Journal of Business and Psychology
https://doi.org/10.1007/s10869-026-10135-x

ORIGINAL PAPER

Faking Ipsative Questionnaires: The Role of General Mental Ability and Ability to Identify Criteria

Miriam Fuechtenhans[1] · **Anna Brown**[1]

Received: 14 July 2025 / Revised: 15 June 2026 / Accepted: 17 June 2026

Abstract

Questionnaires in ipsative (relative-to-self) formats ask respondents to compare several stimuli, for example behaviors or personal characteristics, at a time. Since it is impossible to endorse all desirable stimuli, the ipsative formats are often used in high-stakes personality assessments to reduce impression management (aka 'faking good'). However, their high cognitive complexity may contaminate the intended measurement of personality with cognitive abilities. This study investigated the role of general mental ability (GMA) and the Ability to Identify Criteria (ATIC) in faking personality assessments utilizing a 'graded preference' format, scored normatively using Thurstonian modeling. We recruited $N = 200$ participants who completed the Leadership Style Questionnaire (de Klerk-van Someren et al., 2023) under two conditions: imagining they are applying for a specified job (*high-stakes* condition) and responding normally (*low-stakes* condition). In both completions, participants' response latencies and click counts were recorded. The results indicate that participants can improve their scores on job-relevant traits, but effect sizes are small to medium (0.26 to 0.47). Participants with higher GMA tend to inflate their scores more, and participants with higher ATIC produce personality profiles that closer match the target job when faking. Through cross-classified path analysis, we unpick effects at block and person level and show that criteria recognition is the main driver for targeted increases in personality scores, and strategic editing of responses is the most plausible explanation for increased response latencies and click counts when faking. We also suggest that in repeated measures faking studies, high-stakes condition should precede low stakes.

Keywords Faking · Ipsative questionnaires · Graded preferences · Ability to Identify Criteria (ATIC) · General mental ability (GMA) · Personality assessments · Response latencies

Research has found that 30–50% of job candidates modify their answers on personality assessments to enhance their chances of being hired (e.g., Arthur et al., 2010; Donovan et al., 2013; Griffith & Converse, 2011; Peterson et al., 2011). In the personnel selection context, candidate *faking* can be defined as a motivated response to a self-reported personality measure with the goal of creating a favorable impression (Kiefer & Benit, 2016). Since such motivated responses do not necessarily correspond to the person's true self-image, faking can lead to significant distortions in score distributions, compromise the construct validity, and impact the validity of selection decisions (Pavlov et al., 2019). Failure to either prevent or control these distortions through statistical means poses a considerable threat to the validity and equity of psychological assessments (Kuncel et al., 2011).

Prevention of Faking with Ipsative Response Formats

To prevent candidate faking, the relative-to-self or *ipsative* (from Latin '*ipse*' meaning *self*) response formats have been introduced, where candidates express preferences among two or more stimuli presented simultaneously (e.g., Dilchert & Ones, 2011). In contrast to the traditional *normative* response formats, where one stimulus is rated at a time, with the ipsative format it is no longer possible to endorse desirable options for all items; thus, it has the potential to prevent faking.

Additional supplementary materials may be found here by searching on article title https://osf.io/collections/jbp/discover

✉ Anna Brown
a.a.brown@kent.ac.uk

[1] School of Psychology, University of Kent, Canterbury CT2 7NP, UK

Published online: 02 July 2026

The most popular and well-known type of ipsative format is *forced choice* (FC). In FC pairs, respondents must choose one of two alternatives that describe them best. When three, four or more alternatives are presented in a block (making triplets, quads, etc.), respondents are either asked to rank all items according to how well they describe them or to indicate which item describes them most and which describes them least (Brown, 2015). In either format, the obtained rankings are reduced to simple *binary preferences* in each pair of items involved. For example, the ranking $A > B > C$ reduces to three pairwise binary preference decisions: $A > B$, $A > C$ and $B > C$. Ranking blocks can be constructed to include items measuring the same trait (unidimensional FC), or different traits (multidimensional FC or MFC). Blocks of three or more items are most often multidimensional.

The rationale for using ipsative formats was the hope that when faced with equally desirable alternatives, the candidate will report true preferences (Gordon, 1951). Others have argued that the format facilitates the direct comparison of stimuli desirability, which may even promote faking (Feldman & Corah, 1960). Since then, research has consistently provided evidence that the MFC format can reduce faking substantially (for recent meta-analyses, see Cao & Drasgow, 2019; Martínez & Salgado, 2021; Speer et al., 2023). However, the extent of faking prevention strongly depends on how well items within FC blocks are matched for desirability (Cao & Drasgow, 2019). Earlier studies that failed to control for this important factor yielded inconsistent results (e.g.Christiansen et al., 2005; Heggestad et al., 2006). Later studies have consistently shown that FC measures where items within blocks are matched with respect to their desirability for a given selection situation are much more resistant to faking than measures with poorly matched items (Cao & Drasgow, 2019; Speer et al., 2023; Wetzel et al., 2021).

An important issue of the ipsative formats is that they inherently produce relative-to-self or personally incomparable item scores (Brown & Maydeu-Olivares, 2011). When relative responses within each block are summated to yield personality trait scores, such trait scores are relative to other traits within the person and cannot be used to compare people (Hicks, 1970). However, advancements in Item Response Theory (IRT) models such as Multi-Unidimensional Pairwise Preference Model (MUPP; Stark et al., 2005) or Thurstonian IRT model (TIRT; Brown & Maydeu-Olivares, 2011) have enabled normative trait scores to emerge from the ipsative formats. These developments have enabled the use of ipsative questionnaires in personnel selection, introducing a unique approach to assess personality and other non-cognitive domains.

However, recent research suggests that FC assessments may result in test taker frustration or dissatisfaction (Dalal et al., 2019). It is often found that test takers perceive the FC format more challenging than its single-stimulus (SS) counterparts (Zhang et al., 2019) and react to it less favorably (Converse et al., 2008), particularly when completed for a hypothetical job application. Subsequently, Borman et al. (2023) raised the concern that employers may face a challenge in balancing the reduction of faking and fostering a positive candidate experience.

Further candidate frustration with the FC assessments is caused by the dichotomous nature of choice decisions required of them, while many candidates may want to express the finer nuances of their preferences (Borman et al., 2023). Indeed, whether the candidate prefers A to B decisively or just marginally is not captured in the FC format. To address these limitations, the 'graded preference' and 'proportional preference' ipsative formats have emerged (Brown, 2015). In *graded preference* formats, respondents indicate the extent to which one stimulus is more (or less) true of them than the other using ordered categories, for example, "much more like me" or "a little more like me" or even "equally like me". Ordinal data resulting from this format can be analyzed using, for example, the graded version of TIRT (Brown & Maydeu-Olivares, 2018), deriving normative trait scores. In *proportional preference* (aka *fixed-point-total* or *compositional*) formats, respondents quantify their preferences among several stimuli by distributing a fixed number of points (say, 15) between them. When all stimuli have similar appeals to the respondent, points are distributed more evenly. When one stimulus is decisively more appealing to the respondent, more points are allocated to it, and the proportion of points given to this stimulus relative to others increases, hence the name 'proportional'. Normative traits scores can also be derived from these ratio preferences using the log-linear factor analysis version of Thurstonian model (Brown, 2016).

There has been less research regarding faking in these quantitative preference formats, and the available research is inconclusive about their robustness to faking. While Zhang et al. (2023) found that test takers perceive graded preferences less fakeable than normative graded response items (aka Likert-type items), and Schulte et al. (2024) confirmed less inflation in a graded preference questionnaire than in a Likert-scales questionnaire under faking conditions, it has been argued that respondents have more opportunities to fake the graded preference questionnaire than a FC questionnaire by expressing more extreme preference for more desirable items (Schulte et al., 2024). Given increasing popularity of quantitative preference formats in personality assessments (de Klerk-van Someren et al., 2023; MyPrint Questionnaire, 2018) and their alleged capacity to increase the reliability and improve applicant reactions (Dalal et al., 2019), these ipsative formats should be further researched with regard to their ability to prevent faking.

The Influence of Cognitive Factors on Faking Non-cognitive Assessments

Faking behavior in personality assessments may be moderated by various factors, including individual differences in motivation, attitudes, ability and opportunity as well as characteristics of the situation (Ziegler, 2011). Of most relevance to the MFC measures may be cognitive factors, because of the increased cognitive complexity of preferential judgments compared to absolute judgments. In his book 'Thinking, fast and slow' (2011), Kahneman dedicated a whole chapter (Chapter 33, 'Reversals') to comparative judgments and argued that they elicit effortful, 'slow', thinking. Specific cognitive factors that have been identified in literature as highly relevant to FC assessments are *general mental ability* and *knowledge about the specific attributes being assessed*.

Regarding the general mental ability (GMA), a study by Vasilopoulos et al. (2006) found that GMA was related to the personality scores obtained from the FC measure, but only in the faking condition. Specifically, the high-GMA participants were able to inflate their FC scores by more than three-quarters of a standard deviation compared to the honest condition, while some low-GMA participants failed to increase their scores. Interestingly, no such effects of GMA were observed for the counterpart normative measure. Subsequently, Vasilopoulos and colleagues suggested that FC personality tests represent "measures of personality and cognitive ability". Christiansen et al. (2005) demonstrated very similar effects, and by manipulating job descriptions, they were able to show the mediating role of implicit job theories (IJT, which are beliefs about job demands and relevant traits) in the relationship between GMA and score increases from honest to faking condition for both normative and ipsative questionnaires. Consistent with the idea of increased cognitive load associated with the forced-choice formats, the role of IJT in inflating the test score was greater for ipsative than for normative responses. The significant positive effect of GMA on implicit job theories (the standardized path coefficient was 0.2 in the Christiansen et al. study) suggests that the context-specific insight into job demands to a significant extent depends on the person's general mental ability.

This 'ability-like' nature of a candidate's understanding of constructs being assessed is emphasized in the 'Ability to Identify Criteria' (ATIC) construct, defined as the candidate's ability to recognize the key personal attributes that an employer is seeking to evaluate in an assessment (Kleinmann et al., 2011). Like IJT, ATIC is positively associated with inflation of assessment scores. For instance, a study conducted by König et al. (2007) found that applicants with higher ATIC scores showed greater performance in both assessment centers and structured interviews. Moreover, after controlling for general mental ability, the ATIC scores obtained in one assessment method were predictive of participants' performance on the other assessment method (see also Oostrom et al., 2016). In (normative) personality assessments, ATIC has also been found partly responsible for test takers' tendency to inflate all desirable personality items (referred to as the 'ideal employee factor'). What is more, Klehe et al. (2012) showed that the relationship between the ideal employee factor in a (normative) personality assessment and performance in work simulations can be partly explained by ATIC, even after controlling for GMA. Such findings confirm the role of ATIC as an important factor in faking, related to but distinct from GMA. Examining the role that GMA and ATIC and their interactions play in cognitively complex ipsative assessments seems particularly useful.

The Effect of Faking on Response Latencies and Other Process Data

Beyond actual item responses that candidates provide in normative or ipsative assessments, process data routinely captured in computer-based assessments such as numbers of mouse or keyboard clicks, response latencies, etc. can be used to further understand the faking behavior (Kovačić, 2021). If faking acts as an editing process of one's retrieved responses as many researchers suggest (Brown & Böckenholt, 2022; Fuechtenhans & Brown, 2023; Holtgraves, 2004), then extra mouse clicks may be seen as direct indicators of changing one's mind. Regarding response latencies, the dominant view is that faking should yield longer response latencies in general (e.g.Holtgraves, 2004; Walczyk et al., 2009) since it is a deliberate behavior that requires additional cognitive effort and processing time. Within this view, important person-level factors moderating the speed of response are acknowledged, including individual differences in the tendency to self-deception that makes socially desirable responding faster (Holtgraves, 2004), and individual differences in cognitive ability that makes faking more successful under time pressure (Komar et al., 2010). These findings highlight that the increase in cognitive effort involved in faking does not uniformly affect all individuals. In fact, individuals with high GMA may successfully manipulate a test even when the cognitive effort is heightened.

However, there exist alternative models for response latencies arguing against the expected general increase in response latencies under faking conditions. For instance, Holden et al. (1992) proposed a 'congruence model', which predicts fastest responses to occur when stimulus material is congruent with a respondent's self-schema. This mechanism, however, applies only in 'honest' (low stakes) contexts,

whereas in 'faking' (high stakes) contexts, fastest responses occur when stimulus material is congruent with an adopted schema (respondent's schema regarding favorable impression). Thus, the speed of responding should not only depend on the assessment stakes, but also on the stimulus material, eliciting faster responses to desirable items but slower responses to undesirable items because they are incongruent with the adapted schema of an 'ideal candidate' (Holden & Lambert, 2015). Interestingly, indirect support for the congruence model was found in a recent qualitative study of faking MFC questionnaires (Fuechtenhans & Brown, 2023), where participants reported increased hesitation and difficulty when making forced choices among undesirable alternatives (items that are incongruent with the adopted schema of an 'ideal candidate').

Taken together, these findings underscore the important role that process data and specifically response latencies can play in studying faking behavior, by acting as an additional indicator of cognitive load and thus inherently interlinked with all cognitive factors such as GMA, IJT and ATIC. Measuring response latencies to understand further the cognitively complex faking behavior in ipsative questionnaires may be particularly useful.

Gaps in Research of Faking Behavior in Ipsative Questionnaires

We have identified the following gaps in existing research that the present study will aim to address.

1) The lack of conclusive results about fakeability of graded preferences (and other quantitative ipsative formats) due to their relative novelty on the personnel assessment market.
2) The lack of understanding of how responding to cognitively complex graded preferences under honest and faking conditions impacts response latencies.
3) The lack of understanding of how faking behavior might manifest itself through actual editing of responses captured as number of clicks made in graded preference blocks.
4) The lack of evidence for the (potentially mediating) role of ATIC in the relationships between GMA and score inflation in graded preferences.

We aim to address these gaps through investigating a graded preference questionnaire for direct and indirect outcomes of faking behavior—actual responses, response latencies and numbers of clicks made. We consider GMA and ATIC as potential causes, mediators and moderators of these outcome variables, to provide a more nuanced view of the faking process in ipsative questionnaires, particularly those using quantitative comparative judgments. Given that quantitative ipsative formats capture more nuanced information than forced choice, the role of cognitive factors may be enhanced.

The Present Study

The present study was designed to address the above research gaps and investigate the role of GMA and ATIC in faking ipsative items that capture the extent of preferences for presented alternatives. Within Griffith and Robie's (2013) 'seven questions about faking' framework, the study primarily addressed two questions: "Can people fake?" and "Do people differ in the ability and motivation to fake?", specifically in relation to a unique '7-point-total' format, which, as we explain further in Materials, combines features of both graded preferences and proportional preferences (compositions) and thus provides a unique opportunity to investigate the effects of faking behavior on quantitative ipsative comparisons that may generalize to many specific formats.

We investigated the first question by manipulating assessment stakes and measuring the extent of people's personality score inflation. We investigated the second question by examining effects of GMA and ATIC on people's ability to inflate traits important for the target job. To this end, we utilized a repeated measures design, where all participants completed the "Leadership Style Questionnaire" (LSQ; De Klerk-van Someren et al., 2023) under two conditions: *high stakes* / faking condition (where participants are instructed to complete a personality questionnaire as if applying for a target job) and *low stakes* / honest condition (where participants are instructed to respond honestly). For both personality questionnaire completions, we collected response latencies and click counts. Additionally, we measured the Ability to Identify Criteria and general mental ability using objective tests and included these test scores as covariates. With this collected information, we investigated faking behavior through two different lenses: by looking at actual test responses (and resulting scores) as direct outcomes of the faking process; and by looking at auxiliary information complementing the test responses, namely response latencies and numbers of clicks.

Personality Test Scores

The Leadership Style Questionnaire utilized in the present study measures 24 traits, all of which are formulated as generally desirable qualities for a leader. However, some of these traits will be more important than others for the target job that is presented to the participants in the faking condition. Therefore, we expect that when presenting self

in a positive light, respondents will aim to inflate their scores on these 'job-essential' qualities. Given that cognitive factors play a major role in responding to ipsative personality questionnaires, both GMA and ATIC should influence the assessment scores. In line with the existing research, primarily Christiansen et al. (2005) and Vasilopoulos et al. (2006), we formulated the following hypotheses about the observed personality scores:

H1a: Personality scores on traits essential for the target job will be higher in the faking than in the honest condition.
H1b: Higher GMA will be associated with the larger increase on job-essential personality scores between honest and faking conditions.

Lastly, given the mediating nature of IJT in the relationship between GMA and personality score inflation (Christiansen et al., 2005), we expect that ATIC too will mediate the relationship between GMA and score inflation in our ipsative questionnaire.

H1c: ATIC will mediate the relationship between GMA and the increase in job-essential personality scores between honest and faking conditions.

Response Latencies

Based on existing research (e.g., Holtgraves, 2004), we expected participants to be slower in answering the personality questionnaire under faking conditions, as they engage in a deliberate manipulation of their responses potentially involving complex judgments of what is desirable and undesirable for a given job. However, given the evidence that speeding personality tests reduces faking only for participants low in mental ability (Komar et al., 2010), we expect that participants with higher GMA scores will respond faster than those with low-GMA in the faking condition. However, they may also be faster responding in low-stakes condition given the known associations between GMA and processing speed. Given the mediating role that ATIC plays in the relationship between GMA and ipsative score inflation when people fake (Christiansen et al., 2005), we may reasonably expect that ATIC will also mediate the relationship between GMA and response latency, but the extent and direction of this mediation is yet to be discovered. Thus, we formulated the following hypotheses about response latencies:

H2a: Response latencies will be greater when participants complete the personality questionnaire under the faking condition.
H2b: Higher GMA will be associated with shorter response latencies in both honest and faking conditions.
H2c: ATIC will mediate the relationship between GMA and response latencies in the faking condition. The direction and size of this mediation is yet unknown, and this will be an open exploration.

Click Counts

In addition to investigating response latencies, we wanted to explore whether the number of clicks a participant makes to come to their final response is indicative of their faking behavior. Arguably, if faking acts as an editing process as many researchers suggest (Brown & Böckenholt, 2022; Fuechtenhans & Brown, 2023; Holtgraves, 2004), a larger number of clicks on a given graded response block might be a direct manifestation of participants changing their minds. Furthermore, high-GMA participants might have greater ability to work out the 'right' response in high stakes, thus reducing the number of clicks they need to make, compared to low-GMA individuals. Furthermore, ATIC may mediate the relationship between GMA and the number of clicks, but the direction and size of this mediation is yet unknown. Thus, we formulated the following hypotheses about click counts:

H3a: Click counts will be higher when participants complete the personality questionnaire under the faking condition.
H3b: Higher GMA will be associated with lower click counts in both honest and faking conditions.
H3c: ATIC will mediate the relationship between GMA and click counts in the faking condition. The direction and size of this mediation is yet unknown, and this will be an open exploration.

Method

Design and Procedure

This study was a two-part online study utilizing the survey software Qualtrics. Participants were recruited via Prolific online panel and were granted access to study Part 1 immediately. All participants were first presented with the information sheet followed by the informed consent form. Everyone was asked to provide some basic demographic questions, such as age, nationality, and employment status. After completing Part 1, participants had to wait 72 h before they were given access to Part 2. Importantly, the order of study parts was counterbalanced with 100 participants completing the faking condition first, followed by the honest condition, and 100 participants completed the conditions in reverse order. After completing both parts, all participants received a written debrief. They were compensated for their time with a total of £15 after completing both parts of the study.

In both conditions, participants were first asked to undertake a general mental ability test "designed to measure their ability to work with numbers, identify logical patterns, and deduce accurate information from a passage of text". The GMA measure was administered first to ensure participants were most motivated and least tired when completing this maximum performance test, and no experimental manipulation affected its validity. We emphasized the timed nature of the test and the importance of trying to achieve optimal performance. To help participants become familiar with the task, we provided two examples of each of the three question types and gave feedback on the correct answers. After completing the ability test, which was limited to 25 min, participants were presented with instructions according to the stakes to which they were routed first.

Honest Condition

In this condition, participants were presented with normal instructions for the LSQ: "This questionnaire will explore your personal preferences in a work setting. You will be presented with three statements at a time. Please read them carefully and then compare how each of the three statements describes you as a person at work. When comparing the three statements, we would like you to tell us, out of the three, which is more descriptive of who you are… We encourage you to complete the questionnaire honestly and sincerely." Completing the honest condition took around 60 min.

Faking Condition

In this condition, participants were first presented with the job description of Director of Market Research and Analytics and asked to carefully read it. Then, the participants reviewed 24 personal qualities and rated the importance of each quality for good performance in the target job. After completing this task, participants were presented with personality questionnaire instructions specific to high stakes, which were inspired by work of Mueller-Hanson et al. (2003): "*You have just reviewed a job description for the position of Director of Market Research and Analytics and reviewed several qualities whilst indicating how important they are in relation to good performance on the job. Now, please imagine that you have applied for this position, and you really want this job. The only obstacle that stands in your way is a personality assessment that you need to complete as the last step of the selection process. Based on your personality profile, the hiring manager will decide whether you get the job or not. After study completion, we will evaluate your personality profile and decide whether the profile you presented would render you successful in getting hired. If you are successful, we will enter you into our prize draw with a chance of winning a £50 gift card…*" As argued by Mueller-Hanson et al. (2003), such an instruction mirrors several aspects of a real selection situation, such as a general understanding of the job criteria and characteristics an organization is looking for based on the job description, and an awareness that the assessment results will be instrumental in the selection decision. Additionally, the monetary incentive for a successful application somewhat resembles that of a genuine job application.

After reading these instructions, we asked participants some screening questions, intended to verify that they had read and understood the instructions. The screening questions comprised six statements, with three containing accurate information and three containing false information (see Appendix B); failure to pass these questions prevented participants from proceeding with the next task. All participants passing the screening questions were then presented with the LSQ.

Participants

A total of 200 individuals were recruited via Prolific. Of the participants, 97 (48.5%) were women, and 101 (50.5%) were men. One participant (0.5%) reported non-binary/third gender, and one did not report their gender (0.5%). The age ranged from 18 to 59 years ($M = 29.35$, $SD = 8.53$). Forty participants (20%) reported having a postgraduate or professional degree and a further 90 (45%) having a bachelor's degree. In terms of their employment status, 111 participants (55.5%) reported working full time, 34 (17%) part time and 35 (17.5%) were students, while the remaining participants were not working.

Data screening revealed that two participants provided straight-line responses in their second questionnaire completion (presumably due to loss in motivation). We therefore removed data related to these inattentive completions, which happened to be honest completion for participant ID = 38 and faking completion for participant ID = 124. This resulted in 99 full records for faking and 99 full records for honest condition.

Materials and Measures

General Mental Ability (GMA)

We used the Leadership Ability Quotient (LAQ; de Klerk-van Someren & Rockson, 2023)— mental ability test containing 26 questions measuring participants' deductive, numerical, and abstract reasoning abilities. The test is limited to 25 min, after which the assessment ends automatically, even if not all questions are completed. Participants took the assessment in one sitting and could not pause or exit the test and resume later. Moreover, participants could not go back to

previous questions and could only move forward within the test. Before starting the actual ability test, participants were presented with six practice questions to familiarize themselves with the question types. To measure participants' general mental ability, we obtained their deductive, numerical, and abstract reasoning subscale scores. Overall GMA was operationalized as the sum of three subscales.

Job Description

To provide an ecologically valid context for the high-stakes (faking) condition, we sourced a real-world job description of Director of Market Research and Analytics presented in Appendix A. We then asked participants to complete the LSQ as if they were applying for this job.

Trait Importance Ratings

In faking condition, after reading the job description, participants were presented with definitions of all 24 LSQ scales. Their task was to read each definition and to decide how important having these qualities is to success in the Director of Market Research and Analytics role. Participants rated the importance of each quality on a 4-point Likert scale: 0 = *irrelevant*, 1 = *less relevant*, 2 = *desirable*, 3 = *essential* (to job success).

Ability To Identify Criteria (ATIC)

Prior to the data collection, eight leadership consultants and assessment experts familiar with the LSQ rated the importance of 24 scales to the Director of Market Research and Analytics role using the same 4-point Likert scale as the participants. The mean expert ratings across all 24 scales rounded to the nearest integer served as the 'right answers' against which to evaluate participants' importance ratings. To determine ATIC scores for each participant, we computed the Spearman correlation coefficient between ratings of importance for 24 constructs provided by that participant and the 'right answers' provided by the experts. The Spearman rank-order correlation was chosen because the 24 ratings are ordinal and their distributions within each participant are not expected to be normal. Thus, our measure of ATIC reflected the (relative) order of priorities given to personality traits when faking, a key factor in successful faking forced-choice questionnaires (Fuechtenhans & Brown, 2023). We note that this ATIC measure may confound job knowledge (of those participants who are familiar with the described job) and ability to recognize (unfamiliar) evaluation criteria, further discussed in Limitations.

Leadership-Style Questionnaire (LSQ)

The Leadership Styles Questionnaire (de Klerk-van Someren et al., 2023) is designed to assess personality traits associated with leadership positions in the workplace. The LSQ measures 24 scales, categorized into five domains: Problem Solving, People Focus, Leadership, Execution and Business Drive. A unique aspect of the LSQ is that it utilizes a '7-point total' response format, whereby statements describing ways of behaving at work are presented in blocks of three (triplets), and test takers are asked to indicate how much more, or less, each statement describes them by distributing a total of 7 points between the three statements; however, the maximum number of points a single statement can be assigned is 5, and the minimum is 0. An example block is given in Fig. 1.

Formally, this fixed-point-total format is compositional (Brown, 2016); however, the small number of points available for distribution between statements brings in psychological aspects of graded preferences. Thus, according to empirical research (Brown, 2022), when responding to this format, 81.2% of $N = 250$ respondents reported thinking about *ordered categories*, for example, "4 points felt like 'much more' than 1 point, or 2 points felt like 'slightly more' than 1 point". This contrasts with only 14.4% participants who reported thinking about *ratios* of points, for example,

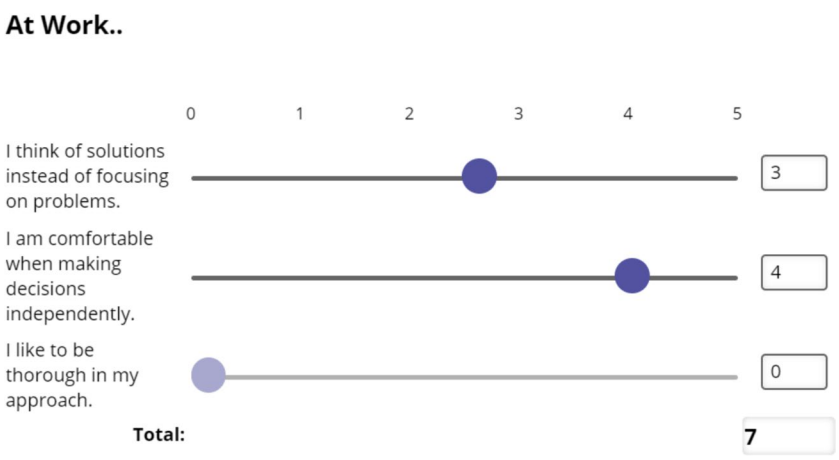

Fig. 1 An example block from the Leadership Styles Questionnaire

"4 points felt like 'twice as much' as 2 points, or 3 points felt like '1.5 times as much' as 2 points". The remaining 4.4% reported of "thinking something else" when responding to the 7-point fixed total format. It appears that psychologically, distributing 7 points between 3 alternatives is a borderline case between ordinal and ratio data, with most participants considering it graded preference (yielding ordinal data) rather than compositions (yielding ratio data). Thus, this research provides an opportunity to investigate the effects of faking behavior on quantitative ipsative comparisons that may generalize to both graded comparisons and compositions.

To derive participants' personality scores, we used the LSQ scoring protocol based on Thurstonian IRT for graded preference data (Brown & Maydeu-Olivares, 2018), where 24 scale scores are estimated maximizing the posterior likelihood distribution (maximum-a-posteriori or MAP). These scores are inherently standardized based on the normative sample described in the LSQ Technical Manual (de Klerk-van Someren et al., 2023) and thus enable meaningful cross-condition comparisons.

Process Data: Response Latencies and Click Counts

Using the Qualtrics integrated timer function, we recorded the time (in milliseconds) between the page download (each LSQ block was presented on a separate page) and the last click to submit final answers on each block (i.e., page submission). Using the Qualtrics integrated click count function, we recorded how many clicks participants made on each block to reach the final selections.

Data Preparation and Planned Analysis

We tested the hypotheses in two different ways. First, we used *observed* summaries of participants' scores across LSQ scales, and of their process data (response latencies or click counts) across LSQ blocks to test between-person effects. This traditional approach promotes simplicity of analysis and interpretation and allows direct comparison with results of previous research studies (e.g., Vasilopoulos et al., 2006). These analyses were conducted in IBM SPSS Statistics version 31.

Second, we used cross-classified multilevel modeling to analyze all available personality or process data and identify effects at their appropriate levels: (1) 'within' (response-per-participant) level; (2) 'between-scale' or 'between-block' level for describing *latent* summaries for LSQ scales or blocks, respectively; and (3) 'between-person' level for describing *latent* summaries of individual differences. Cross-classified models are fairly novel in modeling response process data (Feng & Cai, 2024; Goldhammer et al., 2024) and allow observations (for example, person's response time for a forced-choice block) to be nested into both person and block levels, separating effects due to individual differences from effects due to block features. The key difference with hierarchical multilevel models is that the latter assumes people take different, non-overlapping blocks, which is obviously not the case with a fixed personality questionnaire. Cross-classified analyses were conducted in Mplus version 9 using Bayesian estimator. Figures 2, 3 and 4 use the Mplus user's guide (Muthén & Muthén, 2017) notation for multilevel models.

Observed Summaries of Participants' Personality and Process Data

Within the first 'observed variable' approach, we created the following summary variables. To summarize participants' personality profiles, we averaged each participant's scores on nine LSQ scales[1] rated by the experts as essential to the Director of Market Research and Analytics role (marked with asterisk in Table 1) within each condition. Previous research has used such composite scales (e.g., Jeong et al., 2017) to mimic selection decisions which are often made using composites. The *difference* in this LSQ 'job-essential' score[2] between the faking and honest conditions was used to operationalize the ability to fake successfully and was used in testing hypotheses H1a and H1b.

To create a convenient and robust variable for summarizing response latencies per participant per condition that overcomes the problem of extreme positive skew often present in response latencies when a participant takes a break, we computed the median response time for each participant across all LSQ blocks within each condition. This observed summary of participants' response times was used in descriptive statistics and was further log-transformed to overcome the positive skew for testing hypotheses H2a and H2b.

To summarize click counts per participant per condition, we computed the mean click count for each participant across all LSQ blocks within each condition. This observed summary of participants' click counts was used

[1] We tested for significance the 'job-essential' LSQ score rather than all scales individually to a) reduce the risk of Type I errors, as investigating all scales individually would result in many separate tests, and b) provide a clear operationalization of 'faking success'. We do, however, provide descriptive statistics and effect sizes for all LSQ scores separately in Table 1.

[2] To address potential issues with decreased reliability of the difference score that is observed when scores on repeated measures are strongly correlated (see for example McDonald, 1999), latent difference score analysis following McArdle (2001) was carried out. We report the results of such analysis using the latent difference scores in footnotes to the main results.

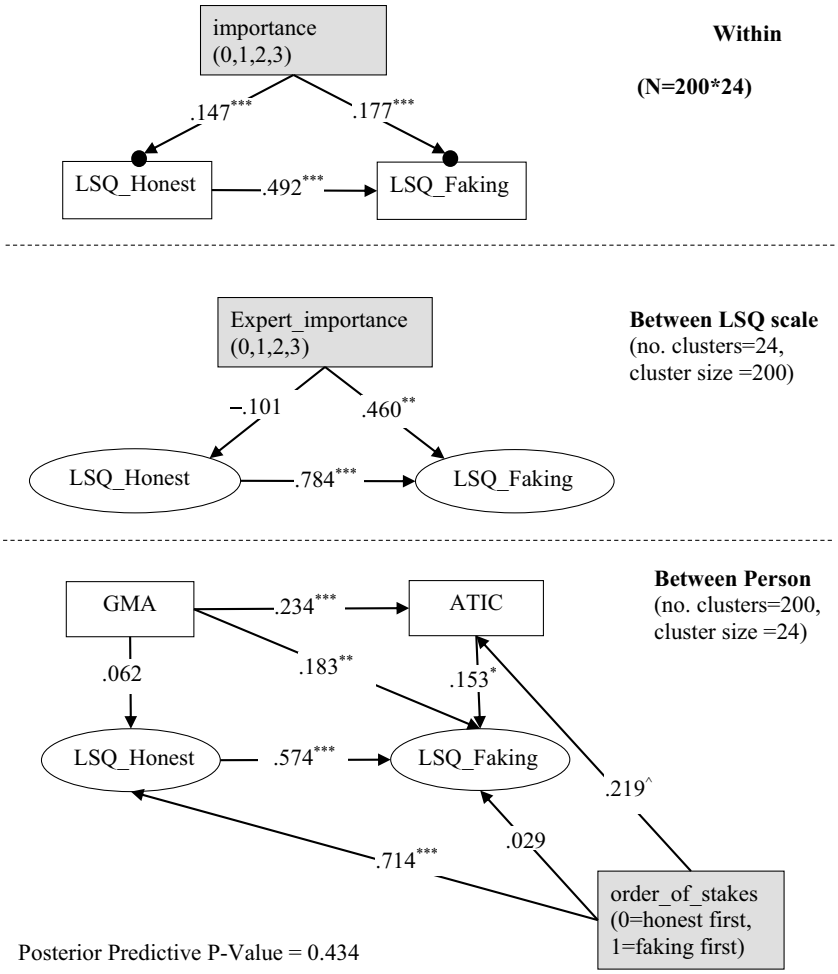

Fig. 2 The cross-classified GMA-ATIC mediation model for LSQ scale scores. *Note*. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, ^ $p < 0.10$ (one-tailed). Coefficients are standardized on all observed continuous variables. For regressions involving predictors in shaded boxes (binary: order_of_stakes; and ordinal: importance and Expert_importance), only DV is standardized, so that the coefficients are interpreted as DV change in DV standard deviation units when predictor changes from 0 to 1

in descriptive statistics, and was further log-transformed to overcome the slight positive skew for testing hypotheses H3a and H3b.

To test hypotheses H1a, H2a and H3a**,** we conducted mixed 2×2 ANOVAs examining the effect of stakes (Within: Honest, Faking) and order of stakes (Between: Honest first, Faking first) on the LSQ job-essential score, log-transformed median response latency and log-transformed mean click count, respectively. The counterbalancing order in which the honest and faking condition was presented to participants was included in this analysis as a potential moderator of relationships between the experimental condition (stakes) and dependent variables. To test hypotheses H1b, H2b and H3b**,** we tested the significance of Pearson correlation between GMA and the LSQ essential score difference from honest to faking condition (or response latency difference, or click counts difference, as appropriate).

Cross-Classified Modeling of Personality and Process Data

To test hypotheses H1c, H2c and H3c**,** we conducted cross-classified path analyses examining the effects of GMA and ATIC on the *latent* intercept of participants' personality scores across all LSQ scales (or response latencies / click counts across all blocks) in faking condition, controlling for the respective score in honest condition. These models are depicted in Figs. 2, 3 and 4. Predicting the outcome in faking condition while controlling for the score in honest condition is a type of auto-regression model (McArdle, 2009), which allows prediction of a base-free measure of change using external variables of interest. This planned hypothesis testing was carried out at the between-person level, where GMA and ATIC are measured, and we added the order of stakes as another between-person predictor of the dependent variables and ATIC (since ATIC was measured after the stakes were manipulated), but not GMA (since this was measured before manipulation of stakes).

The cross-classified models also enabled us to further test hypotheses H1b, H2b and H3b. This time, we used latent (rather than observed) summaries of DVs at between-person level, testing for significance the *regression coefficients* of respective residualized DVs in faking condition on GMA.

Finally, we used cross-classified models to test other relevant predictors of DV at their appropriate levels. In the LSQ

Table 1 Descriptive statistics for LSQ scale scores under honest and faking conditions

Domain	Scale	Expert rating ($N=8$)		Honest ($N=199$)		Faking ($N=199$)		Faking—Honest ($N=198$)	
		Mean	SD	Mean	SD	Mean	SD	d_Z	Corr
Problem solving	Solution focus*	2.50	0.53	−0.03	0.78	0.23	0.76	0.32	0.49
	Analysis*	2.88	0.35	0.13	0.91	0.41	0.87	0.33	0.56
	Curiosity	1.75	0.71	0.07	0.71	0.12	0.64	0.08	0.48
	Foresight	2.38	1.19	0.16	0.85	0.27	0.74	0.13	0.46
	Innovation*	2.75	0.46	0.05	0.73	0.24	0.79	0.26	0.54
People focus	Sociability	0.75	0.89	0.00	0.93	0.17	0.81	0.20	0.52
	Positivity	1.38	0.92	−0.16	0.83	−0.05	0.79	0.17	0.65
	Self-awareness	1.13	0.83	0.08	0.79	−0.16	0.81	−0.29	0.50
	Collaboration*	2.50	0.76	−0.28	0.82	0.05	0.84	0.38	0.43
	Mentorship*	2.88	0.35	−0.28	0.87	0.00	0.94	0.30	0.44
	Empathy	1.50	0.76	−0.46	0.87	−0.32	0.81	0.20	0.66
Leadeship	Independence	1.75	1.04	−0.25	0.86	−0.58	0.85	−0.37	0.45
	Adaptability	1.50	1.07	−0.28	0.78	−0.13	0.76	0.20	0.52
	Influence*	2.50	0.76	−0.21	0.80	0.05	0.73	0.31	0.42
	Control*	2.63	0.52	−0.24	0.99	0.15	0.95	0.42	0.55
	Integrity	1.13	0.83	−0.16	0.80	−0.37	0.84	−0.25	0.54
	Willing to Challenge	1.50	0.93	−0.12	0.80	0.07	0.73	0.25	0.53
	Resilience	0.88	0.64	−0.09	0.88	0.13	0.78	0.29	0.59
Execution	Detail focus	2.13	0.99	−0.06	0.83	0.06	0.76	0.15	0.57
	Organization	1.38	0.52	−0.05	0.70	0.02	0.77	0.10	0.54
	Reliability	2.00	0.53	−0.30	0.89	−0.09	0.87	0.28	0.64
Business drive	Achievement	1.63	0.74	0.12	0.85	0.42	0.75	0.39	0.55
	Opportunity focus*	2.50	0.76	0.37	1.02	0.72	1.00	0.34	0.46
	Vision*	3.00	0.00	0.02	0.80	0.46	0.90	0.47	0.42
	Median	**1.88**	**0.76**	**−0.08**	**0.83**	**0.06**	**0.80**	**0.26**	**0.53**

* LSQ scale with median expert rating of 2.5 and over considered 'essential' for the Director of Market Research and Analytics role

scale model (Fig. 2), we used participants' ratings of importance of each LSQ trait to success in the Director of Market Research and Analytics role as within level predictors of their LSQ scores under faking condition. We should expect that faking LSQ scores will be higher for scales considered more important by the participant. We used the rounded experts' mean importance rating of each LSQ scale to the target job (that was used for computing participants' ATIC) as a predictor of participants' mean score shift at between-scale level. We expect that participants' faking average scores will be higher for scales considered more important by experts.

In the models of response latencies (Fig. 3) and click counts (Fig. 4), we used the mean absolute distance (MAD) between participants' responses to each block (7 points assigned to three items) in the faking and honest conditions as an index of actual 'faking' or 'editing' behavior to predict their response latencies / clicks at within level. If the editing mechanism of faking described by Holtgraves (2004) and Fuechtenhans and Brown (2023) is true, we should see latencies and clicks increase together with increased editing (indexed by MAD). We used block's importance to fulfilling job criteria (operationalized as the maximum of experts' importance ratings among three traits measured by three items in the block) as a between-block predictor of participants' mean latency / clicks. If criteria recognition is indeed important to faking, we shall see increased latencies / clicks in blocks that involve more job-important traits. Finally, we used block order (1 to 88) as another predictor at between-block level. According to a meta-analysis of response time in cognitive tasks (Scharfen, 2018), we may expect that the average time and click counts will reduce with every completed block due to practice effect.

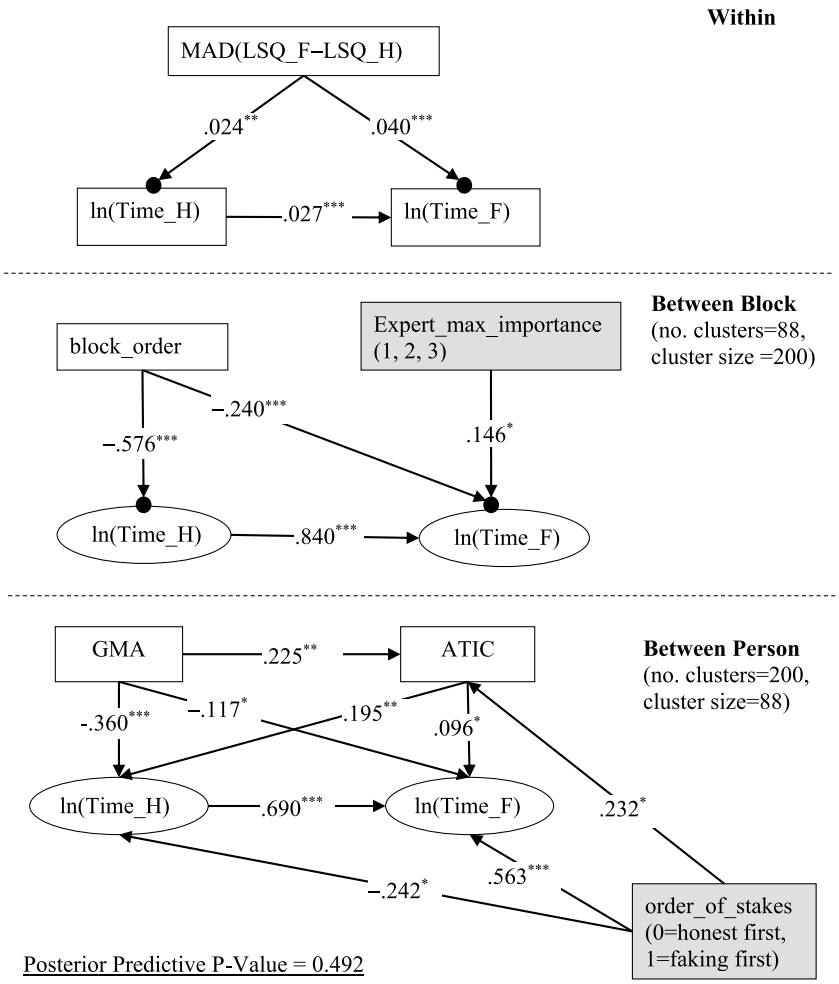

Fig. 3 The cross-classified GMA-ATIC mediation model for response latencies. *Note*. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (one-tailed). MAD = mean absolute discrepancy between faking and honest LSQ responses in the block. Coefficients are standardized on all observed continuous variables. For regressions involving predictors in shaded boxes (binary: order_of_stakes, and ordinal: Expert_max_importance), only the dependent variable is standardized, so that the coefficients are interpreted as DV change in DV standard deviation units when predictor changes from 0 to 1

Results

GMA and ATIC

GMA subscales correlated with each other moderately to strongly; deductive with numerical reasoning $r = 0.47$; deductive with abstract $r = 0.48$, and numerical with abstract $r = 0.46$. The sum of three subscales making up the overall GMA had the reliability $\alpha = 0.723$.

GMA and ATIC scores correlated positively and significantly, $r = 0.24$, $p < 0.001$. This result provides the necessary condition for testing hypotheses about the mediating role of ATIC in relationships between GMA and personality (H1c), GMA and response latency (H2c), and GMA and click counts (H3c). ATIC also correlated with ability subscales as follows: deductive $r = 0.29$ ($p < 0.001$); numerical $r = 0.17$ ($p = 0.015$); abstract $r = 0.17$ ($p = 0.019$), indicating a larger overlap with deductive reasoning.

Since ATIC was measured in either first or second completion depending on when faking condition was administered, we tested for any ordering effect. An independent samples t-test was marginally significant, $t(194) = 1.765$, two-sided $p = 0.079$, with ATIC being higher in faking-first (M = 0.28, SD = 0.196) than in honest-first ($M = 0.23$, SD = 0.204) condition, amounting to small effect size (Cohen's d for independent samples with pooled standard deviation used as the denominator was 0.252).

Personality Scores

Descriptives and Faking vs. Honest Comparisons

The LSQ job-essential score had the reliability $\alpha = 0.712$ in the honest condition and $\alpha = 0.829$ in the faking condition. This suggests that when instructed to respond as if applying for a job, the participants were more consistent in responses to items measuring the nine scales essential for job success than they were in the honest condition.

Table 1 provides descriptive statistics for participants' scores on all 24 LSQ scales under honest and faking conditions, as well as for the differences between faking and honest scores. Under the faking condition, participants increased their scores on average for 21 out of 24 LSQ scales, including all nine LSQ scales rated by experts as

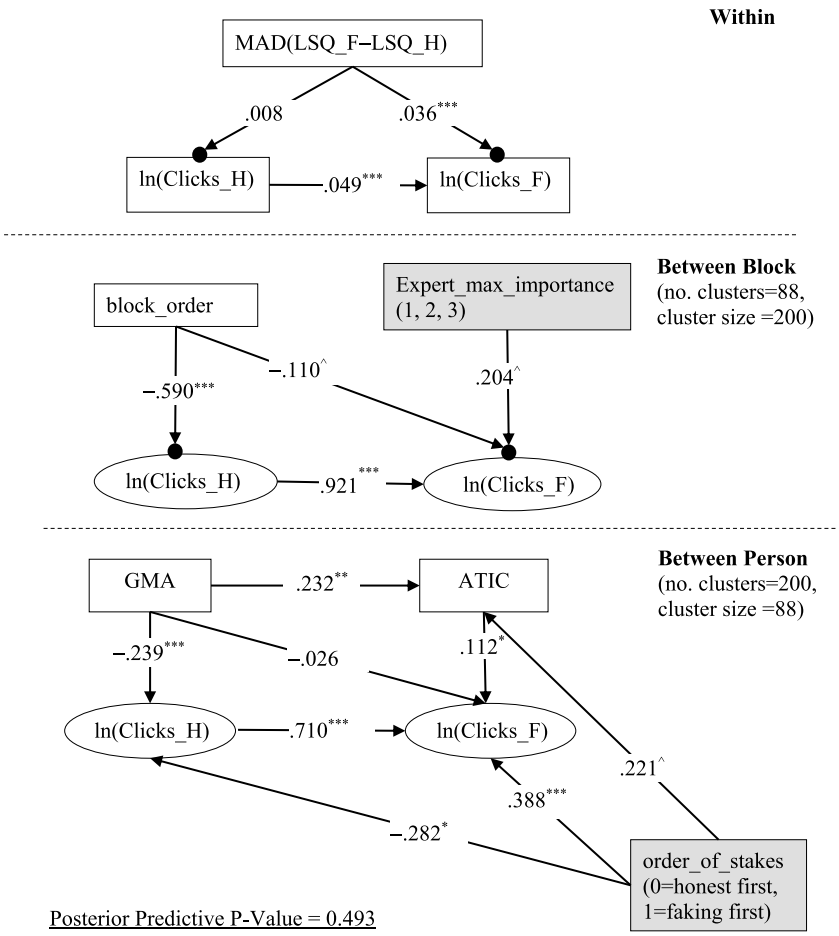

Fig. 4 The cross-classified GMA-ATIC mediation model for click counts. *Note.* $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, ^ $p<0.10$ (one-tailed). MAD = mean absolute discrepancy between faking and honest LSQ responses in the block. Coefficients are standardized on all observed variables. For regressions involving predictors in shaded boxes (binary: order_of_stakes, and ordinal: Expert_max_importance), only the dependent variable is standardized, so that the coefficients are interpreted as DV change in DV standard deviation units when predictor changes from 0 to 1

'essential' (marked with an asterisk in Table 1). However, the effect sizes measured by Cohen's *d*-statistics for paired samples (d_Z), where the standard deviation of the difference scores is used as the denominator, were small (median d_Z 0.26) with the largest inflation $d_Z = 0.47$ for scale Vision. The LSQ 'job-essential' score (the mean of nine LSQ 'essential' scale scores per participant) was distributed relatively normally in both conditions (see Figure S2 in the Supplement) and was inflated to a medium effect ($d_Z = 0.58$) when under faking ($M = 0.26$, $SD = 0.56$), compared to honest ($M = -0.05$, $SD = 0.48$) condition.

A mixed ANOVA examined the effect of stakes (Within: Honest, Faking) and order of stakes (Between: Honest first, Faking first) on LSQ job-essential scores. As expected in **H1a**, the main effect of stakes was significant, $F(1,196) = 67.30$, $p < 0.001$, $\eta^2 = 0.26$, indicating job-essential score increase in the faking condition. The main effect of order of stakes was also significant, $F(1,196) = 19.25$, $p < 0.001$, $\eta^2 = 0.09$, whereby participants completing faking condition first had higher scores across both stakes. No significant interaction was found, $F(1,196) = 0.47$, $p = 0.494$, $\eta^2 = 0.00$. Descriptive statistics for all between and within conditions are given in Table 2,

Table 2 Descriptive statistics for LSQ job-essential score, median response latency and mean click count by stakes and order of stakes

Dependent variable	Stakes	Order of stakes	*N*	Mean	St. dev
LSQ_essential	Honest	Faking first	99	0.10	0.498
		Honest first	100	−0.20	0.401
		Total	199	−0.05	0.475
	Faking	Faking first	100	0.38	0.546
		Honest first	99	0.13	0.559
		Total	199	0.26	0.565
medianTime	Honest	Faking first	99	12.65	5.860
		Honest first	100	13.91	5.441
		Total	199	13.28	5.674
	Faking	Faking first	100	16.96	8.623
		Honest first	99	13.93	6.606
		Total	199	15.45	7.816
meanClicks	Honest	Faking first	99	5.11	1.532
		Honest first	100	5.54	1.578
		Total	199	5.33	1.566
	Faking	Faking first	100	5.64	1.750
		Honest first	99	5.22	1.632
		Total	199	5.43	1.701

Table 3 Correlations of participants' GMA and ATIC with LSQ job-essential score, median response latency, mean click count and their differences between conditions

	DV type	DV	Honest	Faking	Faking – Honest
GMA	Personality	LSQ_essential	−0.057	0.115	0.173*
	Latency	medianTime	−0.352***	−0.345***	−0.108
		ln(medianTime)	−0.309***	−0.296***	−0.003
	Clicks	meanClicks	−0.214**	−0.130	0.087
		ln(meanClicks)	−0.234***	−0.143*	0.106
ATIC	Personality	LSQ_essential	0.094	0.229**	0.152*
	Latency	medianTime	0.057	0.122	0.099
		ln(medianTime)	0.108	0.192**	0.123
	Clicks	meanClicks	−0.050	0.093	0.175*
		ln(meanClicks)	−0.064	0.095	0.186**

***$p<0.001$, **$p<0.01$, *$p<0.05$ (two-tailed). Sample sizes for corelations with GMA were $N=199$ for either Honest or Faking condition, and $N=198$ for their difference. Sample sizes for corelations with ATIC were $N=195$ for either honest or faking condition, and $N=194$ for their difference. The sample size for ATIC was smaller because 4 participants rated all scales 'essential', therefore correlation with expert ratings was not defined for them

and the estimated marginal means plot can be found in the Supplement, Figure S4.

In addition to considering the LSQ score shifts across conditions, we also computed correlations between the honest and faking LSQ scale scores as the measure of stability of participants' ordering across conditions. For all job-essential LSQ scales, these correlations were remarkably similar, from 0.42 to 0.56 (see Table 1), indicating that the graded comparisons assessment was relatively robust to the effects of instructed faking in this study.

Table 1 also provides descriptive statistics for experts' ratings of importance of each LSQ scale to the Director of Market Research and Analytics role. Interestingly, the experts' average ratings of importance across 24 scales correlated very weakly with the participants' average scores in honest condition (Spearman rank-order correlation $\rho=0.10$), but strongly with average scores in faking condition ($\rho=0.47$) and stronger yet with d_z statistic of faking-honest difference ($\rho=0.63$). This suggests that objectively important traits for selection get inflated most.

Relationships with GMA and ATIC

Table 3 reports correlations between GMA, ATIC, and the LSQ job-essential scores by condition, as well as the LSQ score difference between conditions, and Table S3 in the Supplement breaks these down by the order of stakes. While the LSQ essential scores in honest or faking conditions did not correlate with GMA, the LSQ between-condition difference[3] correlated with GMA positively and significantly, $r=0.173$ ($p=0.015$), thus providing support for Hypothesis **H1b** of the positive role of GMA in ipsative test score inflation when faking. The LSQ between-conditions difference also correlated with ATIC, $r=0.152$ ($p=0.034$), providing one essential condition for testing mediation Hypothesis **H1c**. Interestingly, ATIC also correlated with the LSQ essential score in faking condition, $r=0.229$ ($p=0.001$), but not in honest condition.

Cross-Classified Path Analysis

To test effects of predictors on LSQ score inflation, we fitted a cross-classified multilevel path model depicted in Fig. 2. In this model, at each level of analysis, the LSQ scale score in faking condition (or the latent mean of these scores across 24 scales at between-person level, or across 200 participants at between-scale level) is regressed on the same LSQ scale score in honest condition, thus representing a base-free measure of change. The benefit of working with this residualized faking score (rather than manifest faking minus honest difference score) is that the effects of covariates such as GMA or ATIC can be established on both the base (honest) score and the base-free change. We use appropriate covariates at each level to predict dependent variables, as described in the "Cross-Classified Modeling of Personality and Process Data" section. The cross-classified Bayesian model had posterior predictive p-value (PPP) of 0.434, which is close to the ideal value 0.5 and indicates that the model's predictions are consistent with the observed data.

Results at between-person level show a *positive* direct effect of GMA on the residualized mean faking score across all LSQ scales ($\beta=0.183$, one-tailed[4] $p=0.002$), thus

[3] To address potential issues with inflated unreliability of the LSQ difference score, latent difference score analysis following McArdle (2001) was carried out. The resulting corelation with GMA was $r=0.171$, $p=0.014$; and with ATIC $r=0.166$, $p=0.018$.

[4] In Bayesian models, all p-values given are one-tailed.

supporting Hypothesis **H1b**. There is further *positive* indirect effect via ATIC ($\beta = 0.036$, $p = 0.010$). Therefore, ATIC mediates the relationship between GMA and self-reported personality in the faking condition, providing support for Hypothesis **H1c**. While participants with higher GMA tend to produce larger increases on personality traits when faking, their higher Ability to Identify Criteria (ATIC) plays a significant role in this inflation. Order of stakes was found to have a direct positive effect on participants' mean honest LSQ score and no direct effect on faking score, corroborating our ANOVA findings of its main effect on personality scores across conditions.

Results at between-scale level show a significant *positive* effect of expert-rated trait importance on the participants' average residualized faking score on that trait (β_Y[5] $= 0.460$, $p = 0.002$), suggesting that objective trait importance positively influences the score inflation in high stakes. Finally, the within level results show a significant *positive* effect of participant-rated trait importance on both their honest responses ($\beta_Y = 0.147$, $p < 0.001$) and residualized faking responses ($\beta_Y = 0.177$, $p < 0.001$), suggesting that perceived trait desirability influences not only a person's score shift when faking but also their honest response.

Response Latencies

Descriptives and Faking vs. Honest Comparisons

Figure S2 in the Supplement provides the distribution of participants' median response time across blocks by condition, and Table 2 provides respective descriptive statistics. Participants were slower in completing the questionnaire under faking condition ($M = 15.45$, $SD = 7.82$) than honest condition ($M = 13.28$, $SD = 5.67$).

A mixed ANOVA examined the effects of stakes and their order on participants' median response latency.[6] As expected in **H2a**, the main effect of stakes was significant, $F(1,196) = 28.77$, $p < 0.001$, $\eta^2 = 0.13$. The main effect of order of stakes was not significant, $F(1,196) = 1.07$, $p = 0.301$, $\eta^2 = 0.00$, but there was a significant interaction between stakes and order of stakes, $F(1,196) = 26.87$, $p < 0.001$, $\eta^2 = 0.12$, indicating that the response time change from honest to faking condition differed depending on the order in which these conditions were presented. As the observed means in Table 2 and the estimated marginal means in the Supplemental Figure S5 show, when instructed to fake, participants spent significantly longer responding to LSQ only if it was their first completion. If the first completion was under honest instructions, the subsequent faking instruction did not result in longer response times. We will return to this effect in the "Discussion" section.

Relationships with GMA and ATIC

The median latencies correlated with GMA negatively and significantly in both honest ($r = -0.352$, $p < .001$) and faking conditions ($r = -0.345$, $p < .001$), as shown in Table 3. This suggests that participants of higher mental ability take less time when completing ipsative items regardless of condition, corroborating previous research. However, the between-condition time difference did not correlate significantly with GMA, $r = -0.108$ ($p = 0.130$); thus, Hypothesis **H2b** is not supported. The same conclusions could be reached after log-transforming the median latency variables to meet the normality assumption of Pearson's correlation test (see Table 3).

ATIC showed a different pattern, correlating positively but insignificantly with the median latencies in honest and faking conditions and their difference (see Table 3). Correlations with log-transformed median latencies showed the same trend, but statistical significance was reached for the latency in faking condition ($r = 0.192$, $p = 0.007$).

Cross-Classified Path Analysis

To overcome large positive skew in distributions of response times per block, we removed responses with extremely long times over 400 ms, which comprised less than 0.01% of all honest responses and less than 0.02% of faking responses. We then transformed the raw latency values using natural (base e) logarithm. We used level-appropriate covariates of latencies as described in the "Cross-Classified Modeling of Personality and Process Data" section. The resulting cross-classified Bayesian model depicted in Fig. 3 had posterior predictive p-value $= 0.492$, indicating that the model's predictions are consistent with the observed data.

Results at between-person level show a significant *negative* effect of GMA ($\beta = -0.117$, one-tailed $p = 0.011$) on residualized faking latency; and a *positive* indirect effect via ATIC ($\beta = 0.022$, $p = 0.033$). Therefore, ATIC partially mediates the relationship between GMA and response latency in the faking condition, supporting Hypothesis **H2c**, and this mediation effect is in the opposite direction of the total effect of GMA on latencies. While participants with higher GMA are faster in responding generally whether in low or high stakes, their higher ATIC slows their responses in high stakes. Unexpectedly, an *indirect* effect of ATIC

[5] β_Y refers to regression coefficient standardized on DV (continuous variable) but not on IV (binary or ordinal variable), so that the coefficient is interpreted as DV change in DV standard deviation units when predictor changes from 0 to 1.

[6] Despite the slight positive skew in the distribution of median response times, the natural (base e) logarithm transformation did not change the results of mixed ANOVA, therefore the results are reported for non-transformed DV to aid interpretation.

on faking latency via honest latency was found ($\beta = 0.135$, $p = 0.003$). Not including this effect in the model resulted in inferior model fit (PPP = 0.250). This suggests that the Ability to Identify Criteria affects response latencies in the honest condition too, an interesting phenomenon that we will attempt to interpret in the "Discussion" section. With all direct and indirect effects considered, GMA had *negative* **total** effects on both honest latency (total $\beta = -0.360$, $p < 0.001$) and faking latency (total $\beta = -0.314$, $p < 0.001$), thus supporting Hypothesis **H2b**.

Order of stakes was found to have a direct positive effect on participants' mean faking latency and an indirect negative effect via honest latency, corroborating our ANOVA findings of its moderating effect on response latencies and stakes.

Results at between-block level show a significant positive effect of expert-rated importance of traits measured in blocks ('Expert_max_importance' covariate) on the participants' average residualized faking latency ($\beta_Y = 0.169$, $p = 0.004$). There is also a highly significant negative effect of block order on residualized faking latency, directly and via honest latency. Finally, results at within level show statistically significant (due to very large sample size at this level) but a negligible in size positive effect of participant observed faking (mean average distance or MAD between faking and honest responses in the block) on residualized faking latency, directly and via honest latency.

Click Counts

Descriptives and Faking vs. Honest Comparisons

Figure S3 in the Supplement provides the distribution of participant's mean click counts across blocks by condition. Participants made slightly more clicks when completing the questionnaire under faking ($M = 5.43$, $SD = 1.70$) than honest ($M = 5.33$, $SD = 1.57$) condition. A mixed ANOVA examined the effect of stakes and order of stakes on participants' mean clicks.[7] Contrary to the expectations of **H3a**, the main effect of stakes was not significant, $F(1,196) = 1.21$, $p = 0.273$, $\eta^2 = 0.01$. The main effect of order of stakes was not significant either, $F(1,196) = 0.004$, $p = 0.950$, $\eta^2 = 0.00$. However, a significant interaction between stakes and order of stakes was found, $F(1,196) = 22.20$, $p < 0.001$, $\eta^2 = 0.10$, indicating that the number of clicks difference between honest and faking condition depended on the order in which these conditions were presented. According to the descriptive statistics in Table 2 and the estimated marginal means in the Supplemental Figure S5, participants made significantly more clicks responding to LSQ in whatever was their first completion. We will return to this effect in the "Discussion" section.

Relationships with GMA and ATIC

The log-transformed mean click counts in both honest ($r = -0.234$, $p < 0.001$) and faking conditions ($r = -0.143$, $p = 0.044$) correlated with GMA negatively and significantly, thus repeating the pattern of correlations for response latencies. The same trend was observed for mean clicks before log transformation (see Table 3). This suggests that participants of higher mental ability make fewer clicks when completing ipsative items regardless of condition. However, the between-condition click difference did not significantly correlate with GMA, either before or after log transformation, thus **H3b** is not supported.

ATIC did not correlate with the mean click counts in either honest or faking conditions. However, ATIC correlated with the log-transformed between-condition clicks difference positively and significantly, $r = 0.186$ ($p = 0.010$). The same conclusions could be reached based on mean click counts before log transformation (see Table 3).

Cross-Classified Path Analysis

To overcome positive skew typical in count data with small counts over-represented, we used the natural (base e) logarithm transformation before entering click counts per block in a cross-classified path model depicted in Fig. 4. The model had posterior predictive p-value = 0.493, indicating that the model's predictions are consistent with the observed data.

Results at between-person level show that with all direct and indirect effects considered, GMA had *negative* **total** effects on both honest click counts (direct $\beta = -0.239$, $p < 0.001$) and faking click counts (total $\beta = -0.170$, $p = 0.008$), thus supporting Hypothesis **H2c**. The direct effect of GMA on residualized faking clicks was insignificant ($\beta = -0.026$, one-tailed $p = 0.314$), and this was counteracted by an indirect effect via ATIC which was significant and *positive* ($\beta = 0.026$, $p = 0.019$). Therefore, ATIC fully mediates the relationship between GMA and residualized clicks in faking condition, supporting **H3c**. This suggests that while participants with higher GMA make fewer clicks in ipsative blocks across conditions, their better understanding of selection criteria (ATIC) makes them change their answers more in faking condition. Order of stakes was found to have a direct positive effect on participants' mean faking clicks and an indirect negative effect via honest clicks, corroborating our ANOVA findings of its moderating effect on click counts and stakes.

[7] Despite the slight positive skew in the distribution of mean clicks, the natural (base e) logarithm transformation did not change the results of the mixed ANOVA, therefore the results are reported for non-transformed DV to aid interpretation.

Results at between-block level reveal a marginally significant positive effect of expert-rated importance of traits measured in the block on the participants' average residualized faking click counts ($\beta_Y = 0.204$, one-tailed $p = 0.051$). There is also highly significant negative effect of block order on residualized faking click counts, mostly via honest click counts. Finally, within level results show a negligible in size but significant positive effect of MAD between faking and honest responses on residualized faking click count on the block.

Summary of Findings

The aim of this study was to investigate the role of general mental ability and Ability to Identify Criteria in responding to a 'fixed-point-total' ipsative personality questionnaire in high-stakes settings. We found that on average, participants significantly increased the *job-related LSQ scale scores* under high stakes, although the effect sizes for individual LSQ scales were small to medium. The manipulation of stakes resulted in the same average score increase whether high or low stakes were administered first; but those who completed faking condition first, had uniformly higher LSQ scores across conditions.

We found that the ipsative score inflation in high stakes is moderated by GMA so that participants of higher mental ability tend to show a higher increase between their honest and faking scores. Ability to Identify Criteria has a similar effect, with higher ATIC associated with higher personality score inflation when faking. Moreover, ATIC partially mediates the relationship between GMA and personality scores in high stakes, contributing to the overall positive effect of GMA. Our cross-classified modeling supports these conclusions in relation to an **average** LSQ scale, not only the nine job-essential scales. Additionally, this modeling revealed that scales objectively important for the target job are inflated more in high stakes, and that scales perceived as important by participants are inflated not only when faking but also when responding honestly.

In terms of *response latencies*, we found that on average, participants were slower in responding to the questionnaire under high stakes when this was their first questionnaire completion. There was no difference in latency if high-stakes completion followed low stakes. However, participants with higher GMA responded significantly faster to the personality assessment regardless of stakes, but it was ATIC that drove the latency increase in high stakes. Participants with higher ATIC tended to respond slower in high stakes, and this effect was corroborated by our cross-classified modeling of responses to individual blocks. Additionally, this modeling revealed that blocks measuring traits that are objectively more important for the target job caused larger time increases from low to high stakes.

Analysis of *click counts* revealed that there was no significant difference between average click counts across conditions, contrary to expectations. Yet further examination revealed that the effect was masked by counterbalancing of conditions' order. Participants made more clicks in their first completion, whichever stakes that happened to be. Participants with higher GMA tended to make fewer clicks overall, irrespective of stakes. ATIC did not have any significant effect on click counts within conditions; however, it correlated with base-free increase in click counts in high stakes. Cross-classified modeling corroborated these results and suggested that blocks measuring traits that are objectively more important for the target job may have caused larger increase in clicking from low to high stakes. We also found large and decisive effects of block order, whereby both latencies and clicks went down as participants progressed through the questionnaire.

Exploring the interplay between GMA and ATIC, we found that ATIC partially mediates the relationship between GMA and response latencies and fully mediates the relationship between GMA and click counts, with opposing effects to that of GMA. Where GMA speeded up responses, ATIC slowed them down making the overall effect of stakes weaker; and where GMA reduced the number of clicks, ATIC increased them, again making the overall effect weaker.

Discussion

The Effect Sizes of Ipsative Score Inflation in High Stakes

The average effects of faking reported in a recent meta-analysis by Speer et al. (2023) are smaller for forced-choice questionnaires ($d = 0.41$) than for Likert scales ($d = 0.75$). However, these effects are higher for context-desirable traits in both FC ($d = 0.61$) and Likert scales ($d = 0.99$). In our study, we found smaller effect sizes for individual context-desirable traits (ranging from 0.26 to 0.47) than reported by Speer et al. (2023) for FC, but essentially the same effect size (.58) for the composite of all context-desirable scales. We note that overall profile inflation, which was positive but small in this study (median $d_Z = 0.26$), is only possible to observe in properly scaled ipsative measures that yield normative scores (usually with the aid of an appropriate IRT model). This result is reassuring for considering quantitative ipsative formats like the 'fixed-point-total' format for high-stakes assessments. Its resistance to faking appears the same or better than FC measures with their seemingly limited scope for inflating one's scores. Our findings do not

corroborate surprisingly high faking effect sizes (between 1.16 and 2.15) that were reported by Schulte et al. (2024) for a graded preference ipsative measure with a 9-point scale, despite offering similar scope for assigning contrasting numbers of points to statements within blocks.

Most likely, the answer to faking resistance lies in the questionnaire design. While the LSQ described here was created for commercial use with faking resistance in mind, assessing 24 narrow traits with MFC blocks of matched desirability and only very few blocks with negatively keyed items, the graded preference questionnaire described by Speer et al. was created for research, measured 4 broad traits containing nearly half blocks with negatively keyed items and even some unidimensional blocks. All these features can negatively impact the test fakeability as Frick (2022) demonstrates.

The Role of General Mental Ability in Faking Behavior

Our findings of the role of GMA in ipsative scores inflation corroborate those presented by Christiansen and colleagues (2005). Participants with higher GMA were able to inflate their personality scores more, presumably through a greater capacity for processing complex information (Del Missier et al., 2011; Skagerlund et al., 2021), which is particularly relevant when navigating quantitative comparisons, where test-takers must consider not just one item at a time but also how each preference impacts the outcomes for other items (Kuncel et al., 2011; McFarland & Ryan, 2006), and doing so in high stakes where the perceived risk of wrong response is heightened (Fuechtenhans & Brown, 2023).

Using response latencies as an important marker of response behavior, we found a noteworthy trend: participants with higher mental ability responded significantly faster across both stakes and made fewer clicks. The main effect of GMA on clicks reduction corroborates the work of Cokely and Kelley (2009), who found that mental ability is associated with enhanced decision-making, especially in risky environments. Interpreting our findings concerning LSQ scores, response latencies, and click count in concert, we can conclude that participants with greater mental ability are better able to navigate the test, even when the cognitive effort is increased through situational demands (high stakes) or a more complex test format (e.g., fixed-point-total format).

The Role of ATIC in Faking Behavior

One of the main aims of this study was to better understand the role of ATIC in applicant faking. Our findings support the assumption that a better understanding of the criteria being assessed can help fulfill these criteria via manipulating test responses, as higher ATIC was associated with higher LSQ job-essential scores in high stakes and score increases from low to high stakes. Thus, our results are in line with studies that investigated ATIC in assessment centers (Kleinmann, 1993; Preckel & Schüpbach, 2005) or selection interviews (e.g., König et al., 2007).

The influence of ATIC on the faking process manifested in personality scores was also reflected to an extent in response latencies and click counts. Test takers who were more aware of important traits in selection slowed down more in high stakes and increased the number of clicks between the two conditions more. Cross-classified models provided some evidence about mechanisms of these increases. Blocks that measured important traits were associated with longer delays and more clicking, suggesting that purposeful editing likely took place. While mere hesitation, perceived difficulty, extra checking or uncertainty could not be ruled out as potential reasons for the increases in response times and clicks, the evidence is more consistent with strategic editing. Separating block-level effects from person-level effects such as slower devices or internet connections in an online panel strengthens the evidence of strategic responding when faking.

The Interaction of GMA and ATIC in Their Influence on Faking Behavior

Although GMA and ATIC were positively correlated in our study ($r = 0.24$), the direction of their influences on the process data in high stakes was different. While higher GMA meant **faster** responses and fewer clicks in either condition, higher ATIC led to **slower** responses and larger click increase in high stakes specifically. Importantly, ATIC acted as a partial or even full mediator (in the case of clicks) for the effect of GMA. Those with higher GMA also tend to have higher ATIC, and while the general ability (GMA) makes them respond faster, the increased recognition of selection criteria (ATIC) slows them down. Therefore, ATIC partly counteracts the positive effect of GMA on response speed. Similarly, higher GMA leads to fewer edits in responses, but higher ATIC completely counteracts this effect in high stakes.

A surprising finding that emerged from path analysis of response latencies was that ATIC does not only affect response time in high stakes but also in low stakes. This seems strange at first as ATIC should apply only to contexts where there are selection criteria to fulfill. Nevertheless, we think that even though ATIC was measured as the concordance of participants' opinions of important traits for the target job with expert opinion and only made sense under the faking condition, in this case it acted as a proxy for the ability to see *what is being measured* by the questionnaire. Those who are aware of what is being measured as they are

completing a questionnaire under honest instruction take longer to respond. They, however, cannot strategically inflate important LSQ traits because they do not know about them under honest instruction, and do not edit their responses, explaining why ATIC does not have the same direct influence on the LSQ scale scores or click counts in the honest condition in the respective path models.

Overall, these findings present evidence that GMA and ATIC do, in fact, take on distinct roles in the response process under high stakes. We suggest that GMA is a factor for proficiency in navigating the response process regardless of the stakes involved and represents the '*How-To*' mechanism which emphasizes the procedural aspects of achieving a goal or performing an action (i.e., faking). On the other hand, ATIC represents the '*What-To*' mechanism which involves determining what needs to be manipulated based on the recognition and evaluation of relevant criteria. Yet just because a person knows how to fake and what to fake does not necessarily mean they do fake but simply that they would possess the abilities required for successful faking if they decided so.

The Order of Stakes in Within-Subject Faking Designs

In this study, we counterbalanced the order in which honest and faking conditions were presented to participants, to reveal very interesting results. Not only the order of conditions mattered for all the outcome variables – LSQ scores, latencies and clicks, but it also mattered for ATIC as measured by testing participants' understanding of selection criteria after providing them with a job description under faking condition. If this faking condition was the first that the participants completed, they had higher ATIC scores on average, which suggests that they read the job description more attentively and overall took the task more seriously than those for whom it was second completion. This idea is supported by the finding that those who completed faking condition first, had higher LSQ scores. They seemed to have carried this score elevation to the honest condition though, resulting in the same personality score shift across conditions as those who completed the questionnaire honestly first. Thus, administering faking condition first seems to yield better engagement, and despite causing some 'priming' of socially desirable responding, does not alter the size of score shift between stakes.

The order of stakes also moderates the effects of stakes on response latencies and clicks. If the first completion was under honest instruction, the subsequent faking completion did not result in longer response times or more clicks, which again suggests that participants rush through the questionnaire second time around rather than trying hard to 'fake' it. Moreover, more clicks were made in first completion of LSQ, whichever stakes that happened to be. More clicks are needed to strategically 'edit' retrieved responses (Fuechtenhans & Brown, 2023), so administering faking condition first could lead to ascribing the increased clicks increase solely to editing, while administering honest condition first could ascribe the increase solely to indecision/hesitation due to lack of practice, while both reasons are probably at play. The practice effect is evidenced by decreasing latency and clicks as participants go through blocks; while the strategic editing effect is evidenced by increasing latency and clicks in job-relevant blocks (both are block-level effects from cross-classified models). Based on these findings, we tentatively recommend that in studies without counterbalancing, the faking condition should be administered first. Not only is it more ecologically valid, since applicants do not usually get a preview of assessments when applying for jobs, but it also seems to yield more engaged responding consistent with the intended high-stakes manipulation.

Limitations and Future Directions

The first limitation lies in the laboratory nature of our study, in which we did not test real job applicants. Although we successfully raised the stakes, we cannot automatically generalize the findings to a real application process. It is possible that the effects found in our study are stronger than would be in a real application context (Holden, 2008). Yet our experimental manipulation of participants' motivation to fake in a within-participants design combined experimental control with psychological realism. Specifically, a more subtle strategy to induce faking was utilized, associated with higher ecological validity than direct instructed 'faking good' studies.

Moreover, Guenole et al. (2022) raised the concern of utilizing 'faking to job' instructions with a heterogeneous sample, suggesting that some people might be better positioned to fake the predetermined job role because of prior knowledge or experience. Indeed, we did not ask participants whether they had any prior knowledge regarding this job position. Thus, some Prolific participants might have a higher ATIC score because they knew more about the advertised position, helping them to identify essential traits with more ease. Then, the question becomes whether our ATIC measure captured a person's ability to identify any selection criteria or whether the measure captures additional factors such as knowledge/familiarity of the target job. We recommend that in future studies, researchers control participants' prior knowledge, education, and experience in relevant fields to better understand the nature of ATIC.

While utilizing all available records of latencies, clicks and personality scale scores per participant in cross-classified models, the study may still be underpowered at the between-scale, between-block, and between-person levels,

with effective sample sizes of only 24, 88 and 200, respectively. This is particularly salient when it comes to detecting effects specific to order of stakes, with only 100 participants allocated to each order. Therefore, we consider the results as needing replication with larger samples.

Finally, we suggest that future research should capture applicant reactions, particularly to new types of assessments such as the fixed-point-total ipsative questionnaire used here. Failure to do so in this study precluded us from weighing the reduced effect sizes and fakeability provided by the LSQ against participants' feelings toward this assessment.

Conclusions and Practical Considerations

In this within-participants experimental research study we have established that personality test scores are affected by participants' GMA and ATIC. Although GMA was not significantly associated with personality scores in either high or low stakes, it was positively associated with the inflation of job-relevant scores from low to high stakes. This suggests that the effective reduction of fakeability achieved using ipsative formats has a tradeoff – increased cognitive load involved in successfully navigating such assessments. Yet, higher ATIC was significantly associated with higher job-relevant scores in high stakes as well as the score inflation from low to high stakes. Considering the combined effects for GMA and ATIC on personality test scores derived from the cognitively complex quantitative preference test format, our path analysis models suggest that it is GMA that primarily drives score inflation, as the indirect effect via ATIC is relatively small in this study.

Moreover, our findings call for a modification to the argument that 'faking takes time'. While it is true that an average participant spent longer completing the ipsative questionnaire in faking condition, the participants with lower GMA also spent significantly longer completing the questionnaire in honest condition. Therefore, longer response latencies cannot serve as a reliable indicator of faking without considering the participant's perceptual speed (and therefore GMA) in general. Holden and Lambert's (2015) research supports this argument, revealing that a substantial percentage of honest respondents can be misclassified as fakers based on response latencies.

Considering the development of ipsative test formats to counteract faking, a delicate balance must be maintained to avoid excessive cognitive demands, which would favor participants with greater mental ability. This aligns with Pauls and Crost (2005) who observed that personality questionnaires under high situational pressures can resemble achievement tests. Essentially, the challenge lies in preventing the assessment of personality (typical performance) turn into the assessment of abilities (maximum performance). Not only could this disadvantage minority groups thus increasing adverse impact, but also potentially worsen applicant reactions to these cognitively demanding assessments feeling more like "tests" rather than "questionnaires". Ways of keeping balance might be through lowering the stakes, for instance by never using personality assessments as stand-alone screens but rather as interview aids; or by making candidates feel safe to express their true personality, for instance by presenting them with equally desirable alternatives in ipsative blocks (Fuechtenhans & Brown, 2023). When personality assessments are used as part of a broader selection system, this should always be supported by local validation, applicant reaction and fairness monitoring.

Appendix A. Job Description of the 'Director of Market Research and Analytics'

Job Title: Director of Market Research and Analytics

Company Overview: We are a leading market research firm specializing in data-driven insights and strategic decision-making for businesses across various industries. We are seeking a highly skilled and visionary leader to join our team as the Director of Market Research and Analytics.

Job Summary: As the Director of Market Research and Analytics, you will provide strategic direction and leadership in transforming raw data into actionable insights. This role demands a unique blend of technical expertise, analytical skills, and the ability to lead and inspire a team of analysts.

Key Responsibilities:

- Provide vision and leadership in developing and executing the company's market research and analytics strategy.
- Lead and mentor a team of data analysts, guiding their professional growth and ensuring the successful execution of projects.
- Oversee the analysis of complex datasets, identification of market trends, and the generation of strategic recommendations.
- Conduct high-level market research to track industry trends, competitor performance, and consumer behavior.
- Collaborate with clients to understand their specific research needs and provide customized solutions, establishing strong client relationships.
- Oversee the development of predictive models and algorithms to forecast market dynamics and customer preferences.
- Develop strategies for effective report creation and presentation of findings to clients.

The Candidate:

- Master's degree in a relevant field (e.g., Business Analytics, Data Science, Economics).
- Minimum of 8 years of experience in data analysis and market research, with at least 3 years in a leadership role.
- Proficiency in data analysis software (e.g., Python, R, SQL) and data visualization tools (e.g., Tableau).
- Proven leadership and team management skills.
- Exceptional analytical and problem-solving abilities.
- Strong communication and presentation skills.
- Visionary thinking and the ability to drive innovation in market research and analytics.

Appendix B. Screening Questions

Please select TRUE or FALSE for each statement.

1. The personality assessment is the first step of the selection process, followed by the completion of several other tasks.
2. I have applied for a job as a Director of Market Research and Analytics, but I don't really want the job.
3. It does not matter what my personality profile looks like because nobody will look at it.
4. If my profile deems me successful in getting hired, I am in with the chance of winning £50.
5. The personality test is the final step of the selection process, and my profile will be consequential to the hiring decision.
6. I have applied for a job as a Director of Market Research and Analytics, and I really want the job.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10869-026-10135-x.

Funding This work was supported by Innovate UK Knowledge Transfer Partnership [grant number KTP012467], and Young Samuel Chambers (YSC) Limited.

Data Availability Data during this study was collected with accordance to the standard Privacy Policy of Young Samuel Chambers(YSC), which does not permit sharing individual records, only summary statistics. Analysis scripts and outputs including the cross-classified path models are available in the journal's repository https://osf.io/collections/jbp/discover. Test materials for Leadership Styles Questionnaire and Leadership Ability Quotient are proprietary and cannot be shared.

Declarations

Ethics Approval This study was approved by the School of Psychology Ethics Committee at the University of Kent, approval number 202116323169187293.

Conflict of Interest The authors declare that they have no conflict of interest.

References

Arthur, W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*(1), 1–16. https://doi.org/10.1111/j.1468-2389.2010.00476.x

Borman, T. C., Dunlop, P. D., Gagné, M., & Neale, M. (2023). Improving reactions to forced-choice personality measures in simulated job application contexts through the satisfaction of psychological needs. *Journal of Business and Psychology, 39*(1), 1–18. https://doi.org/10.1007/s10869-023-09876-w

Brown, A. (2015). Personality assessment, forced-choice. In *International Encyclopedia of the Social & Behavioral Sciences* (2nd edition). Elsevier. https://doi.org/10.1016/b978-0-08-097086-8.25084-8

Brown, A. (2016). Thurstonian scaling of compositional questionnaire data. *Multivariate Behavioral Research, 51*(2–3), 345–356. https://doi.org/10.1080/00273171.2016.1150152

Brown, A. (2022). *Fixed-total ipsative data - Continuous or ordinal? 87th International Meeting of the Psychometric Society, July 11–15, 2022*. Italy.

Brown, A., & Böckenholt, U. (2022). Intermittent faking of personality profiles in high-stakes assessments: A grade of membership analysis. *Psychological Methods, 27*(5), 895–916. https://doi.org/10.1037/met0000295

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 516–529. https://doi.org/10.1080/10705511.2017.1392247

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347. https://doi.org/10.1037/apl0000414

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment. *Human Performance, 18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803_4

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making, 4*(1), 20–33. https://doi.org/10.1017/s1930297500000067x

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*(2), 155–169. https://doi.org/10.1111/j.1468-2389.2008.00420.x

Dalal, D. K., Zhu, X. (Susan), Rangel, B., Boyce, A. S., & Lobene, E. (2019). Improving applicant reactions to forced-choice personality measurement: Interventions to reduce threats to test takers' self-concepts. *Journal of Business and Psychology*, *36*(1), 55–70. https://doi.org/10.1007/s10869-019-09655-6

de Klerk-van Someren, G., & Rockson, E. (2023). *Leadership ability quotient: Technical manual (First edition).* YSC Consulting - part of Accenture.

de Klerk-van Someren, G., Brown, A., Chandrasekaran, L., & Rockson, E. (2023). *Leadership styles questionnaire: Technical manual (Second edition)*. YSC Consulting - part of Accenture.

Del Missier, F., Mäntylä, T., & de Bruin, W. B. (2011). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making, 25*(4), 331–351. https://doi.org/10.1002/bdm.731

Dilchert, S., & Ones, D. S. (2011). Application of preventive strategies. In *New Perspectives on Faking in Personality Assessment* (pp. 177–200). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0054

Donovan, J. J., Dwight, S. A., & Schneider, D. (2013). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *Journal of Business and Psychology, 29*(3), 479–493. https://doi.org/10.1007/s10869-013-9318-5

Feldman, M. J., & Corah, N. L. (1960). Social desirability and the forced choice method. *Journal of Consulting Psychology, 24*(6), 480–482. https://doi.org/10.1037/h0042687

Feng, T., & Cai, L. (2024). Sensemaking of process data from evaluation studies of educational games: An application of cross-classified item response theory modeling. *Journal of Educational Measurement*, 1–37. https://doi.org/10.1111/jedm.12396

Frick, S. (2022). Modeling faking in the multidimensional forced-choice format: The faking mixture model. *Psychometrika, 87*(2), 773–794. https://doi.org/10.1007/s11336-021-09818-6

Fuechtenhans, M., & Brown, A. (2023). How do applicants fake? A response process model of faking on multidimensional forces-choice personality assessments. *International Journal of Selection and Assessment, 31*(1), 105–119. https://doi.org/10.1111/ijsa.12409

Goldhammer, F., Kroehne, U., Hahnel, C., Naumann, J., & De Boeck, P. (2024). Does timed testing affect the interpretation of efficiency scores?—A GLMM analysis of reading components. *Journal of Educational Measurement, 61*(3), 349–377. https://doi.org/10.1111/jedm.12393

Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *Journal of Applied Psychology, 35*(6), 407–412. https://doi.org/10.1037/h0058853

Griffith, R. L., & Converse, P. D. (2011). The rules of evidence and the prevalence of applicant faking. In *New Perspectives on Faking in Personality Assessment* (pp. 34–52). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0018

Griffith, R. L., & Robie, C. (2013). Personality testing and the "F Word": Revisiting seven questions about faking. In N. Christiansen & R. Tett (Eds.), *Handbook of Personality at Work* (1st Edition, pp. 253–280). Routlege. https://doi.org/10.4324/9780203526910

Guenole, N., Brown, A., & Lim, V. (2022). Can faking be measured with dedicated validity scales? Within-subject trifactor mixture modeling applied to BIDR responses. *Assessment, 30*(5), 1523–1542. https://doi.org/10.1177/10731911221098434

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*(1), 9–24. https://doi.org/10.1037/0021-9010.91.1.9

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin [PsycARTICLES]*, *74*(3), 167–184.http://dx.doi.org.helicon.vuw.ac.nz/10.1037/h0029780

Holden, R. R. (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Personality and Individual Differences, 44*(1), 311–321. https://doi.org/10.1016/j.paid.2007.08.012

Holden, R. R., & Lambert, C. E. (2015). Response latencies are alive and well for identifying fakers on a self-report personality inventory: A reconsideration of van Hooft and Born (2012). *Behavior Research Methods, 47*(4), 1436–1442. https://doi.org/10.3758/s13428-014-0524-5

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63*(2), 272–279. https://doi.org/10.1037/0022-3514.63.2.272

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30*(2), 161–172. https://doi.org/10.1177/0146167203259930

Jeong, Y. R., Christiansen, N. D., Robie, C., Kung, M. C., & Kinney, T. B. (2017). Comparing applicants and incumbents: Effects of response distortion on mean scores and validity of personality measures. *International Journal of Selection and Assessment, 25*(3), 311–315. https://doi.org/10.1111/ijsa.12182

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kiefer, C., & Benit, N. (2016). What is applicant faking behavior? A review on the current state of theory and modeling techniques. *Journal of European Psychology Students, 7*(1), 9–19. https://doi.org/10.5334/jeps.345

Klehe, U.-C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance, 25*(4), 273–302. https://doi.org/10.1080/08959285.2012.703733

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology, 78*(6), 988–993. https://doi.org/10.1037/0021-9010.78.6.988

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work. *Organizational Psychology Review, 1*(2), 128–146. https://doi.org/10.1177/2041386610387000

Komar, S., Komar, J. A., Robie, C., & Taggar, S. (2010). Speeding personality measures to reduce faking a self-regulatory model. *Journal of Personnel Psychology, 9*(3), 126–137. https://doi.org/10.1027/1866-5888/a000016

König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U. (2007). Candidates' ability to identify criteria in nontransparent selection procedures: Evidence from an assessment center and a structured interview. *International Journal of Selection and Assessment, 15*(3), 283–292. https://doi.org/10.1111/j.1468-2389.2007.00388.x

Kovačić, M. P. (2021). Usefulness of item response latencies in separating self-enhancement from impression management. *Current Psychology*. https://doi.org/10.1007/s12144-021-02375-2

Kuncel, N. R., Goldberg, L. R., & Kiger, T. (2011). A plea for process in personality prevarication. *Human Performance, 24*(4), 373–378. https://doi.org/10.1080/08959285.2011.597476

Martínez, A., & Salgado, J. F. (2021). A meta-analysis of the faking resistance of forced-choice personality inventories. *Frontiers in Psychology, 12*, Article 732241. https://doi.org/10.3389/fpsyg.2021.732241

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Strctural Equation Modeling: Present and Future* (pp. 7–46). SSI Scientific Software International.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605. https://doi.org/10.1146/annurev.psych.60.110707.163612

McDonald, R. P. (1999). *Test theory: A unified treatment* (2011th ed.). Erlbaum.

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology, 36*(4), 979–1016.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*(2), 348–355. https://doi.org/10.1037/0021-9010.88.2.348

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide. Eighth Edition.* Muthén & Muthén.

MyPrint questionnaire. (2018). https://help.talentoday.com/assessment/opmi/. Accessed 24 June 2026

Oostrom, J. K., Melchers, K. G., Ingold, P. V., & Kleinmann, M. (2016). Why do situational interviews predict performance? Is it saying how you would behave or knowing how you should behave? *Journal of Business and Psychology, 31*, 279–291. https://doi.org/10.1007/s10869-015-9410-0

Pauls, C. A., & Crost, N. W. (2005). Effects of different instructional sets on the construct validity of the NEO-PI-R. *Personality and Individual Differences, 39*(2), 297–308. https://doi.org/10.1016/j.paid.2005.01.003

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). The effects of applicant faking on forced-choice and Likert scores. *Organisational Research Methods*, *22*(3), 710–739. https://doi.org/10.1177/1094428117753683

Peterson, M. H., Griffith, R. L., Isaacson, J. A., O'Connell, M. S., & Mangos, P. M. (2011). Applicant faking, social desirability, and the prediction of counterproductive work behaviors. *Human Performance, 24*(3), 270–290. https://doi.org/10.1080/08959285.2011.580808

Preckel, D., & Schüpbach, H. (2005). Zusammenhänge zwischen rezeptiver Selbstdarstellungskompetenz und Leistung im Assessment Center. *Zeitschrift Für Personalpsychologie, 4*(4), 151–158. https://doi.org/10.1026/1617-6391.4.4.151

Scharfen, J. (2018). *Response time reduction due to retesting in mental speed tests : A meta-analysis. 2017*. https://doi.org/10.3390/jintelligence6010006

Schulte, N., Kaup, L., Bürkner, P. C., & Holling, H. (2024). The fakeability of personality measurement with graded paired comparisons. *Journal of Business and Psychology*. https://doi.org/10.1007/s10869-024-09931-0

Skagerlund, K., Forsblad, M., Tinghög, G., & Västfjäll, D. (2021). Decision-making competence and cognitive abilities: Which abilities matter? *Journal of Behavioral Decision Making*, *35*(1). https://doi.org/10.1002/bdm.2242

Speer, A. B., Wegmeyer, L. J., Tenbrink, A. P., Delacruz, A. Y., Christiansen, N. D., & Salim, R. M. (2023). Comparing forced-choice and single-stimulus personality scores on a level playing field: A meta-analysis of psychometric properties and susceptibility to faking. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001099

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*(3), 184–203. https://doi.org/10.1177/0146621604273988

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance, 19*(3), 175–199. https://doi.org/10.1207/s15327043hup1903_1

Walczyk, J. J., Mahoney, K. T., Doverspike, D., & Griffith-Ross, D. A. (2009). Cognitive lie detection: Response time and consistency of answers as cues to deception. *Journal of Business and Psychology, 24*(1), 33–49. https://doi.org/10.1007/s10869-009-9090-8

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment, 33*(2), 156–170. https://doi.org/10.1037/pas0000971

Zhang, B., Luo, J., & Li, J. (2023). Moving beyond Likert and traditional forced-choice scales: A comprehensive investigation of the graded forced-choice format. *Multivariate Behavioral Research, 59*(3), 434–460. https://doi.org/10.1080/00273171.2023.2235682

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2019). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods, 23*(3), 569–590. https://doi.org/10.1177/1094428119836486

Ziegler, M. (2011). Applicant faking: A look into the black box. *The Industrial-Organizational Psychologist*, *49*(1), 29–36. http://www.siop.org/tip/july11/06ziegler.aspx