Assessment
2021, Vol. 28(3) 1004–1017
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1073191120914970
journals.sagepub.com/home/asm
SAGE

Caught in the Act: Predicting Cheating in Unproctored Knowledge Assessment

Diana Steger[1], Ulrich Schroeders[2], and Oliver Wilhelm[1]

Abstract

Cheating is a serious threat in unproctored ability assessment, irrespective of countermeasures taken, anticipated consequences (high vs. low stakes), and test modality (paper-pencil vs. computer-based). In the present study, we examined the power of (a) self-report-based indicators (i.e., Honesty-Humility and Overclaiming scales), (b) test data (i.e., performance with extremely difficult items), and (c) para data (i.e., reaction times, switching between browser tabs) to predict participants' cheating behavior. To this end, 315 participants worked on a knowledge test in an unproctored online assessment and subsequently in a proctored lab assessment. We used multiple regression analysis and an extended latent change score model to assess the potential of the different indicators to predict cheating. In summary, test data and para data performed best, while traditional self-report-based indicators were not predictive. We discuss the findings with respect to unproctored testing in general and provide practical advice on cheating detection in online ability assessments.

Keywords

test taking, cheating, honesty, para data, declarative knowledge

Some high school students cheat to get better grades (Cicek, 1999), some applicants fake to get a job (Tippins et al., 2006), and some convicts pretend to suffer from a severe mental disorder in order to escape death penalty (Slobogin, 2005). In psychological assessment, cheating is considered a serious threat to ability testing, and proctored test sessions are regarded as the most effective remedy (Rovai, 2000). With an increasing number of tests administered in unproctored settings—such as Internet-based (Schroeders et al., 2010; Sliwinski et al., 2018) or smartphone-based assessments (e.g., Harari et al., 2016; Steger, Schroeders, & Wilhelm, 2019)—this recommendation has been abandoned in favor of greater dissemination of the tests and accessibility of participants. Consequently, the proneness to cheating is an important characteristic of psychological ability tests administered with digital devices. Conversely, in the assessment of typical behavior, successful faking mostly hinges on participants' faking ability (Geiger et al., 2018) rather than test mode (Gnambs & Kaspar, 2017).

The reasons for cheating on ability tests are manifold and, if it remains undetected, lead to biased test scores (Bressan et al., 2018). Thus, researchers and practitioners proposed different ideas to prevent test takers from cheating—for example, specific instructions (Wilhelm & McKnight, 2002), honor codes and honesty contracts (O'Neill & Pfeiffer, 2012), or the announcement of proctored follow-up tests (Lievens & Burke, 2011). Unfortunately, countermeasures against cheating have only limited success. In a recent meta-analysis, such countermeasures were not suitable to prevent test score differences between proctored and unproctored ability tests (Steger, Schroeders, & Gnambs, 2020). In more detail, results showed that if participants had the opportunity to cheat (e.g., by looking up the correct answer on the Internet), they cheated, irrespective of context (high- vs. low-stakes testing) or whether countermeasures are taken. Because cheating is hard to avoid in the first place, one possibility to secure data quality is to flag irregular responses after testing to evaluate the severity of bias and to allow data cleaning. In the following, we first discuss traditional approaches that rely on self-report data to detect dishonest responding, followed by test data approaches that analyze response patterns. In addition to these classic data formats (Johnson, 2001), we also present more recent approaches that use so-called para data to capitalize on the potential of computer-based assessment.

Self-Report Data or "Lie to Me"

Methods to detect faking in questionnaires have a long tradition: Validity scales were first introduced in the Minnesota

[1]Ulm University, Ulm, Germany
[2]University of Kassel, Kassel, Germany

Corresponding Author:
Diana Steger, Department of Individual Differences and Psychological Assessment, Ulm University, Albert-Einstein-Allee 47, Ulm 89081, Germany.
Email: diana.steger@uni-ulm.de

Multiphasic Personality Inventory (Hathaway & McKinley, 1943), followed by other instruments such as the Sixteen Personality Factor Questionnaire in 1949 (Cattell et al., 1970). For example, in the Minnesota Multiphasic Personality Inventory II (Butcher et al., 2001), up to 12 validity indices could be computed, including scales for lying, social desirability, and infrequent events or rare behaviors. These scales were designed to assess answer tendencies that would lead to false interpretations of results using items about the frequency of either culturally approved behaviors that are unlikely to always occur (e.g., "I always clean up after I make a mess.") or culturally undesirable behaviors that are likely to occur (e.g., "I never pick my nose.").

Contrary to the traditional lie scales and faking indices, which are best applied to detect faking on self-report scales, the Honesty-Humility scale of the HEXACO model (Ashton & Lee, 2007, 2008) represents a measure of a personality trait that has been linked successfully with cheating and other dishonest behavior (Heck et al., 2018). In general, self-reports of Honesty-Humility seem to be valid under low-stakes condition (Ashton et al., 2014; Zettler et al., 2016), although faking might play a role in high-stakes conditions (MacCann, 2013).

Finally, dishonest responding has been linked with overclaiming, which reflects the tendency to claim knowledge about nonexistent items (Paulhus et al., 2003; Phillips & Clancy, 1972). Overclaiming can be assessed by juxtaposing familiarity ratings for a list of items consisting of existing terms (*reals*) and nonexisting terms (*foils*). If overclaiming is understood as participants' response bias, the score may be appropriate to detect response distortion (Paulhus et al., 2003) and to improve the validity of psychological assessment (Bing et al., 2011; but see also Müller & Moshagen, 2019). In practical terms, one might expect people who consciously lie about their knowledge to also boost their test scores by engaging in cheating behaviors, just as one would expect this behavior from people with high self-interest scores (a facet of the dark personality, see also Moshagen et al., 2018). In contrast to social desirability, overclaiming does not seem to be confounded with personality or intelligence measures as such (Bensch et al., 2019), which might allow for a more direct measure of self-enhancement. Taken together, questionnaire-based methods do not depend on the assessment modality: They can be included in both paper-pencil and computer-based tests or self-reports. However, questionnaires can be easily manipulated if a test taker is motivated and capable (Geiger et al., 2018). Especially when they are included in test batteries of cognitive abilities, participants might figure out the purpose of these scales.

Test Data or "The Man Who Knew Too Much"

Whereas cheating detection methods that rely on self-report data demand the implementation of additional instruments, participants' test data itself can be used to detect cheating. In the simplest case, individual test scores can be compared with previous performance to detect unexpected scores and classify participants as potential cheaters (McClintock, 2016). Statistical methods such as the *Z* test or the likelihood ratio tests have been proposed to flag participants with aberrantly high test scores across two testing conditions (Guo & Drasgow, 2010). In personnel selection, proctored follow-up tests are often used to identify suspected cheaters in unproctored screenings (Lievens & Burke, 2011; Nye et al., 2008). Another approach—specifically designed to catch cheaters red-handed—is presenting participants tasks that are virtually unsolvable as in the *word jumble task* (Hoffmann et al., 2015; Wiltermuth, 2011) in which participants are asked to solve anagrams. Some of the anagrams are almost impossible to solve, which identifies participants as cheaters if they report having solved the items.

Furthermore, person-fit statistics can be applied to detect unusual or atypical patterns in a person's responses by taking into account the complete response vector rather than single test scores or responses to single items (Meijer, 1996). More specifically, person-fit statistics can be used to identify participants with spuriously low or high test scores by comparing participants' actual with the expected responses (Karabatsos, 2003). Besides the detection of deliberate cheating (e.g., answer copying; Sotaridona & Meijer, 2002), person-fit indices can help identify careless or random responding, creative responding, and lucky guessing (Meijer, 1996; Niessen et al., 2016). However, in a comprehensive simulation study, Karabatsos (2003) evaluated the performance of 36 different person-fit indices and found that cheating—as compared with other odd response styles such as careless or random responding—was hardest to detect. Unfortunately, most indicators performed only slightly better than chance when trying to detect cheaters. Also, the performance of person-fit statistics varied widely; that is, performance improved with both increasing test length and with decreasing number of cheaters in the sample. Taken together, available methods that focus on the analysis of test data are easy to incorporate and cost-efficient, as they require neither additional testing time nor special technical equipment. In comparison with lying scales and other questionnaire-based methods, they are less obtrusive and in all likelihood more difficult to fake.

Para Data or "Catch Me if You Can"

Technology-based assessment is a generic term for computer- and smartphone-based assessment. It allows the recording of auxiliary data such as reaction times and GPS localization data. Such an enriched assessment has stirred expectations of researchers to measure important aspects of psychological constructs that could not be measured with traditional paper-pencil tests. This expectation, however, has often led to disappointment (e.g., Schroeders et al., 2013). In contrast to previous efforts of supplementing the assessment of psychological constructs—for example, the assessment of intelligence by considering reaction times (Goldhammer & Klein Entink, 2011)—we argue that para data (Couper, 2005) are best used to gain insight into participants' test-taking behavior. Para data include log data (Kroehne & Goldhammer, 2018), response latencies (Holden & Lambert, 2015), or keystrokes and mouse clicks (Kieslich & Henninger, 2017; Olson & Parkhurst, 2013). One major benefit is that collecting incidental para data is supposedly unobtrusive, because it is a mere bycatch of computer-based testing (Couper, 2005). In this sense, response time analyses were used to identify participants who were instructed to fake good or fake bad on a personality test (Holden & Lambert, 2015), resulting in a classification rate of only 60% correctly identified participants. Given the serious consequences of misclassifications especially in many applied contexts, certainly additional indicators (e.g., Buchanan & Scofield, 2018) are needed to improve classification rates.

Another method that relies on para data was introduced by Diedenhofen and Musch (2017). They developed a JavaScript called *PageFocus* that records instances when subjects switch between browser tabs or open a new browser tab: The script records events that are indicative of not focusing on the task at hand. In their study, participants worked on an online knowledge task and a reasoning task. Defocusing events and scores in the knowledge test were positively correlated (r = .37), while there was no significant correlation with an additional figural reasoning task (r = .07). Therefore, the defocusing events could serve as an indicator of cheating, but they cannot be equated with cheating. In summary, the use of para data seems promising for investigating data quality because recording para data is unobtrusive and time- and cost-efficient. However, ethical concerns about recording supposedly unethical behavior remain present. Furthermore, the extent to which notifications about the collection of para data might influence the actual test-taking behavior remains unclear.

The Present Study

In recent decades, technological advances and societal changes have influenced the way we do research and collect data in psychological research (e.g., Yarkoni, 2012). In psychological assessment, web- and smartphone-based measures have been implemented, and at the same time, concerns about the quality of the online collected data have been raised (Krantz & Reips, 2017). Because online knowledge tasks are affected by dishonest participant behavior to a significant degree (Steger et al., 2020), we compare different methods of detecting cheating behavior in an unproctored knowledge assessment. We asked participants to fill out two parallel forms of a knowledge test—once in an online session and once in a lab session. We expect participants who cheated in the unproctored condition to have higher scores than in the proctored condition—in which cheating was not possible. To this end, we employed methods that are based on self-report data (S-data), test data (T-data), and para data (henceforth abbreviated to P-data) to predict cheating behavior, which, in the end, can be used to evaluate data quality of unproctored assessments.

As S-data indicators, we used two scales measuring the HEXACO factor Honesty-Humility and Overclaiming. As T-data indicator, we analyzed participants' performance when answering practically unsolvable knowledge items following the logic of the *word jumble task* (Wiltermuth, 2011), but using a task specifically designed to match the test context of a knowledge assessment. Last, as P-data we used unusual response times and the number of defocusing events to predict cheating in unproctored assessments. Because high levels of honesty are associated with lower levels of various deviant behaviors (e.g., Hilbig et al., 2015; Hilbig & Zettler, 2015; Lee et al., 2013), we expect honesty to be negatively associated with cheating behavior, as participants with lower honesty score might cheat more. Moreover, in line with previous findings (e.g., Fell et al., 2019), we expect participants who tend to overclaim knowledge also to cheat more—resulting in a positive association between cheating behavior and overclaiming. The difficult items we used in this study can be viewed as a direct observation of cheating behavior: Since the test takers did not know that some of the items were almost unsolvable, it was hard to lever out this index. Participants with higher scores on the difficult items are more likely to have cheated during the knowledge test. Similarly, as looking up answers on the Internet takes time (Bloemers et al., 2016) and requires browser tab switches that can be recorded as defocusing events (Diedenhofen & Musch, 2017), we expect cheating behavior to be associated with a larger number of unusually high response times and a larger number of defocusing events.

Method

Design and Participants

The present experiment was part of a large, multicentered study on creative abilities that was conducted at two German universities (i.e., University of Bamberg and Ulm University). In total, 315 participants took part in the comprehensive assessment. Participants were recruited via university mailing lists, posts in local Facebook groups, newspapers, and posters on public notice boards. All participants provided written informed consent. Participants had a mean age of 25.5 years (*SD* = 7.8 years); 226 participants (71.7%) were female.

Data collection took place in two separate sessions: After participants signed up for the study, they received an email with a link to the unproctored online assessment (unproctored condition). During the unproctored online session, participants had to fill out an online knowledge test and a personality questionnaire (description shown below). No time limit was imposed during the online assessment, and the mean testing time was about 1 hour. To increase the propensity of cheating, participants were told that all participants who answer 80% or more of the questions correctly participate in a lottery with the chance to win an Amazon gift card for 25€ just before starting the online knowledge test (see the online supplemental material for the exact wording of the instructions). In the second part of the study (the lab session), participants worked on various cognitive abilities tasks, including a second knowledge assessment and an overclaiming questionnaire (proctored condition). Test time was 5 hours in total. After having completed the online and lab sessions, participants received 70€ as monetary reimbursement. Moreover, participants were also debriefed with regard to the cheating instruction, and the gift card was distributed among all participants at random. The time period between online and lab session varied between 1 day and 3 weeks. To avoid bias due to practice effects, distinct item sets were used for online and lab assessment.

Measures

Declarative Knowledge. We used a computer-based knowledge test, because the solutions to such tasks are especially easy to look up on the Internet (Bloemers et al., 2016; Steger et al., 2020). We used two parallel test forms with 102 items each. Both test forms covered questions from 34 knowledge domains, ranging from the natural, life, and social sciences, humanities, and pop culture domains (see also Table S1 in the online supplement). Questions were sampled from a larger item pool of multiple-choice items (Steger et al., 2019) for two parallel test forms, with both item sets equally covering the broad content domains with comparable mean and range of item difficulties. One parallel constructed test form was administered randomly to participants in the online session; the remaining test form was administered in the lab session to avoid bias due to different item samples or item order effects. Also empirically, both parallel test forms yielded comparable results. In the proctored condition, item difficulty of Form A ranged from .18 to .88 (*M* = .56, *SD* = .14), and item difficulty of Form B ranged from .26 to .85 (*M* = .58, *SD* = .14). Moreover, internal consistency was good for both test forms (Form A: α = .82, Form B: α = .76).

Self-Report Data. First, to assess cheating-related personality traits, we used the German 60-item version of the HEXACO (Moshagen et al., 2014). In the present analysis, we focus on the Honesty-Humility facet as it is reported to be related to dishonest behavior (Ashton et al., 2014; Lee et al., 2005). As we did not expect any influences of the assessment mode on response biases for this self-report (Gnambs & Kaspar, 2017), the HEXACO-60 was administered online to reduce testing time for the lab assessment. For the Honesty-Humility scale, internal consistency was α = .70. Second, to assess overclaiming, we used a newly developed overclaiming questionnaire. Participants were asked to rate their familiarity with 149 terms on a scale ranging from 1 (*never heard of it*) to 5 (*very familiar*). Of these 149 terms, 121 were existing terms (*reals*) and 28 were nonexisting (*foils*). We selected reals to cover a broad range of item difficulty—from terms that most people are at least somewhat familiar with, to terms most people would not know. In turn, we selected foils that sounded similar to terms from the given subject, but were sufficiently different—that is, the terms had to be completely new creations rather than only replacement of one or two letters. Prior to compiling the final questionnaire, reals and foils were rated according to their difficulty and plausibility by six human raters. The domains assessed within the questionnaire matched the 34 content domains assessed in the knowledge test. As an indicator of overclaiming, we used the mean rating of foils (see also Hülür et al., 2011). As expected, mean familiarity of all *foils* was low, ranging from 1.13 to 2.70 (*M* = 1.55, *SD* = 0.43) compared with the mean familiarity ratings of all *reals*, which ranged from 1.16 to 4.46 (*M* = 2.69, *SD* = 0.80). Internal consistency was excellent (α = .90). Subsequently, overclaiming was used as a predictor for cheating behavior in the present study. To prevent participants from looking up terms presented in the overclaiming questionnaire on the Internet, this instrument was included in the proctored lab session.

Test Data. Mixed in with the general knowledge test, both in the online and in the lab condition, we presented participants 34 multiple choice items with four response options that were virtually unsolvable but easy to look up on the Internet (e.g., "When was Cunigunde of Luxembourg born?" or "How high is the north tower of St. Stephen's Cathedral in Vienna?"). To better distinguish between knowledge item types, we label these items as *difficult items*. For these questions, we expect item mean scores of around .25—corresponding to performance on chance level. In practice, these expectations matched our empirical results: In the lab condition, mean item difficulty ranged from .05 to .47 (*M* = .24, *SD* = .10) for Form A and from .07 to .42 (*M* = .22, *SD* = .09) for Form B. Accordingly, all else being equal, the higher the score of participants on these items in the online condition, the stronger the indication that they cheated during the knowledge test.

Para Data. For all knowledge items (both regular and difficult items, as well as in both online and lab condition), we additionally recorded response times and used a JavaScript—similar to the PageFocus script (Diedenhofen & Musch, 2017)—to record the occurrence of defocusing events. For the response times, we used the occurrence of conspicuously long response time as an indicator of potential cheating behavior. For every participant, we counted the number of events in which the participant's response times were three standard deviations above the median response time of the respective item. This item-focused approach takes into account the individual item length as the median is computed for each item separately. To account for individual differences in reading speed, we set the limit for flagged response times at three standard deviations, assuming that even slow readers who work on the items without cheating should achieve response times that fall within the range of unsuspicious response times. On average, participants' median response time across all items was 7.29 seconds (SD = 2.38 seconds) in the lab condition and 9.86 seconds (SD = 3.2 seconds) in the online condition. For the defocusing events, we counted the number of items in which the participant switched browser tabs prior to answering the question. Cases in which the participants switched browser tabs multiple times while answering a question were treated as one single defocusing event. Separate count variables were computed for the online and lab conditions.

Statistical Analyses

Data Cleaning and Missing Data. We screened the data for careless and negligent responding, checking for impossibly low response times and response patterns separately for lab and online data. No participants were excluded from analysis. Because the overall percentage of missings in the dataset is low (1.31%), and reasons for missing scale scores were based on (random) technical malfunctions rather than noncompliance from participants, we did not exclude any of the participants. Instead, we used pairwise complete observations for analyses on the manifest level. For analyses on the latent level, we used *full information maximum likelihood* to account for missingness that is assumed to be completely at random.

Score Computation and Content Aggregates. For analyses on the manifest level, we first computed difference scores between the overall proportion-correct scores of the online and the lab knowledge assessment. The difference score served as an indicator for suspected cheating behavior, with higher score differences between online and lab assessment indicating more cheating during the online session. For computing the scale score of the Honesty-Humility scale, we followed standard procedures (Moshagen et al., 2014) and recoded negatively worded items to subsequently compute mean score across the 10 Honesty-Humility items, with a higher mean score indicating higher honesty levels. Scale scores for Overclaiming were computed using mean familiarity rating of the foils (Hülür et al., 2011), with higher ratings indicating a stronger tendency to overclaim knowledge. For the difficult items, we computed the mean percentage correct score across all 34 items from the online assessment. Last, for both (flagged) reaction times and defocusing events, we computed count variables that indicated the number of occurrences of the respective events during the knowledge quiz. In both cases, the count score indicates the number of items for which participants showed suspicious answer behavior.

For analyses on latent level, we computed aggregates to use as indicators in the measurement models. With the exception of honesty, we computed these aggregate scores based on the content domains for all measures. The assignment of content domains to superordinate factors were based on empirical findings on the dimensionality of knowledge (Steger et al., 2019) and was held consistent in all measures. The computation of the domain aggregates was equivalent to the computation of the overall scores described above. For honesty, we computed three separate aggregates based on item sequence in the questionnaire.

Latent Change Score Models. To model score differences between online and lab assessments on a latent level, we estimated latent change score (LCS) models (McArdle, 2009)—a specific class of structural equation models. Originally, LCS were developed to directly capture and predict interindividual differences in intraindividual change, that is, the difference in scores between two time points as an unobservable (latent) variable in longitudinal data. In the present case, LCS models are used to estimate changes between two experimental conditions (online vs. lab) with lab as a reference, while assuming measurement invariance between conditions and taking into account measurement error.

Open Science. We conducted all analyses using R version 3.5.1 (R Core Team, 2018). Confirmatory factor analyses and LCS models were estimated using the *lavaan* package version 0.6-2 (Rosseel, 2012). To make the present analyses transparent and reproducible (Nosek et al., 2015), we provide all material (i.e., data, syntax, and additional tables and figures) online within the *Open Science Framework*: https://osf.io/74p2w/

Results

Descriptive Analyses

We report scores from the complete sample (see Table 1) because the random presentation of test forms did not affect

Table 1. Descriptive Statistics of the Knowledge Tests, S-Data, T-Data, and P-Data Indicators.

	Descriptives							Correlations									
	N	*M*	*SD*	Minimum	Maximum	Skewness	Kurtosis	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Declarative knowledge																	
(1) Total score (online)	312	.66	.12	.32	.94	−0.11	−0.27										
(2) Total score (lab)	310	.57	.10	.24	.85	−0.25	0.22	.52									
(3) Difference score	307	.08	.11	−.21	.43	0.78	0.35	.61	−.39								
Self-report data																	
(4) Honesty-Humility (online)	311	3.44	0.59	1.30	5.00	−0.19	−0.07	.02	.02	−.01							
(5) Overclaiming (lab)	314	1.55	0.42	1.00	3.62	1.61	3.50	.08	−.02	.12	−.09						
Test data																	
(6) Difficult items (online)	312	.34	.19	.06	.97	1.45	1.55	.56	.00	.63	−.03	.15					
(7) Difficult items (lab)	310	.23	.07	.06	.47	0.34	0.29	−.02	.00	−.02	−.05	−.02	−.02				
Para data																	
(8) Flagged RTs (online)	312	5.79	9.02	0	69	3.10	13.50	.47	−.13	.63	−.02	.07	.60	−.05			
(9) Flagged RTs (lab)	310	0.05	0.23	0	2	5.01	26.77	−.01	−.07	.06	−.04	.08	.04	−.09	.11		
(10) Defocusing events (online)	312	20.70	29.62	0	127	1.63	1.90	.56	−.05	.66	.01	.12	.76	−.02	.59	.06	
(11) Defocusing events (lab)	310	0.09	0.40	0	5	7.61	79.23	−.11	−.05	−.08	.02	−.07	−.02	−.05	.05	.20	−.04

Note. RT = reaction time. For declarative knowledge scales and difficult items, we report the percentage correct answers; for defocusing events and RTs, we report the mean number of defocusing events or flagged reaction times; for overclaiming, we report mean familiarity rating of foils; and for honesty-humility, we report the scale mean. For the correlations, sample size of pairwise-present data ranged between 306 and 314. All correlations $r \geq .12$ are significant ($p < .05$).

knowledge test scores or other characteristics (see Table S2 in the online supplement). As intended by the instruction, both general knowledge scores and difficult items scores were higher in the online condition (see also Figure S1 in the online supplement). In the online condition, participants switched browser tabs on average 21 times during the knowledge assessment, while in the lab condition virtually no defocusing events were logged. We found the same pattern for flagged response times (i.e., response times three standard deviations above the median). This pattern of results suggests that participants did in fact cheat in the online condition to enhance their scores, but they had no chance to do so in the proctored lab condition. Similarly, the correlations showed the same pattern as expected when some participants cheat during the unproctored assessment: The mean correlation between online and lab knowledge scores was moderate ($r = .52, N = 307, p < .01$), indicating low rank order stability. Unsurprisingly, the count data variables (i.e., number of flagged response times and number of defocusing events) had high skewness and kurtosis values.

Knowledge difference scores correlated substantially with the number of defocusing events and the number of flagged response times during the online assessment. This means that participants with higher knowledge scores in the online assessment also tended to leave the test pages more frequently and for longer amounts of time. As a first estimate of the prevalence of cheating, we regressed the online knowledge score on the lab knowledge score and screened for participants whose empirical online score did not lie within the 90% confidence interval of their predicted online knowledge score—resulting in 38 participants (12%) with conspicuously high online knowledge scores.

Cheating Prediction

We conducted a hierarchical multiple regression with manifest indicators to gauge the potential of different indicators to predict cheating. As criterion for cheating, we used the difference score between the lab and the online condition, with higher scores reflecting stronger differences in favor of the unproctored online relative to the proctored lab assessment (Table 2) . In a first step, we included S-data predictors, that is, honesty-humility and overclaiming into the model. In a next step, we added the proportion correct score of difficult items as a T-data predictor in the model. Finally, we added all P-data indicators, that is, response times and defocusing events into the model. In contrast to S-data, both T-data and P-data predict score differences between assessments. In total, the variables included in the final model explain half of the interindividual differences. Since the predictors had high zero-order correlations, we calculated the *variance inflation factor* (VIF; see also Chatterjee & Price, 1991) to check for multicollinearity, which was not the case (i.e., all indicators had VIF < 3, thus falling well below common cutoff scores; for example, see also Hair et al., 1995; Neter et al., 1989). Additionally, we checked for normality of the residuals and homoscedasticity using diagnostic plots (see Figure S1 in the online supplement). Results were robust against outlier removal (see also sensitivity analyses in Table S3 in the online supplement).

To complement the analyses, we also computed an LCS model, which we also extended by several variables to predict the LCS. Before fitting these models, we checked measurement models of all traits for adequate model fit (see Table S4 in the online supplement). To account for the

Table 2. Hierarchical Multiple Regression Analyses Predicting Score Differences Between Online and Lab Knowledge Assessment.

	Model 1			Model 2			Model 3		
	B	*SE* (B)	β	B	*SE* (B)	β	B	*SE* (B)	β
Self-report data									
Honesty-humility	<.01	.01	<.01	<.01	.01	.01	<.01	.01	.00
Overclaiming	.03	.01	.12*	.01	.01	.03	.01	.01	.03
Test data									
Difficult items				.37	.03	.63*	.11	.04	.19*
Para data									
Reaction times							<.01	<.01	.33*
Defocusing events							<.01	<.01	.32*
R^2_{adj}		.01			.40			.53	
ΔR^2_{adj}					.39			.13	
AIC		−478.18			−628.23			−702.70	
BIC		−463.30			−609.63			−676.66	

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion.
*p < .05.

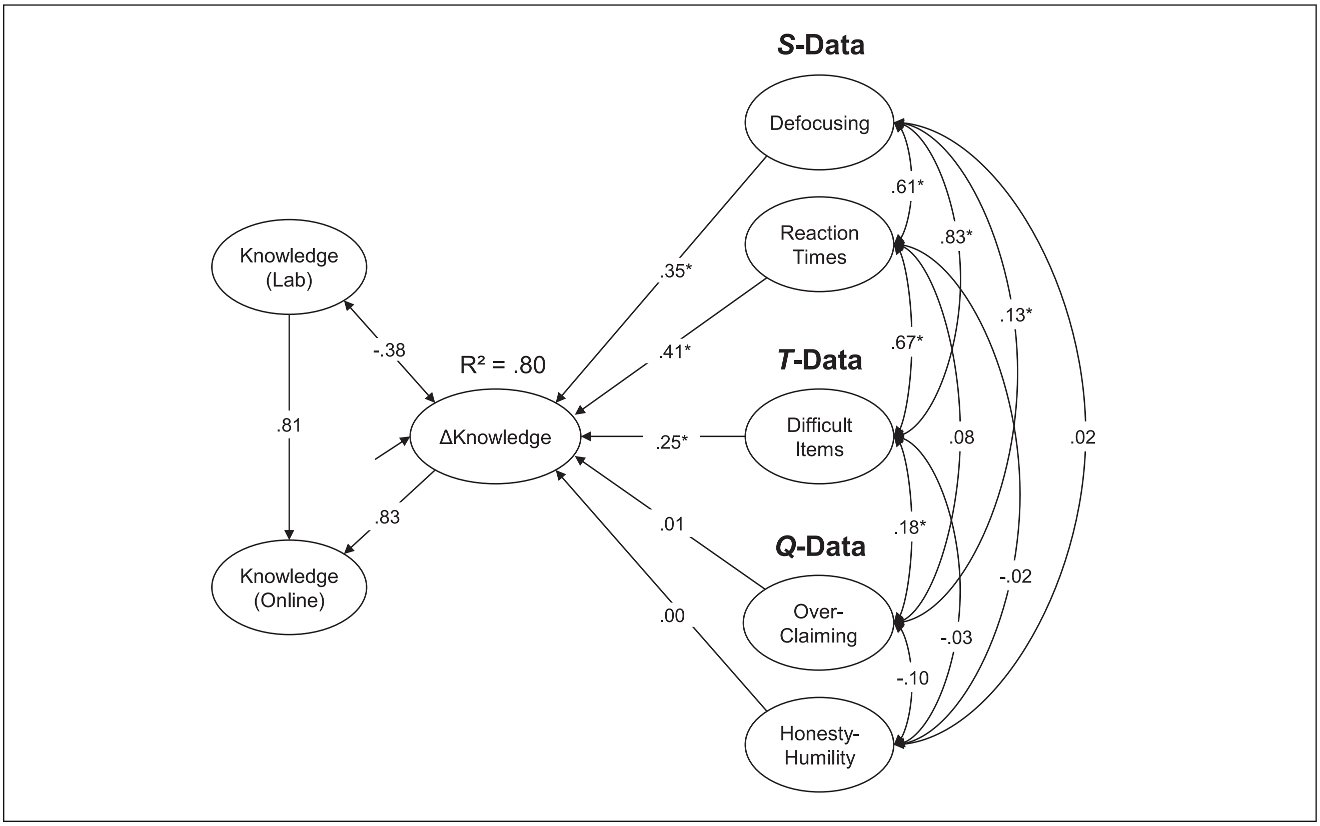

Figure 1. Extended latent change model.
Note. Indicators, residual correlations, and correlations between predictors and the proctored knowledge score were omitted for readability. A complete overview over correlations between latent factors can be found in Table S5 in the online supplement.
*p < .05.

nonnormality in the data, all models were estimated using a maximum likelihood estimator with robust standard errors, which is also suited for nonnormally distributed indicators (Gao et al., 2020). Next, we estimated the LCS model (Figure S3 in the online supplement), which fits the data well ($N = 315$, $\chi^2 = 69.02$, $df = 42$, $p < .01$, comparative fit index (CFI) = .97, root mean square error of approximation (RMSEA) = .05, standardized root mean square residual (SRMR) = .05). The negative correlation ($\rho = -.25$) between the proctored lab knowledge factor and the latent change variable indicates that participants with lower knowledge scores tend to have larger differences between online and lab session. Congruent with previous findings on cheating in academic contexts (Whitley, 1998), this might indicate that participants with lower initial knowledge scores are more likely feel the urge to cheat in order to pass the required knowledge score so that they may enter the lottery.

We extended the LCS model using the previously discussed covariates to predict the latent change. We included all predictors simultaneously (Figure 1). The overall model fit is good ($N = 315$, $\chi^2 = 904.77$, $df = 482$, $p < .01$, CFI = .94, RMSEA = .05, SRMR = .05). Taken together, the indicators explain a total of $R^2 = .80$ of the variance in the latent change variable.

In the extended LCS model, all predictors, except one, are uncorrelated with the lab knowledge score (Table S5 in the online supplement). The only exception is reaction time, which correlates negatively ($\rho = -.15$) with the lab knowledge score: Participants with a higher knowledge score tend to have a smaller amount of flagged response times. With the prediction model, we replicated the findings from the multiple regression analysis: Difficult items, reaction times, and defocusing events predict score differences between lab and online knowledge scores significantly, but honesty and overclaiming do not predict score differences.

Discussion

Data collections in unproctored settings become more and more popular. Current trends include smartphone-based assessments (Pahor et al., 2018; Stieger et al., 2018), online panels (Hays et al., 2015), and large-scale web assessments (Condon & Revelle, 2014). In clinical settings, ambulatory assessment also receives increasing attention (Carpenter et al., 2016; Sliwinski et al., 2018; Wright & Zimmermann, 2019), as it allows to study dynamic processes and to integrate an intraindividual perspective into psychological research. For example, ambulatory assessment can be used to further our understanding of psychological mechanisms underlying mental illnesses (Zimmermann et al., 2019) or for mobile health interventions (see Naslund et al., 2015 for an overview). However, initial enthusiasm about these new data sources was rapidly followed by critical concerns about data quality (e.g., Aust et al., 2012; Buchanan & Scofield, 2018). If we transpose assessments from traditional lab settings to various online platforms, we give up control of test takers' behavior, ultimately leading to the need to flag unusual response patterns post hoc.

In this article, we explore to what extent cheating affects unproctored ability testing. To trigger cheating, we used a declarative knowledge test. Such measures are particularly vulnerable to cheating (Bloemers et al., 2016). As predicted, we found that higher mean scores in the unproctored versus the proctored assessment and the moderate correlations between unproctored and proctored test scores are in line with recent meta-analytic findings (Steger et al., 2020). Accordingly, we interpret the score differences as cheating. A presupposition of this approach is that participants are likely to cheat if they are given incentives and opportunities to do so (Geiger et al., 2018; Moshagen & Hilbig, 2017). In the present study, the major incentive provided was the possibility to participate in a draw for a gift card, which seemed to be sufficiently incentivizing to cheat for a substantial proportion of participants.

A second goal of the current article was to predict cheating behavior with S-data (Honesty-Humility and Overclaiming scales), T-data (extremely difficult items), and P-data (response times and switching between browser tabs). In the following, we discuss the different informational sources in more detail and discuss their potential in detecting cheating.

The link between S-data and deceptive behaviors is widely discussed in the literature (see Heck et al., 2018 for an overview). Although Honesty-Humility seemed to be a promising candidate for predicting cheating, we found honesty to be unrelated not only to cheating but also to every other covariate of the study. At least for the missing link with overclaiming, these results are not surprising, given that previous studies also failed to establish a relation between Honesty-Humility and Overclaiming (Dunlop et al., 2017; Müller & Moshagen, 2019). In the same vein, overclaiming did not contribute substantially to predicting cheating—neither on a manifest nor on a latent level. Based on the present results, we cannot recommend the use of self-reported honesty or overclaiming measures to detect cheating in performance measures. It is up to future research to examine whether other self-report measures perform better in predicting the kind of cheating studied here—as for example, measures assessing current achievement motivation (Freund et al., 2011) or facets of the dark personality (Moshagen et al., 2018). An advantage of current achievement motivation is that it juxtaposes participants' achievement motive (McClelland et al., 1953; see also Steinmayr & Spinath, 2008) as a potential influencing factor of task performance (Freund & Holling, 2011) with situational task characteristics—such as task relevance, task difficulty, or participant's interest in the task. Participants might feel tempted to cheat, for example, if they perceive the given task (or its outcome) as relevant. On a more general stance, participants might cheat more based on their situation-related motivation (Murdock & Anderman, 2006), attitudes (Davy et al., 2007), values (Pulfrey & Butera, 2013), or beliefs (Vohs & Schooler, 2008). This situational-specific approach might also interact with the more person-centered viewpoint of the dark personality, which relies on the individual tendency to maximize one's own benefits at all costs. Accordingly, participants with high scores in dark traits (e.g., self-interest or Machiavellianism) might seek to maximize their scores with minimum effort (Gerbasi & Prentice, 2013), thus engaging in cheating more easily. Additionally, S-data in general might also be more suitable to detect faking in other self-report measures, rather than cheating on ability tests.

As T-data, we used almost unsolvable items containing highly specialized knowledge from various domains, which turned out to be an efficient measure. In the lab setting, performance was on chance level—indicating that difficult items were unaffected by test wiseness (Hartung et al., 2017). In the online condition, performance was on average 1.5 standard deviations higher. The proportion correct score of difficult items in the online condition significantly predicted cheating, with an increment over and above S-data of 39% of explained variance. Correspondingly, on a latent level, performance on the very difficult items in the online condition significantly predicted the LCS. Although these results are promising, the measure we used is highly task-specific: It cannot be readily transferred to other contexts. The general scheme in developing such measures could be to "ask for the impossible" and, thus, to elicit—and, ultimately, observe—deceptive behavior. Possible disadvantages of such measures include additional test time, which is especially problematic in large-scale assessments, and a possible decline in test motivation. These measures are not limited to the application in technology-based settings; they can also be integrated in traditional paper-pencil assessments: Applied alone, T-data serve as a solid predictor of cheating, explaining 40% of the variance.

However, P-data additionally accounted for 13% of the variation in score differences over and above the factor for difficult items, resulting in 53% explained variance. These results illustrate the usefulness of technology-based methods, since P-data often simply come as by-products of computer-based assessments (Couper, 2005; Kroehne & Goldhammer, 2018). Similarly, in the extended LCS model, difficult items, response times, and defocusing events significantly predicted the LCS, explaining 80% of its variance. Response times have been linked to faking behavior in self-report assessments (Maricuț oiu & Sârbescu, 2019; Roma et al., 2019), with participants taking longer to produce dishonest responses. Supposedly, this relation is even more straightforward in ability assessment because searching the web for the correct solution takes time. Furthermore, defocusing events (i.e., switching browser tabs) are a special form of P-data that have been designed to detect cheating behavior in online ability tests (Diedenhofen & Musch, 2017). Nevertheless, neither prolonged response times nor browser tab switches necessarily indicate cheating—we simply do not know what participants are doing when leaving the test page. Only the frequent occurrence of such suspicious behavior might indicate an increased likelihood that people cheat. Definitely, more research is needed in finding aberrant response patterns in complex data. For example, in the case of response times, there might be a u-shaped relationship: Cheating might only occur in a moderate range of response times. Besides, the logic dependence between different P-data sources (e.g., defocusing events and prolonged response times) might result in multicollinearity and biased results, although our checks did not raise concerns in the present case. Clearly, sophisticated models need to be developed to account for the complexity of the data. Other sources of P-data—as for example, mouse clicks (Kieslich & Henninger, 2017), or log data (Boubekki et al., 2016; Kroehne & Goldhammer, 2018)—might be integrated in these models and contribute even further to our understanding of participants' test-taking behavior.

Limitations and Future Research

In this study, cheating was not directly observed; instead, it was computed or modeled as a score difference between two conditions in an experimental setting. These score differences cannot be directly equated with cheating because systematic bias (e.g., declining motivation during a longer lab session) and unsystematic noise (e.g., fluctuation in participants' performance) influence test scores as well. The difference between proctored and unproctored knowledge test scores might also hinge on unmeasured variables such as the reduction of test anxiety when completing the test at home, where the performance pressure might be less prevalent (Stowell & Bennett, 2010). Future studies might find criteria for cheating behavior that are more specific than the difference scores that we used in the present study. Furthermore, these difference scores rely on proctored lab testing as the gold standard to prevent cheating behavior. However, cheating can also occur and succeed in proctored testing (Drasgow et al., 2009). But how can we determine if someone cheated in unproctored settings? Cheating is only directly observable using supervision, sometimes in the form of screen monitoring or webcam surveillance (Karim et al., 2014). Such external control could be perceived as invasive, which might lead to biased test results. Another approach might be to ask participants after the test whether they cheated. Since cheating is a socially undesirable behavior, direct questioning of participants might deliver invalid data. It is very likely that participants substantially underreport their cheating once asked directly (Hoffmann et al., 2015). Therefore, we deem the present indicators superior to an ex post facto self-accusation of cheating. Potentially, indirect questioning approaches such as the randomized response technique (Moshagen et al., 2012) could be applied after the test session. However, this approach does not allow identifying individual cheaters—it allows for an estimate of cheating prevalence in online assessments.

Cheating is a problem in individual settings, but it might also bias the results of applied and basic research that rely on uncleaned data gathered in an unproctored assessment. Unfortunately, our understanding of cheaters is still limited: Who cheats and why? Under which circumstances are aspects of the person more important than the situation and vice versa? What keeps noncheaters from cheating? Or what makes a successful cheater? In the present study, opportunity to cheat was held constant for all participants in both conditions by experimentally varying the level of proctoring. But participants differ in the anticipated costs and utility for the participants (Thielmann & Hilbig, 2018) and also in their ability (Geiger et al., 2018). Generally, participants might engage in cheating behavior when several criteria are met: They must have the opportunity to so, anticipated benefits should outweigh the anticipated costs of possible sanctions, and they must have the necessary skills. Future research should direct attention to the identification of potential cheaters not only because it is a nuisance in psychological assessment but also because it conveys interesting diagnostic information. Importantly, cheating as it was captured here must not be understood as some overarching highly general behavioral disposition that is stable over time. There are many more facets of cheating and honesty, and our understanding of the structure of this domain is still very limited.

Conclusion

Unproctored data collection inevitably provokes the question, "How we can ensure data quality?" Test administrators must be aware that unproctored settings are likely to deliver biased or invalid data for at least some participants (see also Steger et al., 2020) and, accordingly, interpret results with caution. Both researchers and practitioners should keep in mind potential biases that may arise from different test settings. When the stakes are high, proctored testing is still the gold standard to prevent cheating. Obviously, this does not imply that unproctored ability tests cannot be used in practice. However, in low-stakes and high-stakes settings alike, data should be routinely screened for unusual test behavior. In the present study, we demonstrated how this can be done for unproctored knowledge tests. While the S-data indicators we used in the present study failed to predict cheating, T-data and P-data indicators can be used to assess data quality (i.e., estimating the prevalence of cheating in the present data and estimating the extent to which the data are biased) and to develop a transparent procedure of how to deal with potential cheaters. With both T-data and P-data indicators being more or less direct observations of cheating behavior, this result also illustrates the necessity to integrate behavior measures into psychometric research. Ultimately, these data types provide indicators that are almost impossible to fake. Importantly, this applies not only to measures of cognitive abilities but also to measures of typical behavior, even if, in this case, aberrant behavior might look different (e.g., extreme short response times indicating superficial reading). However, more sophisticated models and more appropriate methods are needed.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was generously supported by the Bamberg Graduate School of Social Sciences, which is funded by the German Research Foundation (DFG) under the German Excellence Initiative (GSC1024).

ORCID iD

Diana Steger https://orcid.org/0000-0002-5282-6934

Supplemental Material

Supplemental material for this article is available online.

References

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150-166. https://doi.org/10.1177/1088868306294907

Ashton, M.C., & Lee, K. (2008). The HEXACO model of personality structure and the importance of the H factor. *Social and Personality Psychology Compass*, *2*(5), 1952-1962. https://doi.org/10.1111/j.1751-9004.2008.00134.x

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*(2), 139-152. https://doi.org/10.1177/1088868314523838

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2012). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*, 527-535. https://doi.org/10.3758/s13428-012-0265-2

Bensch, D., Paulhus, D. L., Stankov, L., & Ziegler, M. (2019). Teasing apart overclaiming, overconfidence, and socially desirable responding. *Assessment*, *26*(3), 351-363. https://doi.org/10.1177/1073191117700268

Bing, M. N., Kluemper, D., Kristl Davison, H., Taylor, S., & Novicevic, M. (2011). Overclaiming as a measure of faking. *Organizational Behavior and Human Decision Processes*, *116*(1), 148-162. https://doi.org/10.1016/j.obhdp.2011.05.006

Bloemers, W., Oud, A., & Dam, K. van. (2016). Cheating on unproctored internet intelligence tests: Strategies and effects. *Personnel Assessment and Decisions*, *2*(1), 21-29. https://doi.org/10.25035/pad.2016.003

Boubekki, A., Kröhne, U., Goldhammer, F., Schreiber, W., & Brefeld, U. (2016). Data-driven analyses of electronic text books. In S. Michaelis, N. Piatkowski, & M. Stolpe (Eds.), *Solving large scale learning tasks. Challenges and algorithms* (pp. 362-376). Springer. https://doi.org/10.1007/978-3-319-41706-6_20

Bressan, M., Rosseel, Y., & Lombardi, L. (2018). The effect of faking on the correlation between two ordinal variables: Some population and Monte Carlo results. *Frontiers in Psychology*, *9*, 1876. https://doi.org/10.3389/fpsyg.2018.01876

Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*, 2586-2596. https://doi.org/10.3758/s13428-018-1035-6

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahstrom, W. G., & Kaemmer, B. (2001). *MMPI-2: Manual for administration and scoring*. University of Minnesota Press.

Carpenter, R. W., Wycoff, A. M., & Trull, T. J. (2016). Ambulatory assessment: New adventures in characterizing dynamic processes. *Assessment*, *23*(4), 414-424. https://doi.org/10.1177/1073191116632341.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire*. Institute for Personality and Ability Testing.

Chatterjee, S., & Price, B. (1991). *Regression diagnostics*. Wiley.

Cicek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Lawrence Erlbaum.

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52-64. https://doi.org/10.1016/j.intell.2014.01.004

Couper, M. P. (2005). Technology trends in survey data collection. *Social Science Computer Review*, *23*(4), 486-501. https://doi.org/10.1177/0894439305278972

Davy, J. A., Kincaid, J. F., Smith, K. J., & Trawick, M. A. (2007). An examination of the role of attitudinal characteristics and motivation on the cheating behavior of business students. *Ethics & Behavior*, *17*(3), 281-302. https://doi.org/10.1080/10508420701519304

Diedenhofen, B., & Musch, J. (2017). PageFocus: Using paradata to detect and prevent cheating on online achievement tests. *Behavior Research Methods*, *49*, 1444-1459. https://doi.org/10.3758/s13428-016-0800-7

Drasgow, F., Nye, C. D., Guo, J., & Tay, L. (2009). Cheating on proctored tests: The other side of the unproctored debate. *Industrial and Organizational Psychology*, *2*(1), 46-48. https://doi.org/10.1111/j.1754-9434.2008.01106.x

Dunlop, P. D., Bourdage, J. S., de Vries, R. E., Hilbig, B. E., Zettler, I., & Ludeke, S. G. (2017). Openness to (reporting) experiences that one never had: Overclaiming as an outcome of the knowledge accumulated through a proclivity for cognitive and aesthetic exploration. *Journal of Personality and Social Psychology*, *113*(5), 810-834. https://doi.org/10.1037/pspp0000110

Fell, C. B., König, C. J., Jung, S., Sorg, D., & Ziegler, M. (2019). Are country level prevalences of rule violations associated with knowledge overclaiming among students? *International Journal of Psychology*, *54*(1), 17-22. https://doi.org/10.1002/ijop.12441

Freund, P. A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personality and Individual Differences*, *50*(5), 723-728. https://doi.org/10.1016/j.paid.2010.12.025

Freund, P. A., Kuhn, J.-T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, *51*(5), 629-634. https://doi.org/10.1016/j.paid.2011.05.033

Gao, C., Shi, D., & Maydeu-Olivares, A. (2020). Estimating the maximum likelihood root mean square error of approximation (RMSEA) with non-normal data: A Monte-Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(2), 192-201. https://doi.org/10.1080/10705511.2019.1637741

Geiger, M., Olderbak, S., Sauter, R., & Wilhelm, O. (2018). The "g" in faking: Doublethink the validity of personality self-report measures for applicant selection. *Frontiers in Psychology*, *9*, Article 2153. https://doi.org/10.3389/fpsyg.2018.02153

Gerbasi, M. E., & Prentice, D. A. (2013). The self- and other-interest inventory. *Journal of Personality and Social Psychology*, *105*(3), 495-514. https://doi.org/10.1037/a0033483

Gnambs, T., & Kaspar, K. (2017). Socially desirable responding in web-based questionnaires: A meta-analytic review of the candor hypothesis. *Assessment*, *24*(6), 746-762. https://doi.org/10.1177/1073191115624547

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*(2-3), 108-119. https://doi.org/10.1016/j.intell.2011.02.001

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment*, *18*(4), 351-364. https://doi.org/10.1111/j.1468-2389.2010.00518.x

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (3rd ed.). Macmillan.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838-854. https://doi.org/10.1177/1745691616650285

Hartung, J., Weiss, S., & Wilhelm, O. (2017). Individual differences in performance on comprehension and knowledge tests with and without passages and questions. *Learning and Individual Differences, 56*, 143-150. https://doi.org/10.1016/j.lindif.2016.11.001

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory*. University of Minnesota Press.

Hays, R. D., Liu, H., & Kapteyn, A. (2015). Use of Internet panels to conduct surveys. *Behavior Research Methods*, *47*, 685-690. https://doi.org/10.3758/s13428-015-0617-9

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making*, *13*(4), 356-371.

Hilbig, B. E., Moshagen, M., & Zettler, I. (2015). Truth will out: Linking personality, morality, and honesty through indirect questioning. *Social Psychological and Personality Science*, *6*(2), 140-147. https://doi.org/10.1177/1948550614553640

Hilbig, B. E., & Zettler, I. (2015). When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality*, *57*, 72-88. https://doi.org/10.1016/j.jrp.2015.04.003

Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2015). A strong validation of the crosswise model using experimentally-induced cheating behavior. *Experimental Psychology*, *62*(6), 403-414. https://doi.org/10.1027/1618-3169/a000304

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*, 1436-1442. https://doi.org/10.3758/s13428-014-0524-5

Hülür, G., Wilhelm, O., & Schipolowski, S. (2011). Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement. *Learning and Individual Differences*, *21*(6), 742-746. https://doi.org/10.1016/j.lindif.2011.09.006

Johnson, J. A. (2001). Personality psychology: Methods. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 11313-11317). Pergamon.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298. https://doi.org/10.1207/S15324818AME1604_2

Karim, M. N., Kaminsky, S. E., & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business and Psychology*, *29*, 555-572. https://doi.org/10.1007/s10869-014-9343-z

Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*, *49*, 1652-1667. https://doi.org/10.3758/s13428-017-0900-z

Krantz, J. H., & Reips, U.-D. (2017). The state of web-based research: A survey and call for inclusion in curricula. *Behavior Research Methods*, *49*, 1621-1629. https://doi.org/10.3758/s13428-017-0882-x

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*, 527-563. https://doi.org/10.1007/s41237-018-0063-y

Lee, K., Ashton, M. C., & de Vries, R. E. (2005). Predicting workplace delinquency and integrity with the HEXACO and five-factor models of personality structure. *Human Performance*, *18*(2), 179-197. https://doi.org/10.1207/s15327043hup1802_4

Lee, K., Ashton, M. C., Wiltshire, J., Bourdage, J. S., Visser, B. A., & Gallucci, A. (2013). Sex, power, and money: Prediction from the dark triad and honesty-humility. *European Journal of Personality*, *27*(2), 169-184. https://doi.org/10.1002/per.1860

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology*, *84*(4), 817-824. https://doi.org/10.1348/096317910X522672

MacCann, C. (2013). Instructed faking of the HEXACO reduces facet reliability and involves more Gc than Gf. *Personality and Individual Differences*, *55*(7), 828-833. https://doi.org/10.1016/j.paid.2013.07.007

Maricuțoiu, L. P., & Sârbescu, P. (2019). The relationship between faking and response latencies: A meta-analysis. *European Journal of Psychological Assessment*, *35*(1), 3-13. https://doi.org/10.1027/1015-5759/a000361

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*(1), 577-605. https://doi.org/10.1146/annurev.psych.60.110707.163612

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive*. Appleton-Century-Crofts.

McClintock, J. C. (2016). Reduction in cheating following a forensic investigation on a statewide summative assessment. *Applied Measurement in Education*, *29*(2), 132-143. https://doi.org/10.1080/08957347.2016.1138958

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8. https://doi.org/10.1207/s15324818ame0901_2

Moshagen, M., & Hilbig, B. E. (2017). The statistical analysis of cheating paradigms. *Behavior Research Methods*, *49*, 724-732. https://doi.org/10.3758/s13428-016-0729-x

Moshagen, M., Hilbig, B. E., & Zettler, I. (2014). Faktorenstruktur, psychometrische Eigenschaften und Messinvarianz der deutschsprachigen Version des 60-Item HEXACO Persönlichkeitsinventars [Factor structure, psychometric features and measurement invariance of the German version of the 60-item HEXACO personality inventory]. *Diagnostica*, *60*(2), 86-97. https://doi.org/10.1026/0012-1924/a000112

Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological Review*, *125*(5), 656-688. https://doi.org/10.1037/rev0000111

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods*, *44*, 222-231. https://doi.org/10.3758/s13428-011-0144-2

Müller, S., & Moshagen, M. (2019). True virtue, self-presentation, or both? A behavioral test of impression management and overclaiming. *Psychological Assessment*, *31*(2), 181-191. https://doi.org/10.1037/pas0000657

Murdock, T. B., & Anderman, E. M. (2006). Motivational perspectives on student cheating: Toward an integrated model of academic dishonesty. *Educational Psychologist*, *41*(3), 129-145. https://doi.org/10.1207/s15326985ep4103_1

Naslund, J. A., Marsch, L. A., McHugo, G. J., & Bartels, S. J. (2015). Emerging mHealth and eHealth interventions for serious mental illness: A review of the literature. *Journal of Mental Health (Abingdon, England)*, *24*(5), 321-332. https://doi.org/10.3109/09638237.2015.1019054

Neter, J., Wassermann, W., & Kutner, M. H. (1989). *Applied linear regression models*. Irwin.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality, 63*(August), 1-11. https://doi.org/10.1016/j.jrp.2016.04.010

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1420-1422. https://doi.org/10.1126/science.aab2374

Nye, C. D., Do, B.-R., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment*, *16*(2), 112-120. https://doi.org/10.1111/j.1468-2389.2008.00416.x

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 43-72). Wiley. https://doi.org/10.1002/9781118596869.ch3

O'Neill, H. M., & Pfeiffer, C. A. (2012). The impact of honour codes and perceptions of cheating on academic cheating behaviours, especially for MBA bound undergraduates. *Accounting Education*, *21*(3), 231-245. https://doi.org/10.1080/09639284.2011.590012

Pahor, A., Stavropoulos, T., Jaeggi, S. M., & Seitz, A. R. (2018). Validation of a matrix reasoning task for mobile devices. *Behavior Research Methods*, *51*, 2256-2267. https://doi.org/10.3758/s13428-018-1152-2

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890-904. https://doi.org/10.1037/0022-3514.84.4.890

Phillips, D. L., & Clancy, K. J. (1972). Some effects of "social desirabilty" in survey studies. *American Journal of Psychology*, *77*(5), 921-940.

Pulfrey, C., & Butera, F. (2013). Why neoliberal values of self-enhancement lead to cheating in higher education: A motivational account. *Psychological Science*, *24*(11), 2153-2162. https://doi.org/10.1177/0956797613487221

R Core Team. (2018). *R: A language and environment for statistical computing* (Version 3.5.1). https://www.R-project.org/

Roma, P., Mazza, C., Mammarella, S., Mantovani, B., Mandarelli, G., & Ferracuti, S. (2019). Faking-good behavior in self-favorable scales of the MMPI-2: A study with time pressure. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000511

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. http://doi.org/10.18637/jss.v048.i02

Rovai, A. P. (2000). Online and traditional assessments: What is the difference? *Internet and Higher Education*, *3*(3), 141-151. https://doi.org/10.1016/S1096-7516(01)00028-8

Schroeders, U., Bucholtz, N., Formazin, M., & Wilhelm, O. (2013). Modality specificity of comprehension abilities in the sciences. *European Journal of Psychological Assessment*, *29*(1), 3-11. https://doi.org/10.1027/1015-5759/a000114

Schroeders, U., Wilhelm, O., & Schipolowski, S. (2010). Internet-based ability testing. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 131-148). American Psychological Association.

Sliwinski, M. J., Mogle, J. A., Hyun, J., Munoz, E., Smyth, J. M., & Lipton, R. B. (2018). Reliability and validity of ambulatory cognitive assessments. *Assessment*, *25*(1), 14-30. https://doi.org/10.1177/1073191116643164

Slobogin, C. (2005). Mental disorder as an exemption from the death penalty: The ABA-IRR task force recommendations. *Catholic University Law Review*, *54*(4), 1133-1152.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-Index for detecting answer copying. *Journal of Educational Measurement*, *39*(2), 115-132.

Steger, D., Schroeders, U., & Gnambs, T. (2020). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment*, *36*(1), 174-184. https://doi.org/10.1027/1015-5759/a000494

Steger, D., Schroeders, U., & Wilhelm, O. (2019). On the dimensionality of crystallized intelligence: A smartphone-based assessment. *Intelligence*, *72*, 76-85. https://doi.org/10.1016/j.intell.2018.12.002

Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality*, *22*(3), 185-209. https://doi.org/10.1002/per.676

Stieger, S., Lewetz, D., & Reips, U. (2018). Can smartphones be used to bring computer-based tasks from the lab to the field? A mobile experience-sampling method study about the pace of life. *Behavior Research Methods*, *50*, 2267-2275. https://doi.org/10.3758/s13428-017-0991-6

Stowell, J. R., & Bennett, D. (2010). Effects of online testing on student exam performance and test anxiety. *Journal of Educational Computing Research*, *42*(2), 161-171. https://doi.org/10.2190/EC.42.2.b

Thielmann, I., & Hilbig, B. E. (2018). Daring dishonesty: On the role of sanctions for (un)ethical behavior. *Journal of Experimental Social Psychology*, *79*, 71-77. https://doi.org/10.1016/j.jesp.2018.06.009

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology*, *59*(1), 189-225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will encouraging a belief in determinism increases cheating. *Psychological Science*, *19*(1), 49-54. https://doi.org/10.1111/j.1467-9280.2008.02045.x

Whitley, B. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education*, *39*, 235-274. https://doi.org/10.1023/A:1018724900565

Wilhelm, O., & McKnight, P. E. (2002). Ability and achievement testing on the World Wide Web. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 167-193). Hogrefe & Huber.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, *115*(2), 157-168. https://doi.org/10.1016/j.obhdp.2010.10.001

Wright, A. G. C., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment*, *31*(12), 1467-1480. https://doi.org/10.1037/pas0000685

Yarkoni, T. (2012). Psychoinformatics: New horizons at the interface of the psychological and computing sciences. *Current Directions in Psychological Science*, *21*(6), 391-397. https://doi.org/10.1177/0963721412457362

Zettler, I., Lang, J. W. B., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self- and observer reports: Response processes in personality data. *Journal of Personality*, *84*(4), 461-472. https://doi.org/10.1111/jopy.12172

Zimmermann, J., Ritter, S., Masuhr, O., Jaeger, U., Spitzer, C., Woods, W. C., Happel, M., & Wright, A. G. C. (2019). Integrating Structure and Dynamics in Personality Assessment: First Steps Toward the Development and Validation of a Personality Dynamics Diary. *Psychological Assessment*, 516–531. http://dx.doi.org/10.1037/pas0000625