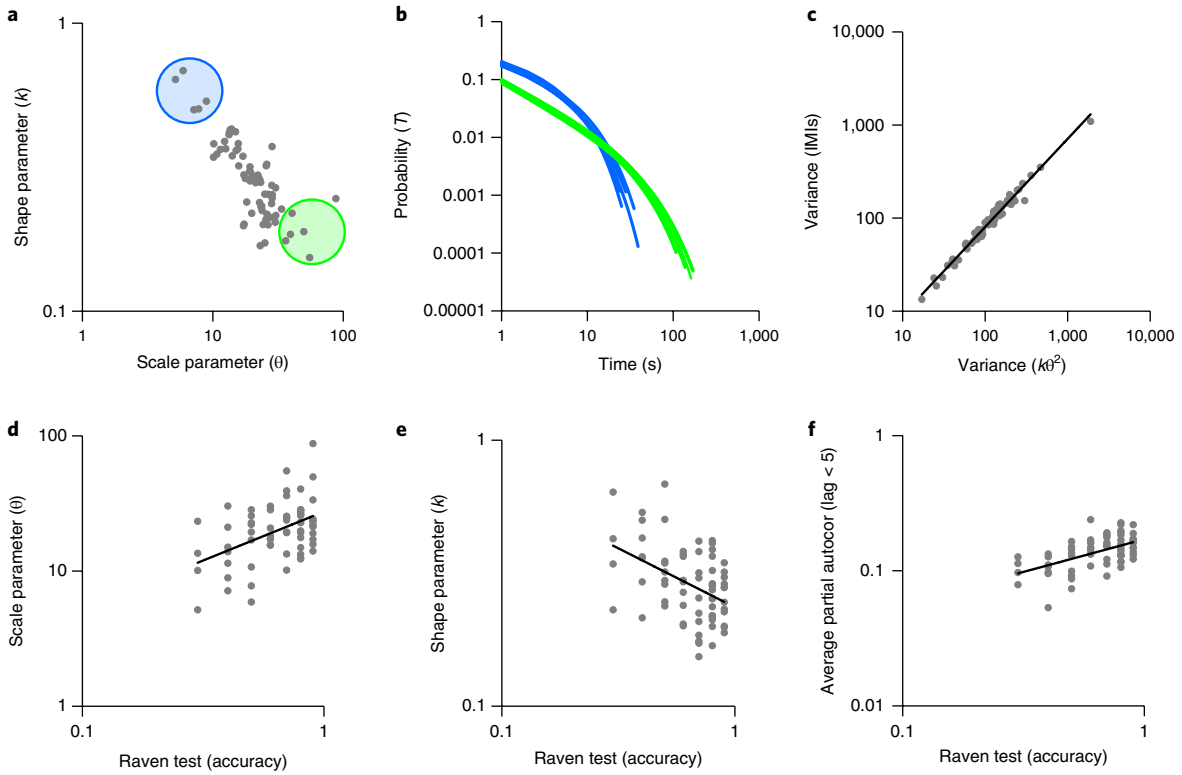

Fig. 3 | Shape parameters, scale parameters, partial autocorrelations as a function of Raven IQ test performance in Experiment 2 ($n = 70$). **a**, Shape and scale parameters for individual participants in Experiment 2. **b**, Rest-interval distributions for two sets of five participants at the ends of the correlated scale–shape spectrum (see green and blue selection in **a**). **c**, Individual differences in variance observed in intermovement intervals (IMIs) as a function of individual differences in variance described by shape and scale parameters. **d**, Shape parameters. **e**, Scale parameters. **f**, Average partial autocorrelations (for lags <5) as a function of Raven test accuracy.

are clustered in between periods of inactivity. This phasic temporal signature can also be observed in the partial autocorrelation function (Fig. 2f), where a dual-mode problem-solver showed correlations for the first ten time-lags, which are absent in a single-mode problem-solver.

How did actual participants perform the task? A representative example is displayed in Fig. 2c. The two-parameter gamma distribution function fit on the aggregated data of all participants showed a heavy tailed distribution of rest intervals ($k = 0.25$, $\theta = 20$, $\bar{x} = 5.61$; Fig. 2e), suggesting that actions were correlated. Indeed, the partial autocorrelation function showed significant correlations for the first six time-lags ($ts > 7$, $Ps < 0.001$; Fig. 2g). Parameter estimates for individual participants confirmed this result: one-sample t-tests indicated that shape parameters (k) for individual participants were significantly smaller than 1, $k_{mean} = 0.29$, $t(69) = 32.81$, $P < 0.001$, 95% CI = [0.27, 0.31] and scale parameters (θ) were significantly larger than the mean $\bar{x} = 5.61$, $\theta_{mean} = 19.93$, $t(69) = 21.51$, $P < 0.001$, 95% CI = [17.72, 22.42]. In addition, the variation in scale and shape parameters revealed large individual differences (Fig. 3a,b), ranging from heavier tailed (green) to more exponentially shaped distributions (blue). Consistent with this, we observed large individual differences in the variance of time intervals between actions (intermovement intervals) and that these individual differences in variances could be accounted for by individual differences in the shape and scale parameters: a simple regression analysis indicated that individual differences in variance observed in the intermovement intervals increased as a function of the individual differences in variance as described by the shape and scale parameters $k\theta^2$ ($t(68) = 55.52$, $P < 0.001$, $b = 0.95$, $SE_b = 0.02$, $\beta = 0.99$, 95% CI = [0.91, 0.98]; Fig. 3c). Importantly, this indicates that the scale and shape of individual distributions were able to capture different strategies used to execute the problem-solving task.

To establish that the execution of external operations was playing a positive cognitive role during problem-solving, we tested whether temporally clustered actions were related to improved test performance by examining shape parameters, scale parameters and average partial autocorrelations (for lags <5) for individual participants. Consistent with our expectations, simple regression analyses indicated that scale parameters increased ($t(68) = 4.28$, $P < 0.001$, $b = 0.72$, $SE_b = 0.17$, $\beta = 0.46$, 95% CI = [0.39, 1.06]), shape parameters decreased ($t(68) = 4.01$, $P < 0.001$, $b = -0.44$, $SE_b = 0.11$, $\beta = -0.44$, 95% CI = [−0.66,−0.22]) and autocorrelations increased ($t(68) = 5.42$, $P < 0.001$, $b = 0.49$, $SE_b = 0.09$, $\beta = 0.55$, 95% CI = [0.31, 0.66]), as a function of Raven accuracy (Fig. 3d–f). This specific pattern of results demonstrates that phasic temporal signatures were indicative of successful problem-solving.

To exclude the possibility that our results were an artefact of the analysis, we examined how the variance of intermovement intervals (that is calculated using unprocessed time stamps) varied with Raven performance. The more evenly spread out actions are over time, the smaller the variance of intermovement intervals. Therefore, if correlated actions are indeed indicative of successful problem-solving, variance should increase as a function of Raven accuracy. A simple regression analysis indicated that variance increased as a function of accuracy ($t(68) = 3.61$, $P = 0.001$, $b = 0.92$, $SE_b = 0.26$, $\beta = 0.40$, 95% CI = [0.41, 1.43]; Fig. 4a), indicating that the systematic relation we observed between phasic task activity and task performance did not depend on our particular analysis.

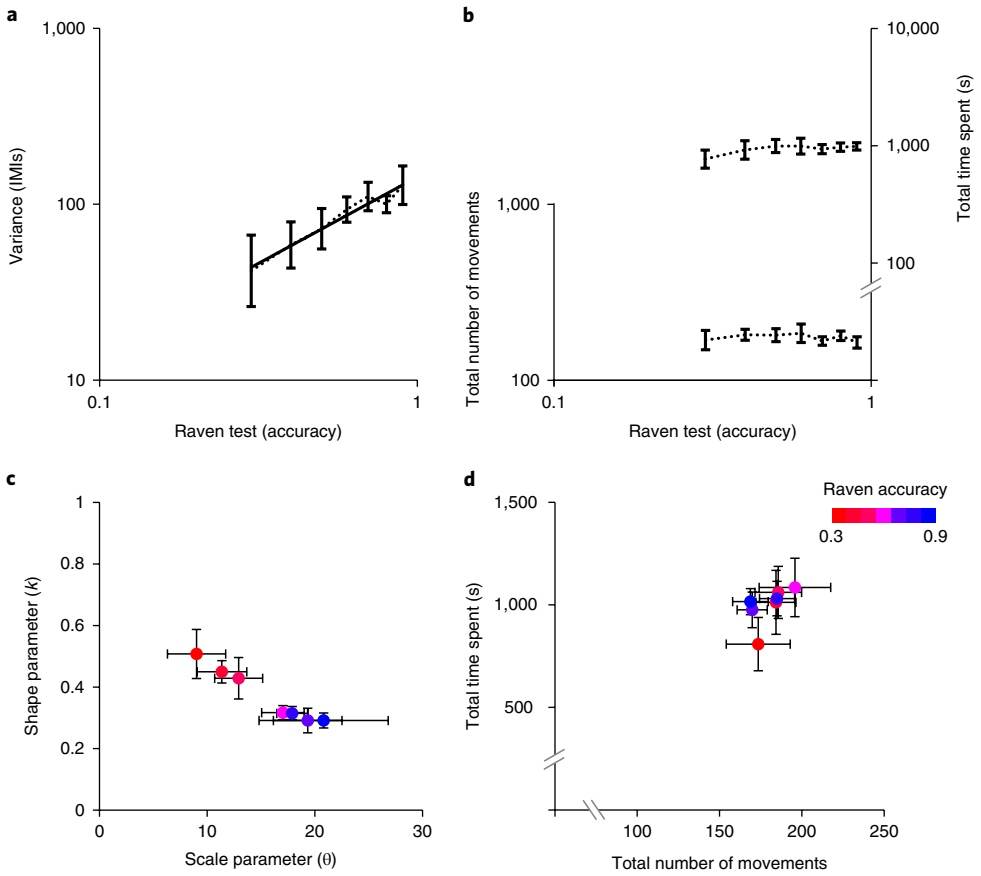

Fig. 4 | Variance of intermovement intervals, total number of movements, total time spent on task as a function of Raven IQ test performance in Experiment 2 ($n = 70$). a, Geometric mean variance of intermovement intervals (IMIs). **b**, total number of movements and time spent as a function of Raven accuracy in the click-and-drag Raven test. **c**,**d**, Error bars represent the mean ± s.e.m. Mean performance levels (Raven accuracy) as a function of scale and shape parameters (**c**) and the number of movements and time spent (**d**). Error bars represent mean ± s.e.m.

Did participants who performed poorly simply lack the motivation to engage with the task (not performing enough actions) or did they give up too soon (not spending enough time on the task)? Our results do not support these explanations: simple regression analyses did not indicate that the total number of actions executed ($t(68) = 0.51$, $P = 0.61$, $b = -0.05$, $SE_b = 0.10$, $\beta = -0.06$, 95% CI = [−0.24, 0.14]) or the total amount of time spent on task ($t(68) = 0.93$, $P = 0.36$, $b = 0.12$, $SE_b = 0.14$, $\beta = 0.11$, 95% CI = [−0.15, 0.40]) changed as a function of accuracy (Fig. 4b). Instead, our results indicate a critical role for the distribution of actions over time. Indeed, whereas poor versus proficient participants could be differentiated on the basis of the temporal distribution of their actions (their shape and scale parameters; Fig. 4c), they could not be differentiated on the basis of the time they spent and the number of actions they performed (Fig. 4d; see Supplementary Information for additional analyses).

Given that in our replication study (Experiment 1b) the P value of the difference between the Fisher r-to-z transformed correlations failed to reach conventional levels of significance (specifically, the observed P value, $P = 0.054$, was larger than our chosen alpha-level, $\alpha = 0.05$), a further—and more highly powered—replication study will be required to substantiate the superior predictive power of the click-and-drag Raven. Interestingly, our findings indicate that an IQ test that allows participants to externalize cognitive operations may be a better predictor of academic achievement than the conventional static IQ test. Why would this be the case? We would suggest that the click-and-drag Raven task provides a better test of a problem-solver's capacities to perform what Kirsh and Maglio dubbed 'epistemic actions'[32]. Whereas pragmatic action is performed with the aim to bring one physically closer to a goal, epistemic action is performed to extract or uncover useful information that is hidden or difficult to compute mentally[20,26,33]. For example, the purposeful reconfiguration of external figures in the click-and-drag Raven task can enable a problem-solver's attentional system to lock onto configural patterns that were previously obscured. By reordering the figures, a featural dimension can become easier to parse, leaving more resources available to discover patterns in the remaining featural dimensions.

In daily life, we perform epistemic actions naturally, for example when we shuffle scrabble tiles in ways that respond to emerging fragmentary guesses while simultaneously cueing better ideas, leading to new shufflings and so on. From this perspective, epistemic actions may be considered part and parcel of the reasoning process[17,20] and are likely to be important in academic contexts. Given that students routinely have to solve complex problems within information-rich, reconfigurable (digital) environments, it seems reasonable to assume that skills at epistemic action may be especially beneficial. The click-and-drag Raven task, we suggest, may a better detector of this kind of crucial cognitive ability than the conventional static Raven task.

Consistent with this interpretation, it has been observed that tasks that allow room for people's natural propensity to perform epistemic actions often have real-world predictive power in various cognitive domains[26]. For instance, Gilbert has shown that an

intention offloading task that allowed the externalization of cognitive operations was a better predictor of real-world intention fulfilment than a task that did not[28]. Also, participants tend to persevere less with suboptimal, idiosyncratic, task-specific strategies in paradigms that allows cognitive operations to be externalized[29–31], which may increase the generalizability of task outcomes.

In a recent paper, Duncan et al. proposed that a critical aspect of fluid intelligence is the function of cognitive segmentation, which is the process of subdividing a complex task into separate, simpler parts[34]. To investigate this, Duncan et al. presented participants with Raven-style matrix problems and asked them to work out the missing figure by drawing figure elements in a blank answer box. This allowed participants to externalize partial solutions to the problem and encouraged them to cognitively segment the problem into its constituent subcomponents. Consistent with the present study, they found that their modified matrix problems showed a slightly higher correlation with a criterion IQ test (0.53) than conventional matrix problems (0.41). These findings raise the following interesting question: was the click-and-drag Raven task better at predicting academic achievement because it helped participants to split the overall problem into simpler subcomponents?

We agree with the claim that cognitive segmentation is a critical function of fluid intelligence. Indeed, we would argue that both in our click-and-drag Raven task and the modified matrix task of Duncan et al., external operations were the means through which participants were able to cognitively segment the problems that were presented to them. However, we would also argue that, in addition to segmentation, external operations enable a problem-solver to recombine task subcomponents in novel ways and perceptually re-encounter them, which, when followed up with critical reflection, allow participants to gain novel insights into the structure of the problem. In other words, external operations not only facilitate the cognitive segmentation of a task but they also produce changes (intended or serendipitous) in the external input which enable an agent to reconceptualize the problem. In this respect, it would be interesting for future research to investigate whether the act of cognitive segmentation is necessarily implemented through external operations (either in the form of active task manipulations or more passive attentional task restructuring[34]).

Given that the click-and-drag Raven task displayed a higher correlation with academic achievement, it would also be interesting to investigate how the temporal profile of problem-solving relates to academic outcomes. To investigate this, one could measure the temporal profiles of task actions and task performance both during the Raven task as well as during a criterion task (for example relating to achievement). Then, one could test whether the type of temporal profiles exhibited during the Raven and criterion task are associated and to what extent this generalization of task strategy can account for the association between Raven and criterion task performance. In other words: to what extent can the association in task outcomes be explained by epistemic strategies that generalize over tasks?

It is important to note two methodological limitations of the present study. Given that we only tested undergraduate students, further research is needed to assess whether our findings are also applicable to the general population. Also, further research is needed to generalize our findings to Raven items other than the particular items we selected for our experiments.

In summary, our work offers a widely applicable approach for investigating how people use their task environment during problem-solving. Our results suggest that an IQ test that allows information processing to be offloaded onto the environment may be better than a more conventional static IQ test at predicting academic achievement. Furthermore, we provide a quantitative demonstration of the degree to which intelligent problem-solvers may benefit from external cognitive operations. The ability to use external objects, props and aids to solve complex problems is considered by many to be a unique feature of human intelligence[16–25,37], which may have provided the core impetus to the advancement of civilization[22–25,37]. Our study supports the emerging view that much of what matters about human intelligence is hidden not in the brain, nor in external technology, but lies in the delicate and iterated coupling between the two[17–25,37,38].

Methods

No statistical methods were used to determine sample size but our sample sizes are similar to those reported in previous publications[4–6,15,27,29–32]. The assignment of participants to between-subjects conditions (click-and-drag versus static Raven task) was randomized and was not blinded to investigators. Both in the click-and-drag and static Raven tasks, items were presented in a fixed order of increasing difficulty for each participant (that is, SPM-D5, SPM-D9, APM-1, APM-8, APM-13, APM-14, APM-17, APM-21, APM-27, APM-28, APM-34). Data collection and analysis were not performed blind to the conditions of the experiments. No participants or data points were excluded from the analyses.

Informed consent. All experiments reported were conducted in accordance with relevant regulations and institutional guidelines and approved by the local ethics committees of the Faculty of Social and Behavioural Sciences, Leiden University and the Erasmus School of Social and Behavioural Sciences, Erasmus University, Rotterdam. All participants signed a consent form before participating in the experiment and received written debriefing after participating in the experiment.

Experimental studies. In Experiment 1a, 211 Leiden University students (156 women, 55 men, M_{age} = 21.4 yr, s.d.$_{age}$ = 3.2 yr) and in Experiment 1b, 284 Erasmus University students (236 women, 48 men, M_{age} = 20.4 yr, s.d.$_{age}$ = 3.1 yr), with normal or corrected-to-normal vision were randomly assigned to either a conventional static Raven IQ test or a click-and-drag Raven IQ test. Academic achievement was assessed using average exam grades on a ten-point scale for a selection of Bachelor of Psychology courses. To validate the Raven Advanced Progressive Matrices tests for fluid intelligence, we selected first-year courses in the Bachelor curricula that were general in their content and that required abstract and logical reasoning. For Leiden University students, we selected the courses Introduction to Psychology, Introduction to Research Methods and Inferential Statistics. For Erasmus University students, we selected the courses Introduction to Research Methods and Practical Statistics. In Experiment 2, we recorded the time course of mouse actions for a new sample of 70 Leiden University students (53 women, 17 men, M_{age} = 20.8 yr, s.d.$_{age}$ = 3.4 yr) performing the click-and-drag Raven IQ test. All participants were undergraduate students participating for course credit or a small monetary reward (€4.00).

Both the static and click-and-drag IQ tests consisted of 11 items taken from the Raven Standard and Advanced Progressive Matrices. In the static test, participants were instructed to inspect the array of figures and decide which figure was missing, whereas in the click-and-drag test participants were instructed to sort these figures into the grid using the mouse, leaving one of the bottom three positions empty. Next, they selected the missing figure from the eight alternatives presented below the array. There was a time limit of 4 min to complete each item and the time remaining to complete the item was displayed at the top of the screen.

In Experiments 1a and 1b, we performed analyses of variance (ANOVAs), linear regression analyses and planned contrasts on the proportion of accurate responses (Raven performance level), the exam grades for the various Bachelor courses and the average exam grade for the courses (academic achievement score). In Experiment 2, we performed linear regressions on log-transformed shape parameters, scale parameters, average partial autocorrelations, variances of intermovement intervals, total number of movements, total time spent (s) and Raven response accuracies. Data distributions of Raven accuracy scores, academic achievement scores and the log-transformed parameters were assumed to be normal but this was not formally tested. All statistical tests conducted in the reported experiments were two-tailed. For further analyses and details of the experimental methods, see Supplementary Information.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

The routines/code that were used to perform the statistical analyses in this study are available from the corresponding author upon request. For the routine/code that was used for simulating the dual-mode and single-mode problem-solvers see Supplementary Code.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

Received: 28 November 2017; Accepted: 5 December 2018;
Published online: 4 February 2019

References

1. Jensen A. R. *The G Factor: The Science of Mental Ability.* (Praeger, Westport, CT, USA, 1998).
2. Deary, I. J., Strand, S., Smith, P. & Fernandes, C. Intelligence and educational achievement. *Intelligence* **35**, 13–21 (2007).
3. Kyllonen, P. C. & Christal, R. E. Reasoning ability is (little more than) working-memory capacity?! *Intelligence* **14**, 389–433 (1990).
4. Engle, R. W., Tuholski, S. W., Laughlin, J. E. & Conway, A. R. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *J. Exp. Psychol. Gen.* **128**, 309–331 (1999).
5. Duncan, J. et al. A neural basis for general intelligence. *Science* **289**, 457–460 (2000).
6. Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J. & Minkoff, S. R. A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence* **30**, 163–183 (2002).
7. Engle, R. W. Working memory as executive attention. *Curr. Dir. Psychol. Sci.* **11**, 19–23 (2002).
8. Kyllonen, P. C. In *The General Factor of Intelligence: How General Is It?* (eds. Sternberg, R. J. & Gigorenko, E. L.) 415–445 (Erlbaum, Mahwah, NJ, USA, 2002).
9. Baddeley, A. Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* **4**, 829–839 (2003).
10. Colom, R., Flores-Mendoza, C. & Rebollo, I. Working memory and intelligence. *Pers. Indiv. Differ.* **34**, 33–39 (2003).
11. Conway, A. R., Kane, M. J. & Engle, R. W. Working memory capacity and its relation to general intelligence. *Trends Cogn. Sci.* **7**, 547–552 (2003).
12. Gray, J. R., Chabris, C. F. & Braver, T. S. Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* **6**, 316–322 (2003).
13. Olesen, P. J., Westerberg, H. & Klingberg, T. Increased prefrontal and parietal activity after training of working memory. *Nat. Neurosci.* **7**, 75–79 (2004).
14. Kane, M. J., Hambrick, D. Z. & Conway, A. R. A. Working memory capacity and fluid intelligence are strongly related constructs. *Psychol. Bull.* **131**, 66–71 (2005).
15. Jaeggi, S. M., Buschkuehl, M., Jonides, J. & Perrig, W. J. Improving fluid intelligence with training on working memory. *Proc. Natl Acad. Sci. USA* **105**, 6829–6833 (2008).
16. Hutchins, E. *Cognition in the Wild.* (MIT Press, Cambridge, MA, USA, 1995).
17. Clark, A. & Chalmers, D. The extended mind. *Analysis* **58**, 7–19 (1998).
18. Clark, A. An embodied cognitive science? *Trends Cogn. Sci.* **3**, 345–351 (1999).
19. Giere, R. in The Cognitive Bases of Science (ed.s Carruthers, P., Stitch, S. & Siegal, M.) 285–299 (Cambridge Univ. Press, Cambridge, UK, 2002).
20. Clark, A. *Supersizing the Mind: Action, Embodiment, and Cognitive Extension*. (Oxford Univ. Press, Oxford, 2008).
21. Rowlands, M. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology.* (MIT Press, Cambridge, MA, USA, 2010).
22. Bocanegra, B. R. Troubling anomalies and exciting conjectures: a bipolar model of scientific discovery. *Emot. Rev.* **9**, 155–162 (2017).
23. Lee, K., & Karmiloff-Smith, A. In *Perceptual and Cognitive Development* (eds. Gelman R. et al.) 185–211 (Academic Press, New York, 1996).
24. Mithen, S. In *Evolution and the Human Mind* (eds. Carruthers, P. & Chamberlain, A.) 207–217 (Cambridge Univ. Press, Cambridge, UK, 2002).
25. Clark, A. *Natural-born Cyborgs: Minds, Technologies and the Future of Human Intelligence*. (Oxford Univ. Press, Oxford, UK, 2003).
26. Risko, E. F. & Gilbert, S. J. Cognitive offloading. *Trends Cogn. Sci.* **20**, 676–688 (2016).
27. Risko, E. F. & Dunn, T. L. Storing information in-the-world: metacognition and cognitive offloading in a short-term memory task. *Conscious. Cogn.* **36**, 61–74 (2015).
28. Gilbert, S. J. Strategic offloading of delayed intentions into the external environment. *Q. J. Exp. Psychol.* **68**, 971–992 (2015).
29. Vallée-Tourangeau, F., Euden, G. & Hearn, V. Einstellung defused: interactivity and mental set. *Q. J. Exp. Psychol.* **64**, 1889–1895 (2011).
30. Vallée-Tourangeau, F., Steffensen, S. V., Vallée-Tourangeau, G. & Sirota, M. Insight with hands and things. *Acta Psychol.* **170**, 195–205 (2016).
31. Weller, A., Villejoubert, G. & Vallée-Tourangeau, F. Interactive insight problem solving. *Think. Reasoning* **17**, 424–439 (2011).
32. Kirsh, D. & Maglio, P. On distinguishing epistemic from pragmatic action. *Cognitive Sci.* **18**, 513–549 (1994).
33. Kirsh, D. Thinking with external representations. *AI Soc.* **25**, 441–454 (2010).
34. Duncan, J., Chylinski, D., Mitchell, D. J. & Bhandari, A. Complexity and compositionality in fluid intelligence. *Proc. Natl Acad. Sci. USA* **114**, 5295–5299 (2017).
35. Kaplan, R. & Saccuzzo, D. *Psychological Testing: Principles, Applications, and Issues.* 8th edn. (Cengage, Boston, 2012).
36. Barabasi, A. L. The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211 (2005).
37. Tomasello, M. *The Cultural Origins of Human Cognition* (Harvard Univ. Press, Cambridge, MA, USA, 2009).
38. Goodale, M. Thinking outside the box. *Nature* **457**, 539–539 (2009).

Acknowledgements

The authors received no specific funding for this work. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

B.R.B., F.H.P. and B.F. designed the experiments. B.R.B. carried out the experiments, simulations and statistical analyses. B.R.B., F.H.P., B.F. and A.C. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0509-y.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to B.R.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): B. R. Bocanegra

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	All experimental tasks were implemented using a locally run version of QRTengine (see Barnhoorn, J., Haasnoot, E., Bocanegra, B.R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. Behavior Research Methods, 47, 918-929.), which made use of the Qualtrics survey platform (www.qualtrics.com).
Data analysis	Data fitting, simulation and visualization was performed using Easyfit v1.0 (Mathwave Technologies) and MS Excel 2013. Statistical analyses were performed using IBM SPSS Statistics 24.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available from the corresponding author upon request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study type is a quantitative experimental study.
Research sample	All participants were bachelor students of the Faculty of Social and Behavioural Sciences, Leiden University and the Erasmus School of Social and Behavioral Sciences, Erasmus University Rotterdam. In Experiment 1a, we tested two-hundred and eleven Leiden University students (156 women, 55 men, Mage = 21.4 years, SDage = 3.2 years). In Experiment 1b, we tested two-hundred and eighty-four Erasmus University students (236 women, 48 men, Mage = 20.4 years, SDage = 3.1 years). In Experiment 2, we tested seventy Leiden University students (53 women, 17 men, Mage = 20.8 years, SDage = 3.4 years). A sample of university students was specifically chosen in order to be able to test adaptive fluid intelligence and complex problem-solving in the context of academic achievement.
Sampling strategy	For each experiment we collected a convenience sample of students through university course-credit administration systems. No statistical methods were used to determine sample size but our sample sizes are similar to those reported in previous publications (see manuscript references 4-6,15,27,29-32).
Data collection	Tests were administered individually on computers in 1-person sound-proof cubicles. Given the fact that the tests had to be administered individually and informal post-experimental questions were asked by experimenters, the assignment of participants to the two randomized experimental conditions was not blinded to experimenters.
Timing	For Experiment 1a, data collection occurred between January 4th-April 22nd, 2016. For Experiment 1b, data collection occurred between January 15th-May 28th, 2018. For Experiment 2, data collection occurred between March 3rd-March 31st, 2017.
Data exclusions	No data was excluded from the analyses.
Non-participation	None of the participants dropped out or declined to participate during the course of the experimental session. All participants signed a consent form prior to participating in the experiment, and received written debriefing after participating in the experiment.
Randomization	The assignment of participants to between-subjects conditions (click-and-drag vs. static Raven task) was randomized.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	The participants enrolled voluntarily to participate in the study through an online university course-credit system. The nature of the study was not disclosed to participants. The study was referred to as simply a computer task to be performed in the behavioral labs of the faculty. Given the fact that students need to acquire a considerable amount of course-credit during their bachelor program (obfuscating their capacity to be selective in their choices), we expect that selection biases will have minor effects in influencing the outcomes of our study.