Technology-Based Assessment

A Theoretical Framework, Psychometric Modeling, and Substantial Issues in the Assessment of Cognitive Abilities

Inaugural-Dissertation

in der Fakultät für Humanwissenschaften

der Otto-Friedrich-Universität Bamberg

vorgelegt von

Diana Steger, geb. Klose

aus

Coburg

Bamberg, den 31.10.2019

URN: urn:nbn:de:bvb:473-irb-477401
DOI: https://doi.org/10.20378/irb-47740

Tag der mündlichen Prüfung:	13.12.2019
Dekan:	Universitätsprofessor Dr. Jörg Wolstein
Betreuer:	Universitätsprofessor Dr. Ulrich Schroeders
Weiterer Gutachter:	Universitätsprofessor Dr. Oliver Wilhelm

Acknowledgements

Over the last few years I set out to write my dissertation, I not only busied myself with my studies, but I also kept a considerable amount of people busy. Without these people, I would have never come so far to write these final lines of my thesis and I'm deeply grateful for all the support I received throughout the years.

First, I would like to thank my supervisors Ulrich Schroeders and Oliver Wilhelm for their advice, encouragement, and the excellent training they provided. I truly learned a lot from them and this dissertation would not have been possible without their support.

I also want to thank all my wonderful colleagues I met on the various steps of my journey in Bamberg, Ulm and Kassel and that became friends over the time. Although the list is much longer than that, I especially want to thank Katharina Dorn (for her unparalleled optimism and cheeriness), Mirijam Böhme, Nadja Wehl, and Regine Schmidt (for being the best writing support and the best company during coffee breaks), Gabriel Olaru (for being the most laid-back landlord and room mate I can imagine), and Selina Weiß (for the wonderful evenings with wine and transcendent conversations).

Next, I thank all the friends, colleagues, and student assistants that assisted me in realizing this dissertation project by constructing, reviewing, or proofreading knowledge questions, or advertising the quiz app.

Last, I want to thank my family, starting with my parents Christine Wittmann-Klose and Volker Klose who always had faith in me and my ability to achieve my goals. I also want to thank my sister Maria Klose, who is my unfailing source of support and laughter. Finally, I want to thank my husband Benedikt Steger for his love, support, and patience that guided me through the years.

Contents

List of Tables

List of Figures

Abstract

New assessment technologies yield the potential to shape the way we conduct research in general and how we assess data in particular. On the one hand, technology-based assessment has many advantages, as for example the accessibility of large and diverse samples, the possibility to collect data on dynamic processes, and the opportunity to assess auxiliary data. On the other hand, potential caveats of technology-based assessment include for example decreasing control over the test setting, multi-dimensional and complex data structures, or concerns about ethics and data security. Whether the benefits or the disadvantages outweigh in a given context should be decided on a case-by-case basis. Hence, in this thesis, I focus on the application of technology-based assessment to the measurement of cognitive abilities, starting with an examination of the impact of proctored versus unproctored settings on cognitive ability tests (*Manuscript 1*). Next, I present a smartphone-based assessment of declarative knowledge (*Manuscript 2*). Declarative knowledge is a psychological construct that is particularly hard to measure with traditional means, which is why smartphone-based assessment is a promising candidate to facilitate its measurement. Lastly, I demonstrate how auxiliary data from technology-based assessment can be used to predict cheating in unproctored knowledge assessments (*Manuscript 3*) and how this approach can be used to secure data quality of assessments conducted in unproctored settings. Taken together, these manuscripts explore substantial issues in the technology-based assessment of cognitive abilities and demonstrate a) that some of the drawbacks can pose a substantial threat to the data quality of technology-based assessment, b) that technology-based assessment, however, has the potential to assess psychological constructs that are hard to measure otherwise, and c) how features of technology-based assessment can be used to restore data quality. In the epilogue, I discuss the findings in light of existing literature on advantages and challenges of technology-based assessments and outline future directions for the technology-based assessment of cognitive abilities in general and declarative knowledge in particular.

Zusammenfassung

Neue Technologien der Datenerhebung verändern die Art und Weise wie wir Forschung betreiben und Daten erheben nachhaltig. Auf der einen Seite haben technologie-basierte Erhebungsmethoden viele Vorteile für die Forschung, wie beispielsweise die Verfügbarkeit von großen und heterogenen Stichproben, die Möglichkeit, Informationen über dynamische Prozesse zu erhalten, ebenso wie die Möglichkeit, zusätzliche Daten zu erheben. Auf der anderen gehören der Verlust der Kontrolle über das Testsetting, multidimensionale und komplexe Datenstrukturen und Bedenken über den Datenschutz zu den Vorbehalten gegenüber technologie-basierten Erhebungsmethoden. Ob die Vorteile oder die Nachteile im jeweiligen Anwendungskontext überwiegen muss von Fall zu Fall entschieden werden. In diesem Zuge befasse ich mich in der vorliegenden Dissertation mit dem Einsatz von technologie-basierten Erhebungsmethoden zur Diagnostik von kognitiven Fähigkeiten. Zu Beginn untersuche ich den Einfluss von beaufsichtigten versus unbeaufsichtigten Testsettings auf die Ergebnisse von kognitiven Leistungstests (*Manuskript 1*). Anschließend wird eine smartphone-basierte Erhebung von deklarativem Wissen vorgestellt (*Manuskript 2*). Bei deklarativem Wissen handelt es sich um ein psychologisches Konstrukt, das mit herkömmlichen diagnostischen Ansätzen nur schwer zu erfassen ist, weshalb smartphone-basierte Erhebung gerade hier einen vielversprechenden Ansatz darstellt um Messung von deklarativem Wissen zu erleichtern. Zuletzt werden Daten, die zusätzlich mit der Hilfe von technologie-basierten Erhebungen erfasst werden können, genutzt um unehrliches Testverhalten in unbeaufsichtigten Wissenstestungen vorherzusagen (*Manuskript 3*). Es wird gezeigt, wie dieser Ansatz genutzt werden kann um die Datenqualität von unbeaufsichtigten psychologischen Testungen zu gewährleisten. Insgesamt untersucht die vorliegende Arbeit substantielle Fragen zu technologie-basierten Erhebungen kognitiver Fähigkeiten und zeigt auf, a) dass teilweise die Nachteile technologie-basierter Erhebungen die Datenqualität substantiell beeinträchtigen können, b) dass technologie-basierte Erhebungen dennoch in der Lage sind die Erhebung schwer zu messender psychologische Konstrukte zu ermöglichen und c) wie Eigenschaften technologie-basierter Erhebungen genutzt werden können um die Datenqualität wiederherzustellen. Im Epilog werden die vorliegenden Ergebnisse vor dem Hintergrund der bestehenden Literatur zu Vor- und Nachteilen technologie-basierter Erhebungen diskutiert und zukünftige Forschungsansätze für die technologie-basierte Erhebung von deklarativem Wissen und kognitiven Fähigkeiten aufgezeigt.

I. Prologue

Introduction

To know that we know what we know, and that we do not know what we do not know, that is true knowledge.

Henry David Thoreau, Walden

The idea of a complete and comprehensive measurement of the human is certainly not new: Already in the 19th century, Francis Galton—founder of the disciplines of differential psychology and psychometrics—sought to measure the entirety of an individual (Galton, 1883). His research did not only include outward appearances like weight and eye color, physical abilities like strength and breathing capacity, and psychological features like reaction time to sound and sight (Galton, 1887a, 1887b), but also more obscure behavioral data—always sticking to his motto: "Whenever you can, count" (Pearson, 1914). In this spirit, he and his assistants recorded a plethora of variables, for example by counting the number of students' yawns in his fellow professors' lectures or the number of beautiful people he saw on the street, drawing a "beauty map" of different regions (Berry, 2003)—everything for his mission to map and ultimately to understand humankind. Back in the late 19th century, this endeavor entailed considerable difficulties: Firstly, apparatuses and instruments were immature and certainly prone to measurement error, and the behavioral observations were time-, cost-, and resource-consuming. Secondly, university management and colleagues were indignant: When in 1877 a psychometric laboratory should be installed at Cambridge, the application was rejected because "[such a laboratory] would insult religion by putting the human soul in a pair of scales" (Bartlett, 1937, p. 98; Sokal, 1972). More than one century later, the tide has turned: What seemed to be a megalomaniac project of an individual back then has become a trend today: Anyone can engage in Hobbies such as "self-tracking" (see also https://quantifiedself.com/) and collect a multitude of data—simply by using their smartphones. The types of data that can be collected using a smartphone seem sheer endless: Smartphone data can give us insight not only into our smartphone usage but also tell us something about the number of steps we are making daily or about our sleep quality. And if the smartphone sensors alone might not suffice, a multitude of different sensory expansions and so-called “wearables” exist, that allow the collection of even more data—all adapted to the individual needs. Overall, one could say that we have come a long way since Francis Galton’s first anthropometric laboratories with its curious apparatuses to our own anthropometric laboratories that we all carry around in our pockets and allow us to quantify our lives. In this context, it only seems logical to connect the idea of scientific—and especially psychometric—endeavors with the use of modern technology.

It has been several years since the call for smartphone-based assessment to answer psychological questions has been raised: In 2012, Miller (2012, p. 221) published his euphoric “Smartphone Manifesto”, in which he advertised the power of smartphones to “revolutionize all fields of psychology and other behavioral sciences”. Indeed, this form of technology-based assessment was touted with many advantages: a) accessibility of large samples and the collection of vast amounts of data (e.g., Dufau et al., 2011), b) ecological validity and reduction of various response biases (Ebner-Priemer & Trull, 2009), c) supplementation of traditional test- or questionnaire data with incidental data (Kroehne & Goldhammer, 2018) and auxiliary data from smartphone sensors (e.g., Mehl, 2017)—just to name a few. However, in psychological research, the possibility to use smartphone technology for research and data assessment was adopted only slowly and the field was left to other disciplines: For example, in medical research, smartphone applications were used to augment traditional approaches by remotely diagnosing falls in elderly people (Abbate et al., 2012; Yavuz et al., 2010) or by monitoring symptoms, such as glucose levels in patients diagnosed with diabetes (Tran, Tran, & White, 2012) or behavior during exercising in patients diagnosed with respiratory diseases (Marshall, Medvedev, & Antonov, 2008). Another discipline that included smartphone-based technology from the early beginnings was transportation research and infrastructure planning, for example to diagnose driving styles (Johnson & Trivedi, 2011), support traffic management

(Campolo, Iera, Molinaro, Paratore, & Ruggeri, 2012), or track pedestrians (Kim, Hyojeong Shin, & Cha, 2012). In the field of Psychology, however, general skepticism subsided more slowly and yet, in Psychology the number of projects that dealt more deeply with smartphone-based assessment grew over time (Figure 1; see also Hamaker & Wichers, 2017 for a trend in ambulatory assessment in general).

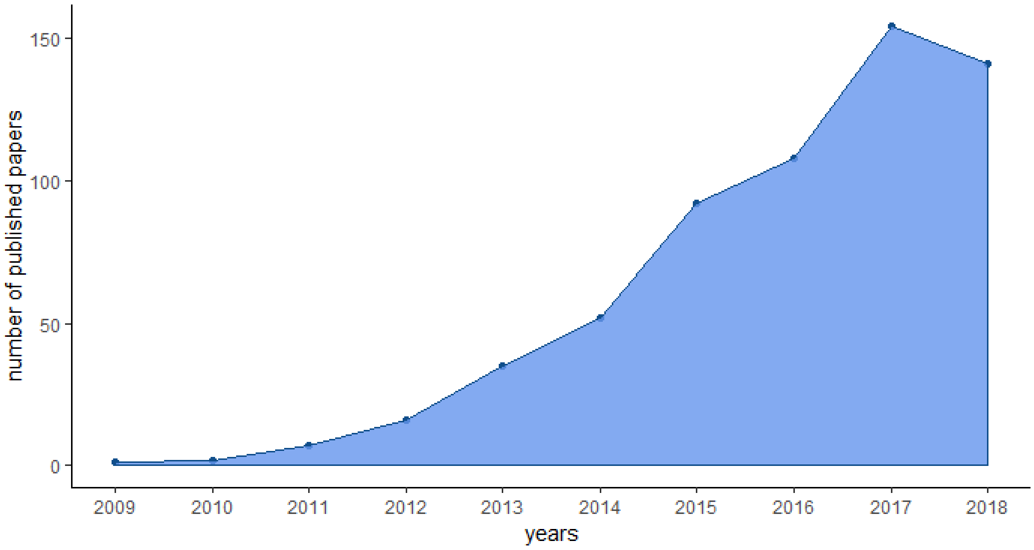

Figure I-1. Papers published on smartphone-based assessment in Psychology.

Note. Papers were identified using the Boolean search term “(smartphone OR smartphone-based) AND (assessment OR testing)” within the databases ERIC, PsycArticles, PsycInfo, and Psyndex.

As depicted in Figure 1, since the publication of Miller’s Smartphone Manifesto in 2012, also in Psychology the number of publications using smartphone-based technology has risen substantially. To date, the main fields of application for smartphone-based research programs has been the sector of eHealth applications for mental health (e.g., Naslund, Marsch, McHugo, & Bartels, 2015) and technology-enhanced education programs (e.g., Kukulska-Hulme & Viberg, 2018). In both cases, the smartphone has become a tool to administer interventions to a specific population. For example, a review of 46 studies on eHealth interventions showed that smartphones were used to support the whole therapeutical process, including psychoeducation, symptom monitoring, compliance, and relapse prevention (Naslund et al., 2015). Apart from these applied contexts, smartphone-based assessment became also more prevalent to study psychological constructs. Especially the possibility to implement intensive longitudinal designs (Bolger & Laurenceau, 2013) was used to study constructs that are less stable over time—such as mood and affective well-being (e.g., Wrzus, Wagner, & Riediger, 2014) or the relationship between state- and trait measures (Rauthmann, Horstmann, & Sherman, 2019). Recently, also mobile sensing studies are gaining more popularity in psychological research. Using the mobile sensing method, data from smartphone sensors are collected and analyzed—offering a more direct measure of human behavior (Harari et al., 2016), which is also less prone to bias than traditional methods. Overall, smartphones in research seem to be a promising tool to get new insights into psychological constructs, or at least to examine them from a new perspective. Smartphones can be used not only to target specific populations and allow for flexible data collection but also to examine dynamic processes over time and to directly measure behavior.

Certainly, there are many more research areas that can benefit from smartphone-based assessment approaches, such as, for example, the area of cognitive abilities. Generally, intelligence is a well-researched subject in psychology, but with an obvious imbalance with regard to the current state of knowledge about the different elements of cognitive abilities. Looking at the two most prevalent components of intelligence—namely, *fluid intelligence* and *crystallized intelligence*—fluid intelligence, on the one hand, seems fairly well established, while crystallized intelligence has gained less attention over time. According to Cattell's (1941, 1943) theory on Fluid and Crystallized Intelligence, crystallized intelligence encompasses skills, knowledge, and language-related abilities in a broad range of domains as crystallized intelligence can be seen as the result of an investment of fluid intelligence in diverse learning situations (Cattell, 1971). In line with Cattell's emphasis on the role of broad declarative knowledge, the ideal measurement of crystallized intelligence should be broad and cover a variety of different knowledge domains (Ackerman, 1996; see also Wilhelm & Schroeders, 2019). In contrast, crystallized intelligence is widely assessed using mostly indicators of verbal abilities, thus covering only a section of possible knowledge domains and therefore systematically neglecting important parts of the overarching factor of crystallized intelligence (Schipolowski, Wilhelm, & Schroeders, 2015)—probably due to the fact that a broad assessment of declarative knowledge is difficult to construct and time-consuming to apply using traditional assessment approaches.

To this end, the present dissertation deals with the overarching question of which new insights into cognitive abilities we can gain from the use of new technology-based assessment techniques, and—more specifically—how it is possible to use smartphone-based assessment to study declarative knowledge. In the following, I first give an overview of new advances in technology-based assessment, discuss potential advantages and disadvantages of these methods, and describe potential applications in the research on cognitive abilities. Second, I focus on declarative knowledge as a potential candidate for the application of technology-based assessment techniques by introducing the state of research in the field, highlighting problems in traditional assessments of declarative knowledge and discussing how technology-based assessments might contribute to improving the measurement of declarative knowledge.

Technology-Based Assessment

Technology-based assessment is a generic term for computer- and smartphone-based assessment. Also, technology-based assessment is not new: It first became relevant in the early 80ies of the last century when computer technology became widely available and also Psychologists began to transfer psychological tests from paper to computer (see also Schroeders, 2010). The first computer labs were established and computer-based testing got more and more refined over the years. Additionally, new assessment tools such as precise measurement of reaction times or adaptive testing were also developed and refined. And while back in the 80ies the introduction of computer-based assessments to Psychology seemed like a technical revolution or even a paradigm shift in psychological assessment, it only took a bit more than a decade until the commercialization of the Internet allowed online-based assessments to be introduced in Psychology (Musch & Reips, 2000). Again with online testing, new technological advances triggered the development of various different new assessment techniques, such as web surveys (Bandilla, 2002), online panels (Göritz, Reinhold, & Batinic, 2002), Internet questionnaires (Gräf, 2002), or online ability tests (Schroeders, Wilhelm, & Schipolowski, 2010; Wilhelm & McKnight, 2002), which were designed, validated, and refined during the years. From there on, online research expanded also to different devices—starting with personal computers in peoples' living rooms, over to pager and other handheld devices and finally to tablets and smartphones—always releasing new cycles of development of new methods, validation, and refinement. To date, smartphone-based assessment (Miller, 2012) is the latest development in technology-based research—facilitating assessment techniques such as mobile sensing methods (Harari et al., 2016), ambulatory assessment (Ebner-Priemer & Trull, 2009), intensive longitudinal designs (Bolger & Laurenceau, 2013), or ecological momentary assessment (Shiffman, Stone, & Hufford, 2008).

Overall, the last decades of equivalence testing suggest that the test medium per se does not influence test scores much. Early on, a meta-analysis on the comparison between paper-based and computer-based cognitive ability tests showed a cross-mode correlation of $r = .97$ (Mead & Drasgow, 1993). Also, later on, comparisons of cognitive ability tests that were delivered either paper-based, computer-based, or on a hand-held device (Schroeders & Wilhelm, 2010, 2011) showed that these tests were largely equivalent across test media. Equally, a meta-analysis comparing self-report questionnaires that were administered either paper-based or computer-based found no score differences between assessment modes (Gnambs & Kaspar, 2017). Consequently, we can assume that the test medium itself is not the most decisive factor when discussing the application of technology-based assessment techniques—regardless of whether notebooks, tablets, smartphones, or future technological devices (that yet have to be developed) are used. Still, it seems reasonable to discuss the comparability of tests across different assessment modes: First, the equivalence of the measure might hinge for example on the specific measure, or sample-specific characteristics (Schroeders & Wilhelm, 2011). For example, robust results on the equivalence of paper-based and online self-report scales might not be readily transferable to ability tests, or what may work for student samples could look different in samples of elderly adults with less exposure to new technological devices. Second, the equivalence of traditional assessment techniques and newly developed approaches should not be the only goal, as a perfect equivalence of these measures also means that the new approach is also only "as good as" the old one. But with advances in assessment methods, we should also strive to *enhance* our assessment techniques rather than finding new ways of achieving what the old techniques were already capable of doing.

In the following parts, I discuss both potential advantages and potential disadvantages of new advances in technology-based assessment. These explications will center around, but will not be limited to, smartphone-based assessment. Rather, I will use the umbrella-term of technology-based assessment, including all assessment approaches that a) use contemporary technological devices (including smartphones, tablets, and other technical devices); b) allow flexible (online) data assessment outside the traditional lab setting; and c) allow the recording of auxiliary data such as log data or other sensory input.

Potential Advantages of Technology-Based Assessment

First, due to its increased flexibility, technology-based assessments offer the possibility to recruit more heterogeneous samples. In contrast, traditional lab-based assessments are often limited to specific groups of participants who live in a particular geographic area and who have time and motivation to take part in psychological assessments, usually subjects from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich, Heine, & Norenzayan, 2010). In psychological research, traditional samples are oftentimes even more limited, as psychology students are usually the easiest group to target. These highly selected samples, however, impede the generalizability of the results. In turn, online studies (Gosling & Mason, 2015) might help to attract a different audience (Gosling, Sandy, John, & Potter, 2010), because these studies are not limited to a certain time or to a specific (university) location. Rather, participants can participate whenever they like, wherever they like. Using appropriate recruitment strategies to target the sample of interest, online studies can facilitate access of larger populations to psychological studies. This way, not only large, nationwide assessments are feasible, but also international research projects are possible (e.g., Dufau et al., 2011). Additionally, smartphone-based assessment technologies can offer even greater flexibility, since smartphones are potentially a valuable assessment tool for psychological research that many people already carry around with them in their daily lives (Miller, 2012).

Second, technology-based assessment cannot only facilitate data collection from large and heterogeneous samples but also allows more flexible item sampling strategies. Using modern assessment tools, alternative methods were introduced using intensive longitudinal designs (Bolger & Laurenceau, 2013), *Ecological Momentary Assessment* (Shiffman et al., 2008), or flexible item sampling approaches such as the *Synthetic Aperture Personality Assessment* (SAPA) approach (Condon & Revelle, 2014; Revelle et al., 2017). These approaches offer new insight into psychological constructs in different ways: On the one hand, using longitudinal studies of participants' behavior in daily life, it is possible to collect information about sequences of events and dynamic processes and to develop idiographic models of human behavior (Wright & Zimmermann, 2019). On the other hand, with a new item sampling approach—such as the SAPA technique—it is possible to collect data on broad and multi-dimensional item samples while keeping effort for participants reasonably low. This is achieved by randomly giving each participant only a small fraction of the original item pool. Although this approach implies data that is *Massively Missing Completely at Random* (Revelle et al., 2017), covariances between scales can be derived from composite scales using covariance algebra. Accordingly, this approach is especially useful if the psychological construct of interest is so broad that it would otherwise require long hours of testing.

Third, technology-based assessments offer the possibility to collect auxiliary data: For example, para data is defined as data that is incidentally collected with almost every technology-based assessment (Couper, 2005), including response times, log data, or mouse clicks. These data can be used to unobtrusively get insights into participants' behavior. But not only para data—that come as a mere by-product of computer- or smartphone-based assessment—can be used to derive behavioral indicators. Using *Mobile Sensing Methods* (Harari, Gosling, Wang, & Campbell, 2015; Harari et al., 2016), other sensory input as well can be combined with traditional psychological measures and para data. Thus, psychological assessment can be augmented with this new type of behavioral data, which helps to complement traditional psychological assessments that usually center around the use of self-reports and measures of cognitive abilities. For example, these additional data can be used to collect additional information about the test setting and the test-taking behavior when psychological assessments are conducted in otherwise uncontrolled settings. For example, response time analyses were used as part of a set of methods to assess bot-generated answers in online surveys (Buchanan & Scofield, 2018), thus identifying invalid cases in online-surveys. Furthermore, mouse clicks can serve as indicators of conflict between choice option or test commitment (Kieslich & Henninger, 2017). Accordingly, technology-based assessment offers the possibility to make new data types accessible for psychological research and complement traditional questionnaire and test data with more direct assessments of human behavior.

Potential Challenges of Technology-Based Assessment

First, technology-based assessment requires elaborate study designs and potentially more preparation than traditional paper-pencil assessments—especially in the fields of data security and ethics, and for the technical realization of smartphone-based studies (Seifert, Hofer, & Allemand, 2018). When using technology-based assessment to discreetly collect data in the background, data protection and the ethical use of data become a central issue in planning such studies. Especially when using mobile sensing methods to extensively collect all sorts of behavioral patterns of an individual, anonymization is difficult to retain and therefore, the implementation of these approaches requires new standards for study ethics and data security (Harari et al., 2016). To this end, a privacy model specifically targeted at the use of mobile data collection was proposed by Beierle and colleagues (2018), including recommendations on transparency measures for the user, data anonymization, and secured data transfer.

Second, technology-based assessment allows the collection of large and more complex data types—including both active data (i.e., self-report measures or tests of cognitive abilities) and passive data (i.e., sensory data or para data; see also Seifert et al., 2018). Accordingly, these complex and extensive data sets fall well within the realm of *Big Data* (Fan, Han, & Liu, 2014), thus requiring more sophisticated statistical approaches for transforming and analyzing such data. Especially when using sensory input, the additional information has to be translated into psychological meaningful variables. In that process, researchers have to monitor the quality of sensory input, the adequacy of data aggregation and transformation techniques, as well as the adequacy of statistical analyses to obtain valid results.

Lastly, with the increased use of online and smartphone-based assessments, we also lose control over the test-setting. Instead of standardized screens and hardware, participants use whatever device they have on hand. Instead of calm and low-stimulus environments, participants might start the assessment while they are on the bus, in a crowded café or in their living room with the TV on. Instead of checking back instructions with the investigator, participants might just quit the test in case something does not work. This lack of standardization might lead to biased results, noisy data, or unwanted dropouts, but especially for unproctored ability assessments, the issue of cheating participants was widely discussed in the literature (Tippins, 2009; Tippins et al., 2006). In contrast, for self-report measures, response distortions are a problem that is independent of the test setting—participants can choose to lie on a personality questionnaire whether a proctor is present or not. In the case of self-report measures, unproctored testing was even discussed to provoke more honest answers because of the anticipated anonymity of the setting. A recent meta-analysis, however, showed that computerized and paper-pencil administered self-reports (Gnambs & Kaspar, 2017) yield comparable mean scores. For ability assessments, proctored testing is still the gold standard against cheating (Rovai, 2000). To date, we do not know much about the extent to which participants cheat or about the conditions that encourage fraudulent behavior even further. First estimates on the prevalence of cheating in high-stakes job recruitment testing surmised cheating rates well below 10% (Lievens & Burke, 2011; Nye, Do, Drasgow, & Fine, 2008) and it was even debated whether cheating is an issue in low-stakes testing at all (Do, 2009). However, in an online knowledge survey, one out of four participants reported cheating in a low-stakes setting that was exclusively for research purposes without any anticipated consequences for overall performance (Jensen & Thomsen, 2014).

Applications for Technology-Based Assessments

In Psychology, one major application field for modern assessment technology centered around the implementation of ambulatory assessment—an umbrella term (Trull & Ebner-Priemer, 2013) for experience sampling methods (Hektner, Schmidt, & Csikszentmihalyi, 2007), Ecological Momentary Assessment (Shiffman et al., 2008), daily diaries (Ellis-Davies, Sakkalou, Fowler, Hilbrink, & Gattis, 2012), or continuous monitoring techniques (Ebner-Priemer & Kubiak, 2007). With the exception of monitoring techniques, which usually collect data via smartphone sensors or wearables, ambulatory techniques primarily use (short) self-report questionnaires to study participants' daily life experiences. Usually, these questionnaires are targeted at recent events, states, or behavior in peoples' every-day life and are administered repeatedly over a pre-defined time period. These real-time assessments are especially beneficial when implementing self-report measures since they may reduce bias due to memory effects (Schwarz, 2012).

But not only self-report instruments can be implemented in a technology-based framework. Cognitive ability measures also benefit from new assessment technology. Usually, the technological implementation of cognitive ability tasks for computers and smartphones is uncomplicated and, compared to traditional paper-pencil assessments, mostly unaffected by mode effects (Schroeders & Wilhelm, 2010, 2011). Additionally, cognitive ability research as well benefits from large data sets from heterogeneous samples (e.g., D. A. Sternberg et al., 2013), insight into intraindividual processes (e.g., Könen, Dirk, & Schmiedek, 2015), or the enrichment with auxiliary data (e.g., Goldhammer, Naumann, & Greiff, 2015), since, to date, the results from research on cognitive abilities are predominantly based on between-person investigations (Schmiedek, Lövdén, von Oertzen, & Lindenberger, 2019) collected in traditional lab settings (D. A. Sternberg et al., 2013).

By expanding technology-based ability assessments, we have the opportunity to, for example, get detailed, longitudinal data about cognitive performance across age groups, with which we could not only describe overall trends but also get insight into intraindividual developmental trajectories of cognitive performance. Alternatively, in a more applied approach, intensive longitudinal data on cognitive performance tasks could be used as indicators for an early onset of neurodegenerative diseases. Another topic, for which considerable knowledge gaps exist, is declarative knowledge. Presumably, these knowledge gaps are due to the fact that declarative knowledge is hard to assess with traditional means, leaving important questions unanswered to date. In the following chapter, I illustrate problems with traditional knowledge assessments using the example of the search for the dimensionality of knowledge, outline potential improvements for knowledge assessments, and present how these improvements could be implemented using a smartphone-based knowledge assessment.

Declarative Knowledge

Declarative knowledge is a substantial facet of crystallized intelligence. However, our knowledge about it is still very limited, presumably because it is difficult to assess using traditional measures. The existing gaps in our understanding of declarative knowledge contrast with its relevance to adult cognitive performance. Therefore, Ackerman (2000, p. 70) used the image of declarative knowledge as the "dark matter of adult intelligence". He implies that declarative knowledge is the substantial part for intellectual performance in adults (in astrophysics, dark matter is hypothesized to account for around 85% of the matter in the universe), but still many open questions revolve around its existence (albeit its ubiquity, dark matter cannot be measured and the nature of dark matter still remains an open question for physical cosmology). He argues that the controversial observation that adult intelligence begins to decline from early adulthood is mainly owed to the fact that with traditional intelligence test batteries—which usually focus on tasks of fluid abilities—a substantial part that constitutes adult intelligence is overlooked: declarative knowledge. This assertion is also mirrored in the historical roots of intelligence theories (Cattell, 1941, 1943; Hebb, 1941), which affirm the duality of fluid and crystallized parts of intelligence. Furthermore, the importance of knowledge and experience becomes also visible in daily life: Presumably, no one would choose a junior doctor over the head surgeon for major heart surgery, although the novice is younger and will therefore outperform the senior physician in common intelligence tasks.

However, when we are trying to assess declarative knowledge, we face substantial difficulties: The measurement of declarative knowledge should be comprehensive, covering broad item samples and domain samples (Ackerman, 1996; Wilhelm & Schroeders, 2019), and yet, there is neither a conclusive conceptual nor a stringent empirical classification of knowledge (Beauducel & Süß, 2011; Rolfhus & Ackerman, 1996), leaving the question of its dimensionality unanswered. Therefore, in the following chapters, I present existing studies on the dimensionality of knowledge, discuss traditional assessments of declarative knowledge, and introduce smartphone-based assessment as a potential alternative to traditional knowledge assessments.

Investigating the Dimensionality of Knowledge

Discrepancies in the results from existing studies on the dimensionality of knowledge start already in the discussion about the mere number of dimensions. Results cover the whole range, starting with studies that state that declarative knowledge is best understood with an overarching general factor (e.g., Wilhelm, Schroeders, & Schipolowski, 2014), to two- (e.g., Hossiep & Schulte, 2008) or four-dimensional models (e.g., Rolfhus & Ackerman, 1999), up to six-dimensional models (e.g., Irwing, Cammock, & Lynn, 2001). One potential reason for these diverging results might lay in the diverse item pools underlying the different studies. Without even focusing on particular thematic focus of these studies, Figure 2 depicts a) the number of domains, b) the number of items per domain, and c) the number of participants—showing that there is considerable variability in the number of domains employed, as well as the number of participants assessed.

Looking at Figure 2, three issues concerning the underlying item samples become apparent: First, Figure 2 shows that all but two studies (Ackerman, 2000; Rolfhus & Ackerman, 1999) only use a very limited number of items per domain, failing to depict the breadth of a given domain. This is especially a problem since with only a limited number of items for a domain, domain content is very likely biased—moving away from the original idea of item sampling in the sense of items being drawn randomly from a universe of items, towards mere expert selection (Loevinger, 1965). The importance of elaborate item sampling strategies was demonstrated by Schroeders, Wilhelm, and Olaru (2016), who illustrated that the well-replicated finding of sex differences in general knowledge (usually finding a male advantage) merely hinges on the item sets analyzed. The effect could not only be extinguished based on the items selected, but it could also be fully reversed, thus demonstrating the need for larger, well-balanced and properly selected test compilations.

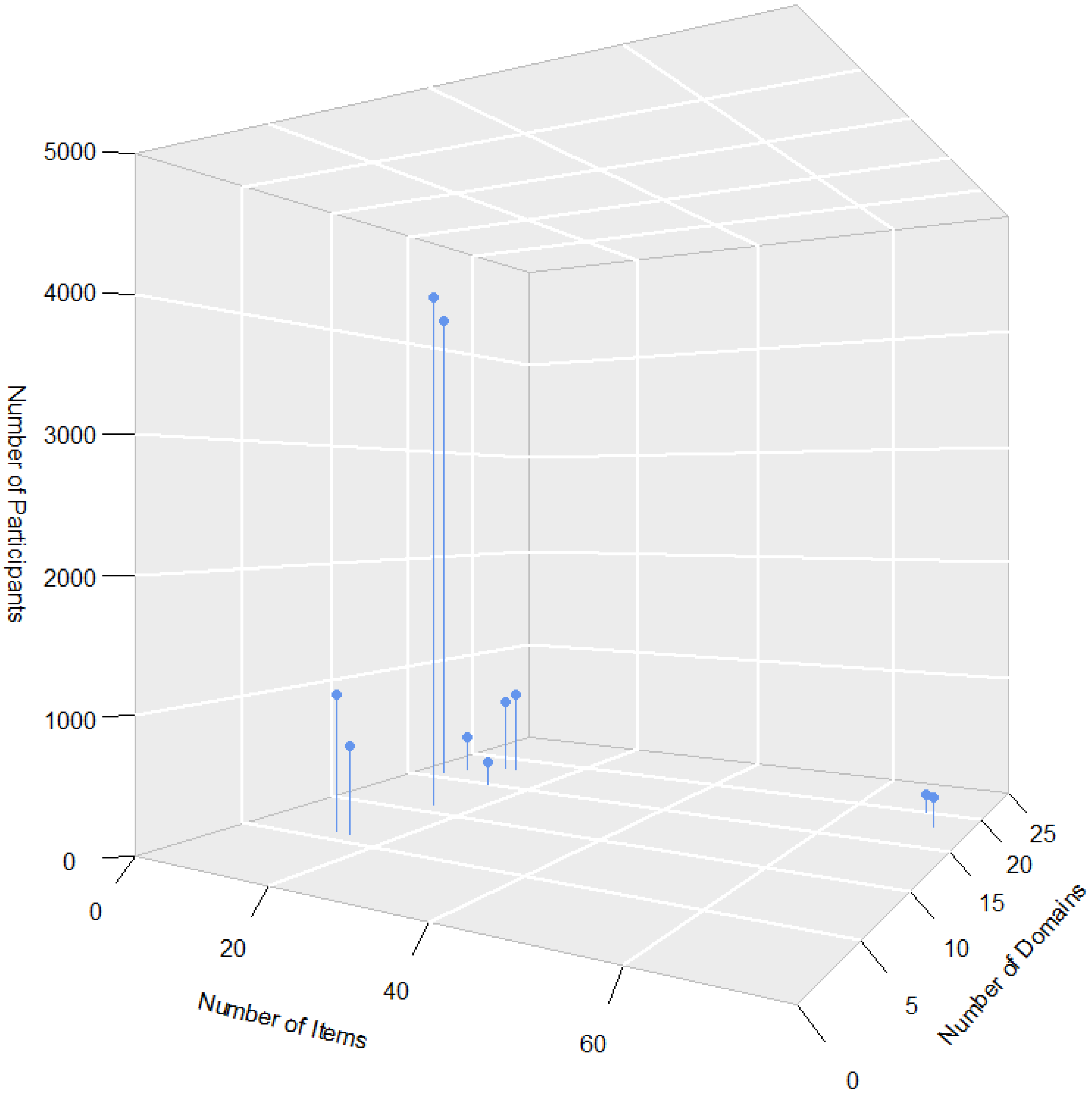

Figure I-2. Studies on the dimensionality of knowledge.

Second, although there are only a few studies that test a larger amount of items per domain, Figure 2 suggests that using traditional lab assessment, a trade-off exists: Either one tests a large sample (N > 4000; see for example Hossiep & Schulte, 2008; Wilhelm et al., 2014) with comparatively few items, or one tests large amounts of items, but with a comparatively smaller sample (N < 250, see for example Ackerman, 2000; Rolfhus & Ackerman, 1999). Certainly, this is a problem of traditional lab assessments, in which practical considerations (e.g., feasibility and reasonableness) usually put constraints to the test design.

Third, looking at the number of domains for each study, it becomes apparent that these studies base their findings on very heterogeneous domain selections, ranging from six domains (e.g., Amthauer, Brocke, Liepmann, & Beauducel, 2001) up to 20 domains (Rolfhus & Ackerman, 1999). It is not surprising that studies with smaller–and probably more general–domain samples might find different factor structures than those that test a broad range of different knowledge domains.

However, the item sampling issues as depicted in Figure 2 are only one aspect of the general heterogeneity between studies: Additionally, studies differ in the considered age range, varying from very narrow samples including only high school students (e.g., Wilhelm et al., 2014) or college freshmen (e.g., Ackerman, Bowen, Beier, & Kanfer, 2001) up to broad adult samples (e.g., Ackerman, 2000). In the same vein, studies also differ in the considered academic background, again from very narrow up to very heterogeneous samples. Lastly, also the measures themselves and assessment designs differ considerably: For example, to realize a broad knowledge assessment covering both a broad item and domain sample, Rolfhus and Ackerman (1999) used a power design. Items were ordered by difficulty within each domain and participants were presented knowledge questions in each domain until they answered three questions incorrectly in a row. However, the design might influence the results, since difficulty estimates might fluctuate across samples and the assessment assumes unidimensionality within a domain, which often is a research question of the study rather than a given fact to start with. Taken together, existing issues in the assessment of declarative knowledge have largely contributed to the discrepancies between studies on the dimensionality of knowledge.

Measuring Declarative Knowledge

The example of studies on the dimensionality of declarative knowledge illustrates issues in the measurement of declarative knowledge. As outlined above, potential reasons for the diverse results include a) differences in breadth and scope of item samples, b) differences in the person samples, especially regarding age and educational background, and c) differences in assessment methods. However, these issues do not arise because past researchers have not paid enough attention to measurement quality. These issues are owed to restrictions that come from traditional knowledge assessments—which are usually conducted in traditional lab settings, with all limitations these settings entail: First, these assessments are limited with regard to test length and test time due to pragmatic concerns. Therefore, we cannot simply test as many knowledge items and domains as we like, limiting the scope of the item sample under investigation. Unfortunately, the smaller the item sample gets, the more prone to bias it is, especially if it was constructed using expert selection (Schroeders et al., 2016). Second, using traditional assessment approaches, we usually only target a very specific person sample. In psychological research, most commonly this sample comprises mostly psychology students that receive course credits for their participation. But even in cases when monetary rewards are offered, the sample is limited to people that have the time and motivation to come to a university lab, usually during weekdays—in the worst case resulting in samples covering students from more diverse disciplines. However, especially in the case of declarative knowledge, results are likely to depend largely on participants' experiences and their educational and professional background—a very unfortunate circumstance when highly selected and homogeneous student samples are the easiest to target.

But how does an ideal assessment of declarative knowledge look like? An ideal assessment of declarative knowledge should certainly cover the breadth and depth of (culturally valued) knowledge—including all areas that an individual could possibly acquire during his life (Ackerman, 1996; Wilhelm & Schroeders, 2019). As this is entirely impossible to achieve, the assessment should at least be based on an item pool that is as large and comprehensive as possible, spanning the whole continuum from general to domain-specific knowledge (Ackerman, 1996, 2000). Furthermore, it should cover both occupational as well as vocational knowledge (Ackerman, 1996)—as opposed to traditional test batteries that were criticized to mainly focus on knowledge acquired during school. The limitation to school knowledge only could lead to biases when assessing adult individuals, as this knowledge usually becomes less relevant for peoples' everyday lives over time and thus fades over time (as described with the term *historical crystallized intelligence* by Cattell, 1971).

Accordingly, an ideal assessment of declarative knowledge should also be kept up-to-date, since clearly the relevance of certain knowledge is likely to change over time. Already the ancient Greeks had ideas about what important knowledge domains are and what general knowledge should include, as for example documented in writings about the *septem artes liberales* (seven liberal arts) or Aristotle's classification of scientific disciplines (Samurin, 1977) into the theoretical subjects, whose goal is pure knowledge development (*episteme*; e.g., analytics, physics, or mathematics), the practical subjects (*techne*; e.g., ethics, economics, or politics), and creative subjects (*poiesis*; e.g., poetics, rhetorics, or arts). Today's ideas about relevant knowledge are surely different from these days. Due to the fast scientific progress and increase in knowledge, it is very likely that a questionnaire that was up-to-date 20 years ago might be completely outdated right now.

A Smartphone-based Assessment of Declarative Knowledge

Against the background of problems associated with traditional assessments of declarative knowledge, I present a smartphone-based approach that yields the potential to address problems of previous studies by a) testing a broad item and domain sample, and b) addressing a broad and heterogeneous person sample, while c) keeping effort for participants reasonable. In the following, I describe the development of a mobile quiz game that was implemented as part of the *IQ App Ulm* (www.iq-app.de).

First, to address the issue of item and domain sampling, we conducted an extensive literature review of existing knowledge test batteries (Amthauer et al., 2001; Hossiep & Schulte, 2008; Irwing et al., 2001; Mariani, Sacco, Spinnler, & Venneri, 2002; Roberts et al., 2000; Schrank, Mather, & McGrew, 2014; Wilhelm et al., 2014), empirical classifications (Engelberg, 2015; Rolfhus & Ackerman, 1996, 1999), the courses in German universities, and various vocational profiles. This search resulted in a list of 34 knowledge domains, covering a wide range of vocational and occupational knowledge. Additionally, we included a measure of current events knowledge from the past two years as well as an additional scale of items that are near to impossible to solve as a potential indicator for cheating behavior. Each domain covered a minimum of 100 items, resulting in a set of 4,054 items total (see Table 1 for a description of knowledge domains and respective number of items per domain). All knowledge items were designed as multiple-choice items with either verbal, figural, numerical, or auditive content. Additionally, item content was checked by an expert in the respective domain and item wording, grammar, and psychometric quality were checked by two independent raters with a background in psychology.

Table I-1. Description of Knowledge Domains

Domain	Covers knowledge about...	*N* Items
Anthropology	diverse cultural groups and their traditions and customs	110
Architecture	architectural style epochs, building types, and building techniques, as well as important architects	100
Arts	national and international artists and their pieces, artistic styles and techniques, as well as famous museums	100
Biology	a broad range of biology, at cellular, organismal, and ecological levels	103
Celebrities	famous people of the modern era, including music and movie stars, influencers and royal families	107
Chemistry	general chemistry, as well as organic and inorganic chemistry	104
Computer science	computers, software, operating systems, file types, and programming languages	101
Ecology	environmental protection and pollution and related technologies	102
Economics	macroeconomical and microeconomical theories, business and management	107
Education	education and basic concepts of educational science, developmental psychology, schooling, and youth protection laws	100
Fashion	clothing styles, types of garment and fabrics, as well as beauty products, hairstyles, and make-up	103
Finances	currencies and transactions, as well as accountancy and investments	103

(continued)

Table I-1. Description of Knowledge Domains *(continued)*

Domain	Covers knowledge about...	*N* Items
Geography	location of countries and cities, regions, mountains and rivers, as well as geology	117
Health	how to stay and age healthy, fitness, home remedies for minor illnesses, and first aid	100
History	German, European, and Western history from antiquity to the modern era	104
Housekeeping	home economics, cleaning and cleaning supplies, gardening, and hand tools	105
Law	legal theories, national and international laws and institutions	107
Linguistics	linguistic and grammatical terms, linguistic history and linguistic families	101
Literature	national and international authors and books, literary forms and literary history	151
Mathematics	basic mathematical concepts and important mathematicians, as well as concepts of algebra, analysis, and geometry	100
Medicine	diseases and their treatment	106
Movies & TV	national and international movies and TV productions, actors and directors	123
Music	national and international composer and their pieces, instruments, and musical theory	100
Nutrition	foods and their preparation, nutrients, and special diets	104
Philosophy	philosophers and philosophical theories from antiquity to the modern era	101
Physics	basics and applications of physical laws and units as well as mechanics, optics, acoustics, thermodynamics, and astronomy	110
Politics	political systems and forms of government, important politicians, parties and political events	105
Pop culture	trends of youth culture, recent scandals and stories, as well as leisure activities and games	109
Pop music	national and international artists, bands, and their pieces, as well as music genres from the 1960ies onwards	118
Psychology	basic concepts of Psychology, psychotherapeutic schools, and mental illnesses	115
Religion	major religious groups and their practices	112
Sports	different disciplines, athletes, and sportive events	116
Statistics	descriptive and inferential statistics, principles of hypothesis testing, and stochastics	101
Technology	automotive technology, aerospace technology, bionics, information and communication technology, and home electronics	107
Current events	events from the years 2014 and 2015	300
Difficult items	unusual facts from all 34 domains	102

Second, to address the issue of person samples, we tried to target a broad audience. We recruited participants using a broad range of different advertising platforms, including advertising flyers and posters, magazine articles, radio interviews, online forums, and Facebook groups. If interested, anyone could download the app via google Playstore (for android) and AppStore (for iOS). After completing the download, participants received an email with information about the study and a consent form. After consenting, participants could play the app wherever and whenever they liked, without any restrictions concerning duration, daytime, test setting, or internet connection.

Lastly, to address the issue of participant motivation, we implemented an item sampling approach similar to the SAPA technique (Condon & Revelle, 2014): Questions were presented in sets of 27 items per round. These sets always covered questions from 9 domains with 3 randomly drawn items from the item pool. Since the items were randomly selected, our analyses are unaffected by the exact amount of rounds a participant decides to play. Furthermore, following the logic of the SAPA technique, our analyses also did not require participants to answer all existing knowledge items, making it possible that participants answer as many questions as they feel motivated to. Nevertheless, to increase participant motivation, participants could also collect badges for good performance as well as regular use of the app (see also Hamari, 2017). Additionally, participants received a detailed report on their performance in all 34 knowledge domains based on the questions they answered. The report was updated after every round they played and could be directly accessed via the app.

Taken together, smartphone-based assessment provides many opportunities to meet the problems in the measurement of declarative knowledge in particular, and of cognitive abilities in general. But are we also generating new problems? What happens if we transfer cognitive ability tests to unstandardized, unproctored settings? How will setting or mode effects influence participants' performance? Will participants try to boost their scores by cheating? And what measures can be taken to secure data quality—*a priori* when choosing an appropriate study design, or *post hoc* by detecting aberrant responses in the data?

Overview of the Dissertation Manuscripts

Technology-based assessment offers many advantages for the measurement of cognitive abilities—for example by offering the possibility to investigate broad item samples in heterogeneous person samples. However, these advantages might also come at a cost: The data retrieved using these methods is more complex—calling for more advanced data processing and analysis techniques—and probably also noisier due to the changed setting. To this end, I present three manuscripts that center around the question of how we can implement technology-based assessments to address open questions in the assessment of declarative knowledge. In the following, I will outline the research questions and methods applied in each manuscript of this dissertation to examine both challenges and opportunities of smartphone-based assessment of declarative knowledge.

Manuscript 1: A Meta-Analysis of Test Scores in Proctored and Unproctored Ability Assessments

The first paper examines the question to which extent online ability assessments in general might be hampered by biases that arise from the lack of a standardized, proctored test environment. While online assessments—an overarching term under which we also count smartphone-based approaches—are widely accepted as a flexible and efficient way to collect data from samples that are otherwise hard to target, this flexibility might also come at the cost of reduced control over the test environment, which ultimately might inveigle participants to cheat to boost their test scores. Despite being verbose, the body of literature on possible biases of online ability assessment was inconclusive. To this end, we conducted a meta-analysis to evaluate the extent to which online assessment might be generally biased and examined possible factors that might moderate this. This manuscript examines both score differences between online and lab ability assessments, as well as a rank order changes for a small subsample of studies. Furthermore, it discusses to which extent test context (i.e., low- vs. high-stakes), countermeasures taken against cheating, test modality (i.e., paper-pencil vs. computer-based tests), and features of the measure itself (i.e., low vs. high *searchability*) influence the results.

Manuscript 2: On the Dimensionality of Crystallized Intelligence: A Smartphone-based Assessment

Manuscript 2 reports the implementation of a mobile quiz app designed to assess knowledge across a broad range of content domains in order to examine the dimensionality of crystallized intelligence. To date, most studies failed to assess the breadth and depth of declarative knowledge across a demographically diverse sample and accordingly, previous studies on the dimensionality of knowledge show diverse results. Using a smartphone app, we evade the need to invite participants to long and tedious on-site test sessions, but rather allow participants to use the app whenever and wherever they want. Not being limited to a specific test location or time might also attract parts of the population that usually do not make it to the lab. We use the so-called "bass-ackwards method" (Goldberg, 2006)—a hierarchical series of principal components analyses—to illustrate the unfolding factor structure of crystallized intelligence. We compare the arising factor structure with previous results on the dimensionality of crystallized intelligence.

Manuscript 3: Caught in the Act: Predicting Cheating in Unproctored Knowledge Assessment

Finally, picking up on the considerable inconvenience of cheating on unproctored ability tests, the third manuscript examines different types of data and their potential to identify cheaters in unproctored knowledge assessment. Cheating is seemingly hard to prevent in the first place, thus with the present study, we focus on approaches to detect cheating to secure or restore data quality. As cheating can be considered a threat particularly in tests that are easy to look up on the Internet, we juxtapose results from both an unproctored online knowledge test on the one hand and a proctored lab knowledge test on the other hand. More specifically, we evaluate the potential of a) questionnaire data, b) test data, and c) para data to predict cheating. Based on our findings, we give recommendations on tailored approaches of how to assess and secure data quality in unproctored ability assessment.

In the following chapters, I will present all three manuscripts and summarize major findings in the epilogue. I will also link them to existing research in psychological assessment and provide suggestions for further research that tie in with the results presented in the present work.

References

Abbate, S., Avvenuti, M., Bonatesta, F., Cola, G., Corsini, P., & Vecchio, A. (2012). A smartphone-based fall detection system. *Pervasive and Mobile Computing, 8*, 883–899. https://doi.org/10.1016/j.pmcj.2012.08.003

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227–257. https://doi.org/10.1016/S0160-2896(96)90016-1

Ackerman, P. L. (2000). Domain-specific knowledge as the "dark matter" of adult intelligence: Gf/gc, personality and interest correlates. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 55*, 69–84. https://doi.org/10.1093/geronb/55.2.P69

Ackerman, P. L., Bowen, K. R., Beier, M., & Kanfer, R. (2001). Determinants of individual differences and gender differences in knowledge. *Journal of Educational Psychology, 93*, 797–825. https://doi.org/10.1037//0022-0663.93.4.797

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R Manual [Manual of the Intelligence Structure Test 2000 R]*. Göttingen: Hogrefe.

Bandilla, W. (2002). Web surveys – An appropriate mode of data collection for the social sciences. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online Social Sciences*, 1–6.

Bartlett, F. C. (1937). Cambridge, England: 1887-1937. *The American Journal of Psychology, 50*, 97–110. https://doi.org/10.2307/1416623

Beauducel, A., & Süß, H.-M. (2011). Wissensdiagnostik: Allgemeine und spezielle Wissenstests [Knowledge assessment: General and specific knowledge tests]. In L. F. Hornke, M. Amelang, & M. Kersting (Eds.), *Serie II, Psychologische Diagnostik: Vol. Methodologie und Methoden* (pp. 235–273). Göttingen:

Hogrefe.

Beierle, F., Tran, V. T., Allemand, M., Neff, P., Schlee, W., Probst, T., ... Zimmermann, J. (2018). Context data categories and privacy model for mobile data collection apps. *Procedia Computer Science, 134*, 18–25. https://doi.org/10.1016/j.procs.2018.07.139

Berry, A. (2003). Whenever you can, count. *London Review of Books, 25*, 23–25.

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods : An introduction to diary and experience sampling research.* New York: Guilford Press.

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

Campolo, C., Iera, A., Molinaro, A., Paratore, S. Y., & Ruggeri, G. (2012). SMaRTCaR: An integrated smartphone-based platform to support traffic management applications. *2012 First International Workshop on Vehicular Traffic Management for Smart Cities (VTM)*, 1–6. https://doi.org/10.1109/VTM.2012.6398700

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing [Abstract]. *Psychological Bulletin, 38*, 592.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*, 153–193. http://dx.doi.org/10.1037/h0059973

Cattell, R. B. (1971). *Abilities: Their Structure, Growth, and Action.* Boston: Houghton Mifflin.

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*, 486–501. https://doi.org/10.1177/0894439305278972

Do, B.-R. (2009). Research on unproctored internet testing. *Industrial and*

Organizational Psychology, 2, 49–51. https://doi.org/10.1111/j.1754-9434.2008.01107.x

Dufau, S., Duõabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., ... Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLoS ONE, 6*, 1–3. https://doi.org/10.1371/journal.pone.0024974

Ebner-Priemer, U. W., & Kubiak, T. (2007). Psychological and Psychophysiological Ambulatory Monitoring. *European Journal of Psychological Assessment, 23*, 214–226. https://doi.org/10.1027/1015-5759.23.4.214

Ebner-Priemer, U. W., & Trull, T. J. (2009). Ambulatory assessment: An innovative and promising approach for clinical psychology. *European Psychologist, 14*, 109–119. https://doi.org/10.1027/1016-9040.14.2.109

Ellis-Davies, K., Sakkalou, E., Fowler, N. C., Hilbrink, E. E., & Gattis, M. (2012). CUE: The continuous unified electronic diary method. *Behavior Research Methods, 44*, 1063–1078. https://doi.org/10.3758/s13428-012-0205-1

Engelberg, P. M. (2015). *Ursachen fÃ¼r Geschlechterdifferenzen in Tests des Allgemeinen Wissens [Causes for gender differences in general knowledge tests]* (Doctoral Dissertation). University of Wuppertal, Wuppertal, Germany.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review, 1*, 293–314. https://doi.org/10.1093/nsr/nwt032

Galton, F. (1883). *Inquiries into the human faculty and its development.* London: Macmillan.

Galton, F. (1887a). *A descriptive list of anthropometric apparatus.* Cambridge: Cambridge Scientific Instrument Company.

Galton, F. (1887b). On recent designs for anthropometric instruments. *Journal of the Anthropological Institute, 16*, 2–8. https://doi.org/ 10.2307/2841732

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

Goldberg, L. R. (2006). Doing it all bass-ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality, 40*, 347–358. https://doi.org/10.1016/j.jrp.2006.01.001

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence, 3*, 21–40. https://doi.org/10.3390/jintelligence3010021

Göritz, A. S., Reinhold, N., & Batinic, B. (2002). Online Panels. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online Social Sciences* (pp. 29–52). Göttingen: Hogrefe & Huber.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences, 33*, 94–95. https://doi.org/10.1017/S0140525X10000300

Gräf, L. (2002). Assessing Internet questionnaires: The online pretest lab. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online Social Sciences* (pp. 73–94). Göttingen: Hogrefe & Huber.

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science, 26*, 10–15. https://doi.org/10.1177/0963721416666518

Hamari, J. (2017). Do badges increase user activity? A field experiment on the effects of gamification. *Computers in Human Behavior, 71*, 469–478. https://doi.org/10.1016/j.chb.2015.03.036

Harari, G. M., Gosling, S. D., Wang, R., & Campbell, A. T. (2015). Capturing situational information with smartphones and mobile sensing methods: Capturing situations with smartphone sensing. *European Journal of Personality, 29*, 509–511. https://doi.org/10.1002/per.2032

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*, 838–854. https://doi.org/10.1177/1745691616650285

Hebb, D. O. (1941). Clinical evidence concerning the nature of normal adult test performance [Abstract]. *Psychological Bulletin, 38*, 593.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method.* Thousand Oaks: Sage.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–83. https://doi.org/10.1017/S0140525X0999152X

Hossiep, R., & Schulte, M. (2008). *BOWIT - Bochumer Wissenstest [BOWIT - Bochum Knowledge Test].* Göttingen: Hogrefe.

Irwing, P., Cammock, T., & Lynn, R. (2001). Some evidence for the existence of a general factor of semantic memory and its components. *Personality and Individual Differences, 30*, 857–871. https://doi.org/10.1016/S0191-8869(00)00078-7

Jensen, C., & Thomsen, J. P. F. (2014). Self-reported cheating in web surveys on political knowledge. *Quality & Quantity, 48*, 3343–3354. https://doi.org/10.1007/s11135-013-9960-z

Johnson, D. A., & Trivedi, M. M. (2011). Driving style recognition using a smartphone as a sensor platform. *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 1609–1615. https://doi.org/10.1109/ITSC.2011.6083078

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

Kim, Y., Hyojeong Shin, & Cha, H. (2012). Smartphone-based Wi-Fi pedestrian-tracking system tolerating the RSS variance problem. *2012 IEEE International Conference on Pervasive Computing and Communications*, 11–19. https://doi.org/10.1109/PerCom.2012.6199844

Könen, T., Dirk, J., & Schmiedek, F. (2015). Cognitive benefits of last night's sleep: Daily variations in children's sleep behavior are related to working memory fluctuations. *Journal of Child Psychology and Psychiatry, 56*, 171–182. https://doi.org/10.1111/jcpp.12296

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

Kukulska-Hulme, A., & Viberg, O. (2018). Mobile collaborative language learning: State of the art: Mobile collaborative language learning. *British Journal of Educational Technology, 49*, 207–218. https://doi.org/10.1111/bjet.12580

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*, 817–824. https://doi.org/10.1348/096317910X522672

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review, 72*, 143–155. https://doi.org/10.1037/h0021704

Mariani, C., Sacco, L., Spinnler, H., & Venneri, A. (2002). General Knowledge of the World: A standardised assessment. Neurological Sciences, 23, 161–175. https://doi.org/10.1007/s100720200057

Marshall, A., Medvedev, O., & Antonov, A. (2008). Use of a smartphone for improved self-management of pulmonary rehabilitation. *International Journal of Telemedicine and Applications, 2008*, 1–5. https://doi.org/10.1155/2008/753064

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological*

Bulletin, 114, 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Mehl, M. R. (2017). The electronically activated recorder (EAR): A method for the naturalistic observation of daily social behavior. *Current Directions in Psychological Science, 26*, 184–190. https://doi.org/10.1177/0963721416680611

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7*, 221–237. https://doi.org/10.1177/1745691612441215

Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. In M. H. Birnbaum (Ed.), *Psychological Experiments on the Internet* (pp. 61–87). https://doi.org/10.1016/B978-012099980-4/50004-6

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*, 24, 321–332. https://doi.org/10.3109/09638237.2015.1019054

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*, 112–120. https://doi.org/10.1080/09639284.2011.590012

Pearson, K. (1914). *The Life, Letters and Labours of Francis Galton.* London: Cambridge University Press.

Rauthmann, J. F., Horstmann, K. T., & Sherman, R. A. (2019). Do self-reported traits and aggregated states capture the same thing? A nomological perspective on trait-state homomorphy. *Social Psychological and Personality Science, 10*, 596–611. https://doi.org/10.1177/1948550618774772

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2017). Web and phone-based data collection using planned missing designs. In N. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE Handbook of Online Research Methods.* Los Angeles, CA: SAGE.

Roberts, R. D., Goff, G. N., Anjoul, F., Kyllonen, P. C., Pallier, G., & Stankov, L. (2000). The Armed Services Vocational Aptitude Battery (ASVAB): Little more than acculturated learning (Gc)!? *Learning and Individual Differences,* *12*, 81–103. https://doi.org/10.1016/S1041-6080(00)00035-2

Rolfhus, E. L., & Ackerman, P. L. (1996). Self-report knowledge: At the crossroads of ability, interest, and personality. *Journal of Educational Psychology, 88*, 174–188. http://dx.doi.org/10.1037/0022-0663.88.1.174

Rolfhus, E. L., & Ackerman, P. L. (1999). Assessing individual differences in knowledge: Knowledge, intelligence and related traits. *Journal of Educational Psychology, 91*, 511–526. http://dx.doi.org/10.1037/0022-0663.91.3.511

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

Samurin, E. I. (1977). *Geschichte der bibliothekarisch-bibliographischen Klassifikation [History of library-bibliographic classification]*. München: Verlag Dokumentation.

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2015). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence, 46*, 156–168. https://doi.org/10.1016/j.intell.2014.05.014

Schmiedek, F., Lövdén, M., von Oertzen, T., & Lindenberger, U. (2019). *Within-person structures of daily cognitive performance cannot be inferred from between-person structures of cognitive abilities* [Preprint]. https://doi.org/10.7287/peerj.preprints.27576v1

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV Tests of Achievement*. Rolling Meadows, IL: Riverside.

Schroeders, U. (2010). *Measurement of cognitive abilities using modern technologies: Artifacts, equivalence, and new constructs* (Doctoral dissertation). Humboldt-Universität zu Berlin, Berlin.

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment, 26*, 284–292.

https://doi.org/10.1027/1015-5759/a000038

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*, 849–869. https://doi.org/10.1177/0013164410391468

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). The influence of item sampling on sex differences in knowledge tests. *Intelligence, 58*, 22–32. https://doi.org/10.1016/j.intell.2016.06.003

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). Washington, D.C.: American Psychological Association.

Schwarz, N. (2012). Why researchers should think "real-time": A cognitive rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of Research Methods for Studying Daily Life.* New York, NY: Guilford Press.

Seifert, A., Hofer, M., & Allemand, M. (2018). Mobile data collection: Smart, but not (yet) smart enough. *Frontiers in Neuroscience, 12*, 971. https://doi.org/10.3389/fnins.2018.00971

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Sokal, M. M. (1972). Psychology at Victorian Cambridge: The unofficial laboratory of 1887-1888. *Proceedings of the American Philosophical Society, 116*, 145–147.

Sternberg, D. A., Ballard, K., Hardy, J. L., Katz, B., Doraiswamy, P. M., & Scanlon, M. (2013). The largest human cognitive performance dataset reveals insights into the effects of lifestyle factors and aging. *Frontiers in Human Neuroscience, 7*, 1–10. https://doi.org/10.3389/fnhum.2013.00292

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology, 2*, 2–10. https://doi.org/10.1111/j.1754-9434.2008.01097.x

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*, 189–225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

Tran, J., Tran, R., & White, J. R. (2012). Smartphone-based glucose monitors and applications in the management of diabetes: An overview of 10 salient "Apps" and a novel smartphone-connected blood glucose monitor. *Clinical Diabetes, 30*, 173–178. https://doi.org/10.2337/diaclin.30.4.173

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9*, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

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). Seattle: Hogrefe & Huber.

Wilhelm, O., & Schroeders, U. (2019). Intelligence. In R. J. Sternberg & J. Funke (Eds.), *The Psychology of Human Thought* (pp. 255–275). Retrieved from https://books.ub.uni-heidelberg.de/index.php/heiup/catalog/book/470

Wilhelm, O., Schroeders, U., & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz fÃ¼r die 8. Bis 10. Jahrgangsstufe (BEFKI 8-10) [Berlin Test of fluid and crystallized intelligence for grades 8–10]*. Göttingen: Hogrefe.

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

Wrzus, C., Wagner, G. G., & Riediger, M. (2014). Feeling good when sleeping in? Day-to-day associations between sleep duration and affective well-being differ from youth to old age. *Emotion, 14*, 624–628. https://doi.org/10.1037/a0035349

Yavuz, G. R., Kocak, M. E., Ergun, G., Alemdar, H., Yalcin, H., Incel, O. D., ...

Ersoy, C. (2010). A smartphone-based fall detector with online location support. *Proceedings of Phone Sense*, 31–35.

II. Manuscript 1

A Meta-Analysis of Test Scores in Proctored and Unproctored Ability Assessment

Diana Steger

Ulm University, Germany

Ulrich Schroeders

University of Kassel, Germany

Timo Gnambs

Leibnitz Institute for Educational Trajectories, Germany

Johannes Kepler University Linz, Austria

Status: Published

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*, 174–184. https://doi.org//10.1027/1015-5759/a000494

Abstract

Unproctored, web-based assessments are frequently compromised by a lack of control over the participants' test taking behavior. It is likely that participants cheat if personal consequences are high. This meta-analysis summarizes findings on context effects in unproctored and proctored ability assessments and examines mean score differences and correlations between both assessment contexts. As potential moderators, we consider (a) the perceived consequences of the assessment, (b) countermeasures against cheating, (c) the susceptibility to cheating of the measure itself, and (d) the use of different test media. For standardized mean differences, a three-level random-effects meta-analysis based on 108 effect sizes from 49 studies (total N = 100,434) identified a pooled effect of $\Delta = 0.20$, 95% CI [0.10, 0.31], indicating higher scores in unproctored assessments. Moderator analyses revealed significantly smaller effects for measures that are difficult to research on the Internet. These results demonstrate that unproctored ability assessments are biased by cheating. Unproctored assessments may be most suitable for tasks that are difficult to search on the Internet.

Keywords: meta-analysis, unproctored assessment, cognitive ability, cheating

Introduction

Recent technological developments changed the way researchers collect psychological data in general (Miller, 2012) and conduct psychological assessments in particular (Harari et al., 2016). Gathering data outside the laboratory in an unproctored setting, for example, using mobile devices or web-based tests serves as an ecologically valid (Fahrenberg, Myrtek, Pawlik, & Perrez, 2007) and economic method (Buhrmester, Kwang, & Gosling, 2011) to collect psychological data on large, heterogeneous samples (Gosling, Sandy, John, & Potter, 2010). Therefore, unproctored, web-based testing has become the dominant assessment mode in market and public opinion research (Evans & Mathur, 2005) and is similar popular in the academic realm (Allen & Seaman, 2014) or in personnel selection (Lievens & Harris, 2003; Tippins, 2011). The advantages of unproctored testing, however, come at a cost: the lack of supervision results in less standardized test taking conditions and less control over test-takers' behavior (Wilhelm & McKnight, 2002). Therefore, the question arises if the opportunity for dishonest behaviors in unproctored assessments leads to biased scores and threatens the usefulness of these tests (Rovai, 2000; Tippins et al., 2006). To this end, a meta-analysis is presented that compares scores from proctored and unproctored ability tests across assessment contexts and examines potential moderating influences thereon.

Mode Effects in Ability Assessments

While scores of self-report instruments can be considered equivalent for proctored and unproctored testing (Gnambs & Kaspar, 2017), results for tests of maximal performance are rather inconclusive (Do, 2009): Some studies found no systematical differences between self-selected web samples and traditional lab samples (e.g., Ihme et al., 2009), whereas others reported significantly higher scores for unproctored tests (e.g., Carstairs & Myors, 2009) or, occasionally, for proctored tests (e.g., Coyne, Warszta, Beadle, & Sheehan, 2005). Inconsistent results were also reported for the prevalence of cheating: Some studies found low cheating rates varying from below 2.5% (Nye, Do, Drasgow, & Fine, 2008) to 7.0% (Tendeiro, Meijer, Schakel, & Maij-de Meij, 2013). Conversely, in an online survey, every fourth participant reported cheating on knowledge task without being offered performance-dependent incentives (Jensen & Thomsen, 2014).

One reason for the heterogeneous results are the varying settings that unproctored assessments were administered in (Reynolds, Wasko, Sinar, Raymark, & Jones, 2009), such as personnel selection (Bartram, 2006; Tippins, 2009), educational contexts (Allen & Seaman, 2014), and research contexts, in which the feasibility, equivalence, and validity of web-based assessments are examined (e.g., Jensen & Thomsen, 2014; Wilhelm & McKnight, 2002). These settings differ in the perceived consequences of assessment, the countermeasures that are taken to prevent cheating, and the measured cognitive domain. In industrial and organizational (I/O) psychology, ability testing often takes place in high-stakes settings with hiring decisions linked to the individual test results. Thus, test-takers have a strong motivation to perform well to increase their chances of employment. To maximize the benefits for applicants and employers (Gibby, Ispas, Mccloy, & Biga, 2009), countermeasures against cheating are implemented to discourage participants from faking their test scores in recruitment procedures. In educational assessments, online placement tests or exams are commonly knowledge tests that are tailored to the curriculum. In research settings, however, test-takers' performance in unproctored assessments usually have no severe consequences, thus, participants are expected to cheat less (Do, 2009). In contrast to the applied contexts, a wide range of different measures are examined, such as reasoning tests (e.g., Preckel & Thiemann, 2003), perception tasks (e.g., Williamson, Williamson, & Hinze, 2016), and knowledge tests (e.g., Jensen & Thomsen, 2014). Accordingly, the current meta-analysis investigates whether there are systematic score differences in proctored and unproctored ability assessments depending on the aforementioned differences in the test environment.

Research Questions

The aim of this meta-analysis was to investigate to what extent a lack of supervision undermines psychological assessment of cognitive abilities. Given that unproctored assessment procedures are on the rise (Gosling & Mason, 2015), it is crucial to know whether the mode of test administration influences test scores. Our outcome variables are standardized mean differences and correlations between proctored and unproctored ability assessments. We take into account all test situations without a human supervisor present (Tippins, 2009). Accordingly, a setting is proctored if a human supervisor is present or remotely proctored if the testing is supervised via web-cam. Additionally, this meta-analysis considers various moderators to explain the heterogeneous findings reported in the literature.

First, test-takers' cheating motivation can be influenced by the perceived consequences of a test result. If participants anticipate severe consequences such as hiring or university admission, they are most likely more motivated to cheat. Therefore, proctored assessments are still viewed as the gold standard in high-stakes testing (Rovai, 2000). Do (2009) hypothesized that cheating is not as prevalent in low-stakes contexts, even though previous results point in a different direction (Jensen & Thomsen, 2014). We expect that in case important consequences are directly linked to the participant's performance, test-takers might be more likely to cheat. Conversely, test-takers are presumably less motivated to cheat if no consequences are linked to the test results. Thus, we expect higher score differences in high-stakes settings *(Hypothesis 1)*.

Second, test administrators can implement countermeasures that overcome participants' motivation to cheat. Especially in high-stakes contexts, administrators are advised to use honesty contracts or follow-up verification tests (International Test Commission, 2006). Honesty contracts include explicit policies and negative consequences of cheating. Usually, such honesty contracts are presented to the test-taker prior to the testing and must be signed to indicate commitment. Verification tests are proctored follow-up tests that help to identify participants with aberrant test scores (Guo & Drasgow, 2010; Tendeiro et al., 2013). To work as a countermeasure designed to lower the test-takers' motivation to cheat, it is important to inform test-takers about the follow-up tests in advance. These procedures are often used in personnel selection (Lievens & Burke, 2011; Nye et al., 2008). In academic settings, institutions often implement honor codes not only to raise students' awareness of cheating, but also to call attention to the consequences linked to unethical behavior (McCabe & Treviño, 2002; O'Neill & Pfeiffer, 2012). Furthermore, other researchers suggested the use of specific instructions to reduce cheating that can contain the note that test results, or feedback, are only valid if the test-taker does not cheat (e.g., Wilhelm & McKnight, 2002). These precautions are intended to lower participants' cheating motivation, thus should result in reduced score differences *(Hypothesis 2).*

Third, the measurement instrument itself can affect participants' opportunity to cheat. Diedenhofen and Musch (2017) investigated cheating in an unproctored assessment, comparing a knowledge quiz and a reasoning task. They found that participants switched between browser tabs more often when answering knowledge questions that can be looked up on the Internet. Moreover, a positive relationship between page switches and test performance was found for the knowledge task, whereas no significant relationship was found for the reasoning test. These findings are in line with other studies reporting that cheating was most effective for subtests that assess abilities such as vocabulary and numeracy, in which performance can be enhanced through the use of a web search, dictionaries, or calculators (Bloemers, Oud, & Dam, 2016). In contrast, tasks that assess fluid abilities such as reasoning are less susceptible to cheating. Therefore, score difference should be higher for tests with a high *searchability (Hypothesis 3).*

Lastly, a factor that can lead to test score differences is the use of cross-mode comparisons. Unproctored assessments are usually administered over the Internet and, therefore, computer-based. Most studies compared these web-based assessments to proctored, computer-based assessments (e.g., Germine et al., 2012).

However, not all studies adopted identical test modes in both contexts: Some studies compared unproctored, computerized tests to proctored, paper-and-pencil assessments (e.g., Coyne et al., 2005). Although computer-based and paper-and-pencil ability assessments are considered equivalent for non-speeded measures (Mead & Drasgow, 1993), Schroeders and Wilhelm (2010) suggested differences in perceptual and motor skills as potential influencing factors. These differences, however, might lead to biased scores when proctored and unproctored assessments are compared across test media. If substantial mode differences exist, cross-mode comparisons are expected to result in larger mean differences between proctored and unproctored settings *(Hypothesis 4)*.

However, the equivalence of test scores across proctored and unproctored ability assessments should not be solely based on the comparison of mean scores. From a psychometric perspective it is important to ensure that test scores are only dependent on the trait in question and independent of testing conditions. The comparability of test scores gathered in different settings should be carried out using latent variable modeling (Schroeders & Wilhelm, 2010, 2011). However, such strict psychometric procedures require raw data, which is usually not available for meta-analysis. One of the simplest statistic indexing the similarity of the test-takers' ranking across conditions are correlation coefficients (Mead & Drasgow, 1993). A low correlation indicates differences across conditions in the assessment of test-takers' ability. If examinee ranking is invariant across modes (i.e., high cross-mode correlations are obtained), mean scores can be converted using linear transformations (Green, 1991; Hofer & Green, 1985). Therefore, we additionally examine correlations between ability test scores in proctored and unproctored settings.

Method

To make the present analyses transparent and reproducible (Nosek et al., 2015), we provide all material (i.e., coding protocol, data, syntax, additional tables and figures) in the Electronical Supplemental Material and online within the *Open Science Framework* (Center for Open Science, 2017): https://osf.io/xf8dq/

Literature Search and Study Selection

An overview of the literature search is depicted in Figure S1 (https://osf.io/3kaf8). In total we identified 101 potentially relevant studies, searching in major scientific databases, screening reference lists, and contacting authors. Subsequently, these studies were examined regarding the following criteria to be included in the meta-analysis: (a) The study reported a comparison of test scores obtained in a (remotely) proctored setting versus an unproctored setting, (b) administered cognitive ability measures, (c) was published during the last 25 years (1992–2017), (d) was written in English, and (e) reported appropriate statistical information that allowed the calculation of an effect size. Studies only reporting latent mean scores were excluded from the analyses. Furthermore, studies were excluded from the analyses, if (a) participants were actively instructed to cheat (e.g., Bloemers et al., 2016), (b) participants underwent different training phases prior to the assessments (e.g., online vs. traditional classes), or (c) different tools and aids were allowed across testing conditions (e.g., open vs. closed book exams; Brallier & Palm, 2015; Flesch & Ostler, 2010). After applying these criteria, 50 studies were considered eligible for the meta-analysis (see Table S1 for an overview of all studies included in the analysis; https://osf.io/3kaf8). Although we planned to include other assessment modalities that are discussed in the literature (e.g., smartphones; Harari et al., 2016), all but one study included in our analysis only reported paper-and-pencil or computer-based assessments. While assessments that used advanced security checks such as web-cams or security biometrics (Khan et al., 2017) were coded as remotely proctored ($n = 3$), studies that only used specific testing platforms that impeded browser-tab switches or returning back to previous questions were coded as unproctored assessments (see Table S1, column 7 for additional information on the type of proctoring; https://osf.io/3kaf8).

Coding Process

We developed a standardized coding protocol assessing descriptive information, effect sizes, and the moderator variables. For each study, we coded the type of publication (i.e., peer-reviewed journal, contribution to an edited book, master or doctoral thesis, conference presentations, or unpublished manuscripts), year of publication, mean age, percentage of female participants, sample type (i.e., children or adolescents up to 11th grade, college or university students, or mixed/ adult samples), the assessment context (i.e., academic research, educational, or I/O context), and research design (i.e., within- or between-subject).We extracted the sample sizes, means, and standard deviations of the ability scores in the unproctored and proctored setting as well as the correlation coefficients between test scores, and any other information that could be used to calculate an effect size (e.g., t-values). Moreover, we recorded whether test-takers expected consequences of the test results (such as a hiring decision or grading). If test performance yielded important consequences for the test-taker, the assessment was coded as high-stakes. To examine the usefulness of countermeasures against cheating, we coded different procedures (i.e., honesty contracts, honor codes, announcement of verification tests, instructions, or a combination of them). We also rated the proneness of the measure for cheating, that is, whether the searchability was high (e.g., for knowledge tests) or low (e.g., for figural matrices tests). Finally, we noted whether identical presentation modes (i.e., computerized or paper-and-pencil) were used in both assessment conditions.

All studies were coded twice by three independent raters. To evaluate the coding process, Cohen's (1960) κ was calculated. Intercoder agreement is considered strong for values exceeding .70 and excellent for values greater than .90 (LeBreton & Senter, 2008). The pairwise intercoder reliability ranged from .70 to .92. All discrepancies were discussed until consensus was reached.

Statistical Analyses

Calculation of effect sizes. As mean differences between scores assessed in proctored and unproctored settings were the primary topic of interest, the standardized mean difference Hedge's (1981) g was calculated with positive effect sizes indicating higher scores in the unproctored condition. For studies not reporting information necessary to calculate g, we applied transformation formulas to derive g from t values (Morris & DeShon, 2002). Studies that only reported multiple regression weights were excluded from the analysis (Aloe, 2015). For a subsample of studies reporting within-group comparisons, we additionally pooled Pearson correlations between the two test contexts to investigate the effects of mode differences on the rank ordering of test-takers. Extreme effect sizes were identified using internally studentized residuals (Viechtbauer & Cheung, 2010). Two extreme effect sizes with standardized residuals larger than 3 (Tukey, 1977) were removed from the analyses.

Meta-analytic model. Effect sizes were pooled using a random-effects model with a restricted maximum likelihood estimator (Viechtbauer, 2005). To account for dependent effect sizes (e.g., if a study reported more than one effect size for a given sample), we conducted a three-level meta-analysis (Cheung, 2014), in which individual effect sizes are nested within samples: Level 1 refers to the individual effect sizes, Level 2 refers to the effect sizes obtained using different instruments within a sample (with random Level 2 variance indicating the heterogeneity of effects due to the use of different tests of cognitive abilities), and Level 3 refers to the different samples (with the random Level 3 variance indicating the heterogeneity of effect sizes across samples after controlling for the different instruments at Level 2). To account for sampling error, we used different weighting procedures for the analysis of standardized mean differences and the correlational analysis. For the analysis of standardized mean differences, each effect size was weighted by the inverse of its variance, which is superior to other weighting procedures and results in more precise estimates of the mean effect (Marín-Martínez & Sánchez-Meca, 2010). Correlations were weighted using sample size weights, which is the most accurate procedure (Brannick, Yang, & Cafri, 2011). Heterogeneity in the observed effect sizes was quantified by the I^2 statistics (Higgins & Thompson, 2002), which describes the proportion of total variation in study estimates that is due to heterogeneity. Although I^2 does not measure heterogeneity on an absolute scale, higher values reflect more inconsistent results (Higgins, Thompson, Deeks, & Altman, 2003). We examined moderating effects on the pooled effect size using mixed-effect regression analyses of the R package metafor version 1.9.9 (Viechtbauer, 2010).

Results

The meta-analysis of mean differences was based on 49 studies[1] that were published between 2001 and 2017, mainly in peer-reviewed journals (67%). Unpublished work comprised master and doctoral theses (11%), conference proceedings (19%), and unpublished reports (3%). The meta-analytic database included 65 independent samples providing 109 effect sizes, with each sample reporting between 1 and 7 effect sizes. Overall, the meta-analysis covered scores from 100,434 participants (range of samples' *n*s: 19 to 24,750). Most studies were conducted in an educational (43%) or research context (41%); fewer studies reported on I/O contexts (16%). Low-stakes settings were reported more often than high-stakes settings (62% vs. 38%). In 29% of the samples countermeasures against cheating were implemented. Approximately half of the reported effect sizes (48%) were based on highly searchable tasks. In all cases that reported cross-mode comparisons (29%) the proctored assessment was paper-and-pencil, whereas the unproctored assessment was computerized.

The subsample reporting rank order stabilities comprised 5 studies published in peer-reviewed journals between 2005 and 2009. The studies included 7 independent samples providing 15 correlations. The total sample size was 1,280 (range of the samples' *n*s: 29 to 856). The subsample covered articles from all settings described

[1] One study only reported correlation coefficients and was therefore only included in the meta-analysis of rank order stability

above, with three studies being conducted in a research context and one each in educational and I/O context.

Mean Score Differences between Proctored and Unproctored Assessments

The pooled mean difference between proctored and unproctored settings was $\Delta = 0.20$ ($SE = 0.05$), 95% CI [0.10; 0.31]; thus, on average, test-takers achieved slightly higher scores in unproctored settings (Table 1). The between-cluster heterogeneity was $I^2 = .80$ and the within-cluster heterogeneity was $I^2 = .17$, indicating pronounced variability between samples, but negligible differences within samples. Furthermore, between-cluster variance–an absolute indicator of variability–was $\sigma^2_{(2)} = .14$, also indicating large heterogeneity according to common rules of thumb (Tett, Hundley, & Christiansen, 2017). To quantify the influence of a potential publication bias, we compared effect sizes from published sources (i.e., journal articles) to effect sizes from unpublished sources (i.e., theses, conference proceedings, and unpublished manuscripts). The respective mixed-effects regression analysis identified no significant difference between effect sizes extracted from both sources, $\gamma = 0.09$, $SE = 0.11$, $p = .43$. Furthermore, funnel plot analyses (Figure S2, see https://osf.io/3kaf8/) and a rank correlation test ($\tau = .12$, $p = .07$; Begg & Mazumdar, 1994), which tests the distribution of effect sizes for asymmetry, revealed no evidence of a potential publication bias. Although the funnel plot illustrated pronounced heterogeneity of the effect sizes, this most likely reflects the effects of moderators on score differences in proctored and unproctored settings.

Table II-1. Meta-Analysis of Mean Differences and Separate Moderator Analyses

	k_1	k_2	N	$Md_{(n)}$	g	SD_g	Δ	SE_Δ	z	Q_M	$\sigma^2_{(2)}$	$\sigma^2_{(3)}$	$I^2_{(2)}$	$I^2_{(3)}$
Overall	109	65	100,434	269	0.19	0.46	0.20	0.05	3.85*		.14	.03	.80	.17
Stakes									1.62	2.64				
High	67	42	79,203	312	0.31	0.47	0.27	0.07	4.02*		.15	.03	.80	.18
Low	42	23	21,231	212	-0.01	0.35	0.09	0.08	1.06		.13	.03	.75	.18
Countermeasures									1.64	2.68				
Yes	32	19	6,518	128	0.33	0.49	0.35	0.11	3.13*		.18	.04	.80	.18
No	77	46	93,916	298	0.13	0.43	0.15	0.06	2.53*		.13	.03	.79	.17
Searchability									3.73	13.95*				
High	51	34	21,407	187	0.40	0.50	0.38	0.08	4.76*		.14	.06	.65	.28
Low	58	34	79,863	336	0.00	0.32	0.02	0.05	0.44		.05	.03	.62	.33
Modality									1.73	3.01				
Cross mode	31	18	16,409	276	0.27	0.55	0.39	0.14	2.89*		.30	.02	.90	.07
Same mode	78	50	84,428	247	0.16	0.41	0.15	0.05	2.87*		.08	.04	.64	.33

Note. k_1 = Number of effect sizes; k_2 = Number of samples; N = Total sample size; $Md_{(n)}$ = Median of studies' sample size; g = Observed mean difference; Δ = Weighted standardized mean difference; SE_Δ = Standard error of Δ; $z = \Delta/SE_\Delta$; Q_M = test statistic for the omnibus test of coefficients (*df* = number of moderator categories − 1); $\sigma^2_{(2)}$ = between-cluster variance; $\sigma^2_{(3)}$ = within-cluster variance; $I^2_{(2)}$ = proportion of between-cluster heterogeneity; $I^2_{(3)}$ = proportion of within-cluster heterogeneity. * $p < .05$

To quantify the influence of moderators on the pooled effect, a mixed-effects regression analysis was conducted to examine the effects of test setting, countermeasures, searchability, and test media. The correlations among the moderators varied between r_Φ = -.18 and r_Φ = .44 (Table 2), indicating negligible multicollinearity. Together, the four moderators explained about 18% of the random variance (Table 2). Searchability was the only significant moderator (γ = 0.26, *SE* = 0.09, p < .01); mean score differences between proctored and unproctored settings were significantly larger for tasks that could be easily solved using the Internet (Δ = 0.38, *SE* = 0.08, p < .001) as compared to measures for which correct solutions were difficult to identify using ordinary web searches (Δ = 0.02, *SE* = 0.05, p = .66). Moderator analyses yielded the same results when each moderator was examined individually (Table 1). No significant effects were found for the other moderator variables, suggesting that the score differences between proctored and unproctored assessments are not affected by anticipated consequences of test results, the implementation of countermeasures against cheating, or a change of test media.

Table II-2. Moderator Analysis Including all Four Moderator Variables Simultaneously

	Moderator Analysis			Correlations		
	γ	SE_γ	z	(1)	(2)	(3)
Intercept	-0.04	0.09	-0.49			
(1) Stakes (1 = high, 0 = low)	0.08	0.11	0.69			
(2) Countermeasures (1 = yes, 0 = no)	0.12	0.11	1.05	.43		
(3) Searchability (1 = high, 0 = low)	0.26	0.09	2.87*	.44	.24	
(4) Modality (1 = cross, 0 = same)	0.14	0.10	1.50	-.17	-.18	.10
Q_M	17.62*					
$\sigma^2_{(2)}/\sigma^2_{(3)}$	0.11 / 0.03					
k_1/k_2	109 / 65					

Note. Phi coefficients of correlations for dichotomous variables are displayed (n = 109 effect sizes). γ = Fixed effects regression weight; SE_γ = Standard error of γ; Q_M = test statistic for the omnibus test of coefficients (df = number of moderator categories − 1); $\sigma^2_{(2)}$ = between-cluster variance; $\sigma^2_{(3)}$ = within-cluster variance; k_1 = Number of effect sizes; k_2 = Number of samples. * p < = .05

Rank Order Stability between Proctored and Unproctored Assessment

We identified a pooled correlation of $\rho = .58$ (SE = .10), 95% CI [.38, .78] (Figure 1). This result suggested a moderate relationship between test scores obtained in proctored and unproctored assessment, indicating substantial rank order changes for the different testing conditions. The between-cluster heterogeneity was $I^2 = .80$, and the within-cluster heterogeneity was $I^2 = .12$, indicating a large variability of the pooled effect sizes between samples. As the meta-analysis of correlation coefficients was based on a small number of effects, we did not pursue further moderator analyses.

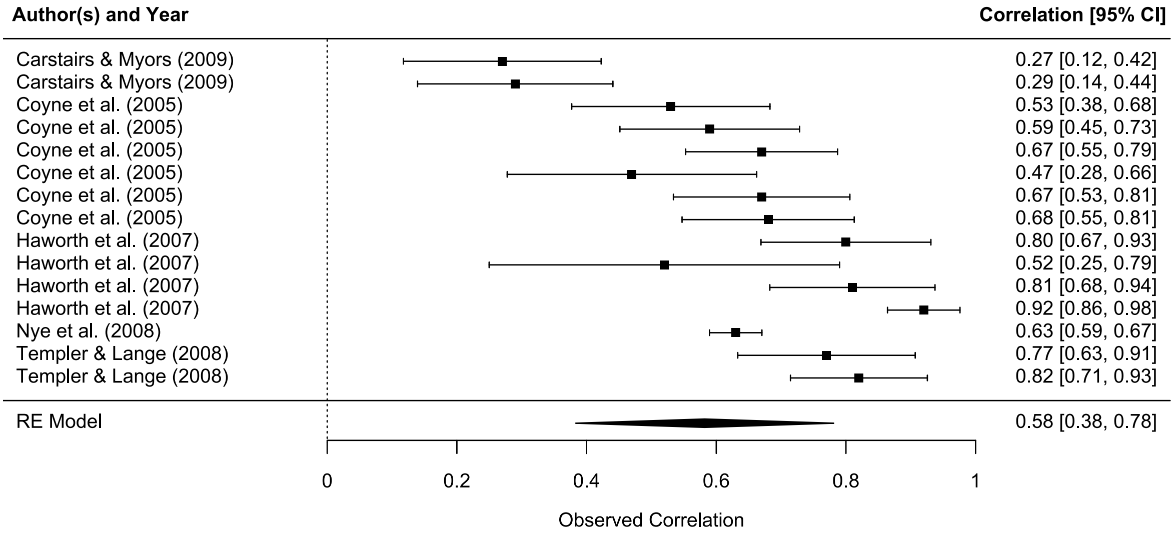

Figure II-1. Forest plot of the results of the random-effects model for the analysis of correlation coefficients.

Discussion

Unproctored, web-based assessments are typically faced with highly unstandardized settings that allow limited control over the participants' test-taking behavior. A pressing issue in this regard pertains to the question whether test scores from unproctored assessments can be readily compared to test scores from proctored lab sessions. Although a growing number of studies addressed score differences between proctored and unproctored settings, they reported rather inconclusive results (see also Do, 2009). Therefore, the current meta-analysis provided a comprehensive overview of the existing findings and studied various moderators of potential cross-mode differences. Overall, the meta-analysis revealed significantly higher scores on cognitive tests in unproctored settings as compared to proctored test contexts. However, with a standardized mean difference of $\Delta = 0.20$ the respective effect was rather small. Because the comparison of mean scores does not warrant conclusions about the equivalence of two measurements (AERA, APA, & NCME, 2014; Schroeders, 2009), we also analyzed correlations between scores of proctored and unproctored ability assessments for a subset of studies. This analysis showed a relationship of $\rho = .58$, indicating changes in the rank order of participants. These results suggest that participants' relative standing within a group does not solely depend on their ability, but also on other factors such as their motivation or their ability to cheat. However, since only 5 studies were included in the analyses, this result should be interpreted with caution.

In general, the effect sizes exhibited a large heterogeneity between samples. Therefore, we examined the influence of moderators on the observed score differences between proctored and unproctored ability assessments. Using a meta-regression approach, we found significant effects for the searchability of a task. If correct solutions were not easily identifiable over the Internet, mean score differences were approximately zero. This finding corroborates previous research suggesting that some tasks are more prone to cheating than others (Diedenhofen & Musch, 2017; Karim, Kaminsky, & Behrend, 2014). For instance, Bloemers and colleagues (2016) investigated cheating strategies for various subtests of a web-based cognitive ability test battery. They demonstrated that cheating was most effective for subtests that could be tampered through Internet searches, while cheating did not affect tasks that required complex reasoning. Interestingly, moderator analyses found no significant effect for score differences between proctored and unproctored settings for high and low-stakes testing. This finding does not support the prevailing assumption that cheating only corrupts high-stakes settings (Arthur, Glaze, Villado, & Taylor, 2010; Do, 2009) whereas it can be ignored in low-stakes testing. Furthermore, moderator analysis showed no significant effect for the implementation of countermeasures against cheating. Despite the vast body of research that advocates the implementation of countermeasures to improve data quality in unproctored assessments (Bartram, 2009; Bryan, Adams, & Monin, 2013; Dwight & Donovan, 2003; O'Neill & Pfeiffer, 2012), we found no empirical evidence for their effectiveness. Conversely, on a descriptive level, mean score differences appeared to be higher when countermeasures were implemented. Finally, differences in the test modes did not have a significant effect on the mean score differences. This finding is in line with previous results on the equivalence of paper-pencil and computerized ability tests (e.g., Mead & Drasgow, 1993; Schroeders & Wilhelm, 2010, 2011).

Recommendations for Unproctored and Proctored Assessment

Unproctored, web-based or mobile assessments promise a low-cost opportunity to reach large, heterogeneous, and geographically scattered samples (Fahrenberg et al., 2007; Gosling et al., 2010) and, thus, increasingly complement or even replace traditional data collection techniques. However, our results demonstrate considerable differences in the mean and variance-covariance structure between proctored and unproctored assessments. Based on our findings some words of caution are warranted if results obtained in one specific setting are to be generalized to the other. We also recommend against relying on countermeasures to overcome effects of cheating. What makes matters worse, the present data does not support the assumption that cheating is limited to high-stakes testing and can be ignored in low-stakes settings, including research contexts. Taking a pessimistic view, one might conclude that some participants will always cheat if they have the opportunity, regardless of countermeasures or anticipated consequences. On a more positive stance, participants will not cheat if they are not given the opportunity. Accordingly, a straightforward recommendation for ability assessments in unproctored settings is the development of test batteries that are limited to measures with a low searchability. In any case, administrators of unproctored assessments are encouraged to adopt post hoc strategies to identify potential cheaters, for example, using incidental data (Couper, 2005) such as reaction times or non-reactive behavioral data (Diedenhofen & Musch, 2017), seriousness checks (Aust, Diedenhofen, Ullrich, & Musch, 2012), or data-driven anomaly detection (Karabatsos, 2003). However, these analytical methods are no panacea, since identifying and excluding cheaters results in selective and most likely biased samples.

Limitations and Implications for Future Research

Some limitations to the present meta-analysis must be noted. First, most research on the comparability of ability scores in proctored and unproctored assessments focused on mean score differences, which do not allow drawing inferences about the equivalence of a measure. Measurement invariance is best studied with a latent variable approach (Raju, Laffitte, & Byrne, 2002; Schroeders & Wilhelm, 2011). We analyzed correlation coefficients as a proxy indicator for the equivalence of proctored and unproctored settings (Mead & Drasgow, 1993). Despite an extensive literature search, we only identified five studies that reported correlation across conditions. Therefore, we stress that the analysis is tentative and results must be interpreted with caution. Also, the correlations analyzed in the present meta-analyses were highly heterogeneous, ranging from $r = .27$ to $r = .92$, leaving open the question of potential moderator variables. Future research should also focus on the covariance structure by meta-analyzing raw data (Kaufmann, Reips, & Merki, 2016).

Second, the present research makes no inference about the extent of cheating in unproctored settings. Against the background of the data available for the study, we were able to ascertain that ability scores, on average, are higher in unproctored settings. Although dishonest behavior is one of the major concerns in unproctored settings (Tippins, 2009), the increased test scores might also be the result of reduced test anxiety, since participants might feel more comfortable if they are able to freely choose their testing environment (Stowell & Bennett, 2010). Further research might also address cheating directly by investigating appropriate means for the detection of dishonest behavior in ability tests. These measures include traditional approaches, such as scales measuring personality traits or integrity (McFarland & Ryan, 2000), or over-claiming (Bing, Kluemper, Kristl Davison, Taylor, & Novicevic, 2011), as well as data-driven approaches (Couper, 2005; Diedenhofen & Musch, 2017).

Finally, our data does not allow conclusions about groups of people that are more likely to cheat than others. We assume that individual differences in personality, moral beliefs, and social norms are predictive of cheating behavior. For example, some studies suggested culture-dependent differences in cheating behavior (Chapman & Lupton, 2004; McCabe, Feghali, & Abdallah, 2008). Future research might focus on test-takers who show large differences between an unproctored and a proctored assessment. For applied contexts, this might exert valuable diagnostic information (e.g., faking ability, Geiger, Sauter, Olderbak, & Wilhelm, 2016).

Conclusion

The presented meta-analysis identified higher mean scores for unproctored ability assessments, independent of the test setting (high- vs. low-stakes) and whether countermeasures were taken. However, mean score differences highly depended on the administered measure itself and its proneness to cheating. Mean differences were more pronounced for tasks that are easy to look up on the Internet, while no mean differences were found for other tasks. These findings, however, do not imply that unproctored ability assessments are not feasible *per se*. Based on the present meta-analysis, we recommend to carefully evaluate task characteristics when developing or choosing test instruments for an unproctored test battery. For example, the measurement of declarative knowledge seems better conducted in a proctored setting, whereas figural reasoning tasks might be comparably administered in unproctored contexts. We also caution researchers to generalize statements across test conditions and encourage test users to further examine the equivalence of proctored and unproctored ability tests with appropriate statistical methods.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Allen, E., & Seaman, J. (2014). *Grade Change: Tracking Online Education in the United States.* Newburyport, MA: Babson Survey Research Group.

Aloe, A. M. (2015). Inaccuracy of regression results in replacing bivariate correlations: *Inaccuracy of Regression Results. Research Synthesis Methods, 6*, 21–27. https://doi.org/10.1002/jrsm.1126

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*, 1–16.

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

Bartram, D. (2006). Testing on the Internet: Issues, Challenges and Opportunities in the Field of Occupational Assessment. In D. Bartram & Hambleton, R. K. (Eds.), *Computer-Based Testing and the Internet: Issues and Advances* (pp. 13–37). Hoboken: John Wiley & Sons.

Bartram, D. (2009). The international test commission guidelines on computer-based and internet-delivered testing. *Industrial and Organizational Psychology, 2*, 11–13. https://doi.org/10.1111/j.1754 9434.2008.01098.x

Begg, C. B., & Mazumdar, M. (1994). Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics, 50*, 1088–1101. https://doi.org/10.2307/2533446

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, 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*, 21–29.

Brallier, S., & Palm, L. (2015). Proctored and unproctored test performance. *International Journal of Teaching and Learning in Higher Education, 27*, 221–226.

Brannick, M. T., Yang, L.-Q., & Cafri, G. (2011). Comparison of weights for meta-analysis of r and d under realistic conditions. *Organizational Research Methods, 14*, 587–607. https://doi.org/10.1177/1094428110368725

Bryan, C. J., Adams, G. S., & Monin, B. (2013). When cheating would make you a cheater: Implicating the self prevents unethical behavior. *Journal of Experimental Psychology: General, 142*, 1001–1005. https://doi.org/10.1037/a0030655

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5. https://doi.org/10.1177/1745691610393980

Carstairs, J., & Myors, B. (2009). Internet testing: A natural experiment reveals test score inflation on a high-stakes, unproctored cognitive test. *Computers in Human Behavior, 25*, 738–742. https://doi.org/10.1016/j.chb.2009.01.011

Center for Open Science. (2017). https://cos.io/ [Last access September 29, 2017].

Chapman, K. J., & Lupton, R. A. (2004). Academic dishonesty in a global educational market: a comparison of Hong Kong and American university business students. *International Journal of Educational Management, 18*, 425–435. https://doi.org/10.1108/09513540410563130

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*, 211–229. https://doi.org/10.1037/a0032968

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and*

Psychological Measurement, 20, 37–46. https://doi.org/10.1177/001316446002000104

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

Coyne, I., Warszta, T., Beadle, S., & Sheehan, N. (2005). The impact of mode of administration on the equivalence of a test battery: A quasi-experimental design. *International Journal of Selection and Assessment, 13*, 220–224. https://doi.org/10.1111/j.1468-2389.2005.00318.x

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

Do, B.-R. (2009). Research on unproctored internet testing. *Industrial and Organizational Psychology, 2*, 49–51. https://doi.org/10.1111/j.1754-9434.2008.01107.x

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1–23. https://doi.org/10.1207/S15327043HUP1601_1

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research, 15*, 195–219. https://doi.org/10.1108/10662240510590360

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory assessment - monitoring behavior in daily life settings. *European Journal of Psychological Assessment, 23*, 206–213. https://doi.org/10.1027/1015-5759.23.4.206

Flesch, M., & Ostler, E. (2010). Analysis of proctored versus non-proctored tests in online algebra courses. *MathAMATYC Educator, 2*, 8–14.

Geiger, M., Sauter, R., Olderbak, S., & Wilhelm, O. (2016). Faking ability: Measurement and validity. *Personality and Individual Differences, 101*, 480. https://doi.org/10.1016/j.paid.2016.05.147

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*, 847–857. https://doi.org/10.3758/s13423-012-0296-9

Gibby, R. E., Ispas, D., Mccloy, R. A., & Biga, A. (2009). Moving beyond the challenges to make unproctored internet testing a reality. *Industrial and Organizational Psychology, 2*, 64–68. https://doi.org/10.1111/j.1754-9434.2008.01110.x

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

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences, 33*, 94–95. https://doi.org/10.1017/S0140525X10000300

Green, B. F. (1991). Guidelines for Computer Testing. In T. B. Gutkin & S. L. Wise (Eds.), *The Computer and the Decision-Making Process* (pp. 245–273). Hillsdale: Lawrence Erlbaum Associates.

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*, 351–364.

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*, 838–854. https://doi.org/10.1177/1745691616650285

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107.

https://doi.org/10.2307/1164588

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology, 53*, 826–838.

Ihme, J. M., Lemke, F., Lieder, K., Martin, F., MÃ¼ller, J. C., & Schmidt, S. (2009). Comparison of ability tests administered online and in the laboratory. *Behavior Research Methods, 41*, 1183–1189. https://doi.org/10.3758/BRM.41.4.1183

International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*, 143–171.

Jensen, C., & Thomsen, J. P. F. (2014). Self-reported cheating in web surveys on political knowledge. *Quality & Quantity, 48, 3343–3354.* https://doi.org/10.1007/s11135-013-9960-z

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 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

Kaufmann, E., Reips, U.-D., & Merki, K. M. (2016). Avoiding methodological biases in meta-analysis: Use of online versus offline individual participant data (IPD) in psychlogy. *Zeitschrift FÃ¼r Psychologie, 224*, 157–167.

https://doi.org/10.1027/2151-2604/a000251

Khan, S. M., Suendermann-Oeft, D., Evanini, K., Williamson, D. M., Paris, S., Qian, Y., ... Davis, L. (2017). MAP: Multimodal Assessment Platform for Interactive Communication Competency. In S. Shehata & J. P.-L. Tan (Eds.), *Practitioner Track Proceedings of the 7th International Learning Analytics & Knowledge Conference.* Vancouver, CA: SoLAR.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815–852. https://doi.org/10.1177/1094428106296642

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*, 817–824. https://doi.org/10.1348/096317910X522672

Lievens, F., & Harris, M. M. (2003). Research on internet recruiting and testing: Current status and future directions. In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (pp. 131–165). Chichester: Wiley.

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by Inverse Variance or by Sample Size in Random-Effects Meta-Analysis. *Educational and Psychological Measurement, 70*, 56–73. https://doi.org/10.1177/0013164409344534

McCabe, D. L., Feghali, T., & Abdallah, H. (2008). Academic dishonesty in the Middle East: Individual and contextual factors. *Research in Higher Education, 49*, 451–467. https://doi.org/10.1007/s11162-008-9092-9

McCabe, D. L., & Treviño, L. K. (2002). Honesty and honor codes. *Academe, 88*, 37. https://doi.org/10.2307/40252118

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812–821. https://doi.org/10.1037//0021-9010.85.5.812

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7*, 221–237. https://doi.org/10.1177/1745691612441215

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105–125. https://doi.org/10.1037//1082-989X.7.1.105

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 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*, 112–120. https://doi.org/10.1080/09639284.2011.590012

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*, 231–245. https://doi.org/10.1080/09639284.2011.590012

Preckel, F., & Thiemann, H. (2003). Online- versus paper-pencil-version of a high potential intelligence test. *Swiss Journal of Psychology, 62*, 131–138. https://doi.org/10.1024//1421-0185.62.2.131

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517–529. https://doi.org/10.1037/0021-9010.87.3.517

Reynolds, D. H., Wasko, L. E., Sinar, E. F., Raymark, P. H., & Jones, J. A. (2009). UIT or not UIT? That is not the only question. *Industrial and Organizational Psychology, 2*, 52–57. https://doi.org/10.1111/j.1754-9434.2008.01108.x

Rovai, A. P. (2000). Online and traditional assessments: What is the difference?

The Internet and Higher Education, 3, 141–151. https://doi.org/10.1016/S1096-7516(01)00028-8

Schroeders, U. (2009). Testing for equivalence of test data across media. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment. Lession learned from the PISA 2006 computer-based assessment of science (CBAS) and implications for large scale testing* (pp. 164–170). JRC Scientific and Technical Report EUR 23679 EN.

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment, 26*, 284–292. https://doi.org/10.1027/1015-5759/a000038

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*, 849–869. https://doi.org/10.1177/0013164410391468

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

Tendeiro, J. N., Meijer, R. R., Schakel, L., & Maij-de Meij, A. M. (2013). Using cumulative sum statistics to detect inconsistencies in unproctored internet testing. *Educational and Psychological Measurement, 73*, 143–161. https://doi.org/10.1177/0013164412444787

Tett, R. P., Hundley, N. A., & Christiansen, N. D. (2017). Meta-analysis and the myth of generalizability. *Industrial and Organizational Psychology, 10*, 421–456. https://doi.org/10.1017/iop.2017.26

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology, 2*, 2–10. https://doi.org/10.1111/j.1754-9434.2008.01097.x

Tippins, N. T. (2011). Overview of technology-enhanced assessments. In N. T. Tippins & S. Adler (Eds.), *Technology-Enhanced Assessment of Talent* (pp.

1–19). San Francisco, Ca: Wiley.

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*, 189–225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*, 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48.

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*, 112–125. https://doi.org/10.1002/jrsm.11

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). Seattle: Hogrefe & Huber.

Williamson, K. C., Williamson, V. M., & Hinze, S. R. (2016). Administering spatial and cognitive instruments in-class and on-line: Are these equivalent? *Journal of Science Education and Technology, 26*, 12–23. https://doi.org/10.1007/s10956-016-9645-1

III. Manuscript 2

On the Dimensionality of Crystallized Intelligence: A Smartphone-Based Assessment

Diana Steger

Ulm University, Germany

Ulrich Schroeders

University of Kassel, Germany

Oliver Wilhelm

Ulm University, Germany

Status: Published

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

Abstract

Crystallized intelligence (*gc*) is a prominent factor in consensual theories on the structure of intelligence. Although declarative knowledge is arguably a core aspect of *gc*, little is known about the dimensionality of knowledge in adults; the proposed dimensional models vary broadly from unidimensionality, to three-dimensional models (science, humanities, and civics), to a six-dimensional model with an overarching g-factor. While previous studies were mostly based on narrow item samples once administered to a specific sample within a restricted time frame, we used a smartphone-based approach to investigate the dimensionality of knowledge based on a large set of items administered to a heterogeneous sample. More specifically, questions were randomly drawn from a pool of 4,050 items from 34 subject domains such as chemistry, arts, and politics and administered to an age- and ability-heterogeneous sample of 1,117 participants. We calculated Weighted Likelihood Estimates separately for each domain and then estimated a series of principal component analyses with increasing number of factors. The component solution at different levels match models reported in previous studies on the dimensionality of knowledge. We conclude that the dimensionality of declarative knowledge highly depends on the item and person sample. Finally, we discuss different approaches to model *gc*, give advice on the measurement of *gc* in general and discuss weaknesses and strengths of mobile assessments.

Keywords: crystallized intelligence, declarative knowledge, dimensionality, smartphone-based assessment

Introduction

Crystallized intelligence (*gc*) was originally defined as "acculturation knowledge" measured with "tasks indicating breadth and depth of the knowledge of the dominant culture" (Horn & Noll, 1997, p. 69). But what is meant by the expression "breadth and depth of knowledge"? What are relevant facets of knowledge? What is the dimensionality of *gc*? Although *gc* is a prominent factor in consensual theories on the structure of intelligence—such as the Extended Gf-Gc theory (Horn & Noll, 1997), Carroll's Three Stratum theory (1993), and Cattell-Horn-Carroll theory (McGrew, 2009)—our knowledge about *gc* is still limited; presumably because it is complex or even impossible to measure *gc* in an age-heterogeneous sample that had very different learning environments using a traditional declarative knowledge test. To this end, we present a smartphone-based assessment of *gc* operationalized as a test of declarative knowledge. In the introduction, we discuss the difficulties in the assessment of *gc*, summarize existing studies on the dimensionality of declarative knowledge and, finally, present a smartphone-based approach to measure declarative knowledge.

Challenges in the Assessment of gc

Different theoretical frameworks of intelligence provide different definitions and conceptualizations of *gc*: Cattell (1941, 1943) described *gc* as a combination of skills, knowledge, and language-related abilities in a broad range of domains, emphasizing the importance of declarative knowledge. Carroll (1993), in contrast, focused on language-related abilities such as vocabulary, reading, and writing skills and the Cattell-Horn-Carroll theory of human cognitive abilities (McGrew, 2009) contains two knowledge factors—a comprehension-knowledge factor (labelled *gc*), and a factor assessing domain-specific knowledge (*gkn*). Although described as two distinct broad abilities, *gc* and *gkn* obviously build a continuum: spanning from general knowledge that is shared by all individuals of a respective culture, to highly specialized professional knowledge that goes far beyond the knowledge gained during regular schooling. But this is not the only attempt to describe different nuances of declarative knowledge: Cattell (1971) used the terms *historical gc* and *present gc* to distinguish between knowledge acquired in homogeneous learning environments during childhood and adolescence and knowledge based on individualized learning opportunities following formal schooling. Although McGrew's (2009) concepts of (general) *gc* and (specialized) *gkn* overlap to a certain degree with Cattell's (1971) *historical* and *present gc*, Cattell describes knowledge with a focus on the time during the life course when the knowledge was acquired. While *present gc* is the knowledge that an individual has recently acquired or used, *historical gc* might become fogged over time since it was last relevant.

Even though many theorists would follow Cattell's original definition of *gc* as declarative knowledge, often verbal indicators—predominantly vocabulary measures—are used in the assessment of gc (see Schipolowski, Wilhelm, & Schroeders, 2015 for further discussion). One reason for this narrow operationalization of *gc* might be the difficulties one has to face when adopting Cattell's view: Firstly, *gc* is extremely broad; it covers knowledge that we acquired throughout different stages of our educational and professional careers, and at different levels of specialization. It also encompasses avocational knowledge that we acquire outside of our professions, for example, through our favorite hobbies (Ackerman, 2000). Thus, everybody over time develops his or her idiosyncratic profile of knowledge and expertise and no one even tried to exhaustively list (let alone test) domains of knowledge and expertise. Cattell (1971) regarded the assessment of declarative knowledge possible only if the individual learning history is taken into account, which becomes clear in his statement "that crystallized ability begins after school to extend into Protean forms and that no single investment [...] can be used as manifestation by which to test all people" (Cattell, 1971, p. 121).

Keeping these difficulties in mind, how should an ideal declarative knowledge test—that covers breadth and depth of knowledge—look like? A decent *gc* assessment should cover a variety of domains which are representative for the knowledge in a given culture. Moreover, a conclusive assessment of *gc* should also include specialized knowledge gained, for example, through professional experience or hobbies to more exhaustively depict the heterogeneity of the construct. Usually, researchers assess the more homogenous *historical gc*, neglecting domain-specific knowledge that has been acquired after school. Unfortunately, without knowing the dimensionality of knowledge, it is hard to judge whether an assessment is broad (i.e., many different domains) and deep (i.e., many items within a given domain).

Empirical Findings on the Dimensionality of Declarative Knowledge

Empirical findings on the dimensionality of declarative knowledge vary broadly and include a unidimensional model, two, three or four (or even six) factor models (see Table 1 for an overview). Wilhelm, Schroeders, and Schipolowski (2014) proposed a unidimensional model of declarative knowledge in the manual of the *Berlin Test of Fluid and Crystallized Intelligence*. The test, however, was originally constructed with three separate scales, that is, science, humanities, and social studies (Schipolowski, 2009)—a classification that can be also found for the academic knowledge scale of the Woodcock-Johnson IV Test of Achievement (Schrank, Mather, & McGrew, 2014). Other empirical investigations suggest a two-dimensional model of declarative knowledge, splitting a general factor in a *Natural* and a *Social Sciences* factor (e.g., Hossiep & Schulte, 2008). The four factor models reported in Table 1 do not only differ in the factor composition—different models may include a *Mechanical Knowledge* factor (Rolfhus & Ackerman, 1999), a *Business and Law* factor (Ackerman, 2000), or a *Psychology and Biology* factor (Ackerman, Bowen, Beier, & Kanfer, 2001)—but they may or may not assume an overarching general factor of knowledge. In the six-dimensional models, the an even broader factor spectrum is found: For example, Irwing, Cammock, and Lynn (2001) report not only three factors that closely resemble the above-mentioned factors, but also three factors that contain less curriculum-oriented knowledge (i.e., Fashion, Family, and Physical Health and

Recreation).

But are these models really as diverse as they appear? Most authors of the studies reported in Table 1 would agree on the existence of an overarching factor of declarative knowledge alongside of a set of content factors, even though samples largely varying in terms of age distribution and educational background (e.g., high school students or Psychology freshmen, or mixed adults samples; see columns 3 and 4). To the degree that those range-restricted context variables are correlated with declarative knowledge, the conclusions concerning dimensionality are compromised.

Table III-1. Overview of Studies Examining the Dimensionality of Knowledge

		Sample						
Study	*N*	Age *M* (*SD*)	Characteristics	Domains	Average items per domain	Identified factors	Model	Method
Ackerman (2000)	228	34.2 (10.6)	middle-aged adults	18	71.0	(1) Science (2) Civics (3) Humanities (4) Business/Law	CF	EFA
Ackerman, Bowen, Beier, & Kanfer (2001)	320	19.0 (0.4)	college freshman	19	NA	(1) Physical Science (2) Biology/Psychology (3) Humanities (4) Civics	CF	EFA
Amthauer, Brocke, Liepmann, & Beauducel (2001)	661	28.0 (9.6)	mainly graduates	6	14.0	(1) Science (2) Economics (3) Math (4) Culture (5) Geography/History (6) Everyday life	CF	MDS, CFA
Engelberg (2015)	202	NA	students (grade 10 – 11, university)	15	13.7	2 factors (no labels provided by the author)	CF	EFA, CFA

(continued)

Table III-1. Overview of Studies Examining the Dimensionality of Knowledge *(Continued)*

Study	N	Sample		Domains	Average items per domain	Identified factors	Model	Method
		Age M (SD)	Characteristics					
Hossiep & Schulte (2008)	4,224	36.5 (12.3)	mainly graduates	11	14.0	(1) humanities/social studies (2) natural/technical sciences	CF	PCA
Irwing, Cammock, & Lynn (2001)	718	20.9 (5.3)	mainly undergraduates	18	12.0	(1) Current affairs (2) Fashion (3) Family (4) Physical Health and Recreation (5) Arts (6) Science	HO	EFA, CFA
Lynn & Irwing (2002)	1,047	20.5 (3.3)	undergraduates	6	12.0	(1) General factor of semantic memory	UN	CFA
Lynn, Irwing, & Cammock (2001)	636	20.4 (3.3)	undergraduates	18	10.1	(1) Current affairs (2) Fashion (3) Family (4) Arts (5) Science (6) Physical Health and Recreation	HO	PCA, CFA

(continued)

Table III-1. Overview of Studies Examining the Dimensionality of Knowledge *(Continued)*

Study	*N*	Sample Age *M* (*SD*)	 Characteristics	Domains	Average items per domain	Identified factors	Model	Method
Lynn, Wilsberg, & Margraf-Stiksrud (2004)	302	17.6 (0.9)	students (grade 12)	17	5.6	4 factors (with limited interpretability)	OF	PCA
Rolfhus & Ackerman (1999)	141	19.1 (1.2)	undergraduates	20	68.4	(1) Humanities (2) Science (3) Civics (4) Mechanical	HO	EFA
Wilhelm, Schroeders, & Schipolowski (2014)	4,215	NA	high school students	16	4.0	(1) general factor of crystallized intelligence	UN	CFA

Note. CF = correlated factors; HO = higher-order; OF = orthogonal factors; UN = unidimensional; CF = Correlated Factors; EFA = Exploratory factor analysis; CA = Cluster analysis; CFA = Confirmatory factor analysis; PCA = Principal component analysis; MDS = multidimensional scaling.

Overall, the existing studies on the dimensionality of knowledge are difficult to compare as they vary in terms of the sample, the measure of declarative knowledge, and the study design. As these factors contribute to the diverging results presented in Table 1, we discuss each of them in the following: First, the factor structure of declarative knowledge is likely to change depending on sample characteristics, such as the educational background or the level of declarative knowledge itself. The ability-differentiation hypothesis, also known as Spearman's (1927) law of diminishing returns, states that the strength of the correlations amongst a set of cognitive ability indicators decreases with higher ability level. Another sample-related differentiation hypothesis, the age-related differentiation, argues that the factor structure of cognitive abilities might vary with age (e.g., Li et al., 2004). Surprisingly little is known about the dimensionality of crystallized abilities in children and adolescents: One study that examined the mean and covariance structure of *gc* of in an adolescent sample found a linear increase for mean levels of crystallized ability, but no evidence for the development of a more nuanced knowledge profile during this age period (Schroeders, Schipolowski, & Wilhelm, 2015). While it is only reasonable to assume that more specialized knowledge (as conceptualized in *gkn*) develops only after school (see also the trajectories of occupational and avocational knowledge suggested by Ackerman, 1996), the dimensionality of more general *gc* might evolve in the first life decade. Overall, conclusive evidence concerning the validity of the differentiation hypotheses is sparse (Molenaar, Dolan, Wicherts, & van der Maas, 2010; Schroeders, Schipolowski, & Wilhelm, 2015). Recent research indicates that when choosing adequate sample size and methods, no evidence for the differentiation hypothesis can be found (Hartung, Doebler, Schroeders, & Wilhelm, 2018; Kievit, Fuhrmann, Borgeest, Simpson-Kent, & Henson, 2018). Although no systematic differentiation is expected for declarative knowledge, studies that use only very narrow samples (e.g., Psychology freshmen) underestimate the true variability of the construct, which affects both reliability of the measure and bias the dimensional analyses.

Second, characteristics of the measure itself influence the dimensionality of knowledge. Although it is often assumed that tests are compiled from a set of randomly drawn items from an infinite item universe and thus can be used interchangeably, this romantic idea is simply not true. Schroeders, Wilhelm, and Olaru (2016b) demonstrated the impact of item sampling on sex differences in a declarative knowledge test using metaheuristics to select items in such a manner that latent mean differences are maximized. Depending on the specific non-randomly drawn item set, the compiled knowledge tests favor either females or males, which emphasizes Loevinger's (1965) notion that item selection is always expert selection. Moreover, declarative knowledge items are seldom purified measures of a single domain. For example, the item "*When was Mozart born?*" could be assigned to both History and Music. Ultimately, classifying indicators into domains will often result in opaque assignments and inconsistencies if done by several raters. Therefore, it seems impossible to define a unambiguous and exhaustive catalogue of subject domains (see also Loevinger, 1965).

Third, in large-scale studies with item samples, research designs implemented to reduce the individual workload might contribute to the inconclusive results reported in Table 1. For example, some studies (e.g., Ackerman, 2000; Rolfhus & Ackerman, 1999) used a so-called power design, that is, the assessment is aborted after answering a pre-defined number of consecutive items incorrect. Such designs require sorting of the items in ascending difficulty; however, difficulty estimates might fluctuate across samples and the assessment assumes unidimensionality within a domain, which often is a research question of the study rather than a given fact to start with. Furthermore, the domain tests in a power-design vary considerably in length for the different individuals, as participants that are more knowledgeable will answer more questions until the test is stopped. Therefore, their response vector will provide more dependable information, resulting in more precise measures. In the following, we introduce a smartphone-based assessment as an alternative to traditional lab assessment that enables researchers to test a large number of items and reach a large and diverse sample to reduce biases that arise from traditional assessment methods.

A Smartphone-Based Knowledge Assessment

In psychological research, smartphone-based assessments were primarily used for implementing measures of typical behavior (e.g., affective well-being in different contexts; Riediger, Wrzus, & Wagner, 2014), rather than for tests of maximal performance. But how can a smartphone-based assessment complement or enhance the traditional assessment of *gc*? Possible advantages of smartphone-based assessments are: a) the cost-efficient collection of data from a large and heterogeneous sample (Condon & Revelle, 2014; Dufau et al., 2011) , b) the administration of a large item set (Revelle et al., 2017), and c) the possibility of "gamification" (Deterding, Dixon, Khaled, & Nacke, 2011; Kim & Shute, 2015; Sailer, Hense, Mayr, & Mandl, 2017) to increase participants' motivation.

First, smartphone-based assessments offer the possibility to recruit participants that are heterogeneous in terms of age and educational background. Traditional lab-based assessments are often limited to specific groups of participants who live in a particular geographic area and who have time and motivation to take part in psychological assessments, usually attracting White, Educated, Industrial, Rich, and Developed (WEIRD) subjects (Henrich, Heine, & Norenzayan, 2010). The assessment of heterogeneous samples is especially important, since previous studies on the dimensionality of declarative knowledge often used high selective samples in terms of age or educational background, which severely affects the generalizability of the results.

Second, traditional knowledge assessments usually use only a limited number of items per domain, depicting only a small part of the construct. Setting aside the three studies by Ackerman (2000), Ackerman and colleagues (2001), and Rolfhus and Ackerman (1999), the other studies depicted in Table 1 used knowledge tests that vary between only 4 and 14 items per domain. Smartphone-based knowledge assessments, in contrast, can be used to implement a broad item sample for every domain or even subdomains. For example, the domain of history can include subdomains such as the ancient Greek civilization, the Middle Ages, the Enlightment, World War I and II, as well as national and international recent history with a sufficient number of items for each subdomain.

Third, to increase participants' motivation when answering many questions, smartphone-based solutions also allow for "gamification", for example by using badges to incentivize user activity (Hamari, 2017). This is especially important since in the present smartphone-based approach, participants are self-paced and free to answer as many questions as they like.

To approach problems of traditional assessments of declarative knowledge in the present study, we report on the development of a large item pool and a quiz app and provide psychometric results for the knowledge domains. We further examine the factor structure of declarative knowledge using a sequential approach (Goldberg, 2006). Doing so we want to analyze the generalizability of the different models reported in the literature and demonstrate possible convergences and divergences with the present model. Finally, we report on an exploratory factor analysis to further investigate the underlying factor structure for the given set of knowledge domains.

Method

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) in the Electronical Supplemental Material and online within the *Open Science Framework* (Center for Open Science, 2017): https://osf.io/es35a/

Design and Participants

We developed a mobile quiz app to assess a broad range of declarative knowledge with 4,050 items. We recruited participants via online forums, Facebook groups, magazine articles, advertising flyers, and radio interviews. The analyzed data set is a snapshot (collected from October 2016 to February 2018) of an ongoing data collection. After downloading, participants received an email providing information about the study and a consent form. After consenting, participants could play the quiz wherever and whenever they liked, without any restrictions concerning duration, daytime, testing context, or internet connection. Participation was voluntary and no monetary incentive was offered. In the present analysis, we analyzed data of 1,117 participants, of which 58% (N = 647) were female with a mean age of 36.6 years (SD = 15.3). Regarding their educational background, 2.7% indicated to have no degree, 4.9% indicated to have a degree from a vocational track school (*Hauptschulabschluss*), 18.6% indicated to have a degree from an intermediate track school (*Realschulabschluss*), 31.8% indicated to have a degree from an academic track school (*Abitur*), and 42.0% indicated to have a university degree. Generally, the sample is balanced in terms of gender and—compared to traditional lab samples—more heterogeneous in terms of age. Regarding educational background, academic qualifications are rather overrepresented. Based on our diverse recruiting strategies and the download statistics, we assume that still a large proportion of our sample consists of students (most prominently psychology undergraduates), which is also in line with the sample characteristics given above. On average, participants worked on 1,230 items (MD = 797; SD = 1,074; min = 189, max = 4,050). We found a weak relationship between participants' age and the number of questions answered (N = 1,117, r = .16, p < .001). In contrast, we found no differences in the number of questions answered for women (M = 1,233, SD = 1,086) and men (M = 1,226, SD = 1,059). No clear pattern emerged for the number of items answered by participants' educational background (see Table S1 in the online supplement for further information).

Measures and Apparatus

For the item development, we reviewed existing knowledge test batteries (Amthauer, Brocke, Liepmann, & Beauducel, 2001; Hossiep & Schulte, 2008; Irwing et al., 2001; Mariani, Sacco, Spinnler, & Venneri, 2002; Roberts et al., 2000; Schrank et al., 2014; Wilhelm et al., 2014), empirical classifications (Engelberg, 2015; Rolfhus & Ackerman, 1996, 1999), the courses in German universities, and various vocational profiles to ensure that our knowledge test covered a wide range of knowledge gained by post-secondary education and training. This search resulted in a list of 34 knowledge domains (see Table S2 in the online supplement for further information). Since our measurement intention was to assess declarative knowledge in breadth and depth, we aimed at developing at least 100 items for each domain. To evaluate quality, items were reviewed with respect to content and domain fit by experts (i.e., people with a job, degree, or experience in a related field), with respect to grammar, wording, and psychometric quality by graduate Psychology students, and revised accordingly. All items were multiple-choice with only one valid solution and three distractors.

The knowledge tests was part of a larger study assessing cognitive abilities via smartphone app (https://www.iq-app.de), which was available for Android and iOS in German. Knowledge items were presented in sets of 27 questions per round (9 domains with 3 items each randomly drawn from the item pool). It was not possible to skip questions. Upon answering an item, participants received feedback on their performance, and a detailed knowledge report covering all 34 domains, updated after every round the participants completed. To reduce the competitiveness participants only received their individual scores without social comparison. As incentive, participants could collect badges for good performance (e.g., a strike of 10 correctly answered items), good overall performance (e.g., 20 correctly answered items in a specific domain) as well as regular use of the app (e.g., three days in a row, see also Hamari, 2015). Participants could choose how many rounds of the quiz they wanted to play, similar to the *Synthetic Aperture Personality Assessment* (SAPA) approach (Condon & Revelle, 2014). As a result, the data matrix had large proportions of missing data, also known as data *Massively Missing Completely at Random* (Revelle et al., 2017)[2].

[2] Massively Missing Completely at Random (MMCAR) is the continuation of Lord's (1955) idea of

Statistical Analyses

To compute domain scores for the selected domains, we calculated *Weighted Likelihood Estimates* (WLE; Warm, 1989) based on two-parameter logistic item response models separately for each knowledge domain. For the computation of domain scores, we used data from all participants that answered at least 15 items in the respective domain (equals five rounds with 3 items each). This cut-off criterium allowed the estimation of domain scores with sufficient accuracy while retaining a sufficient sample size.

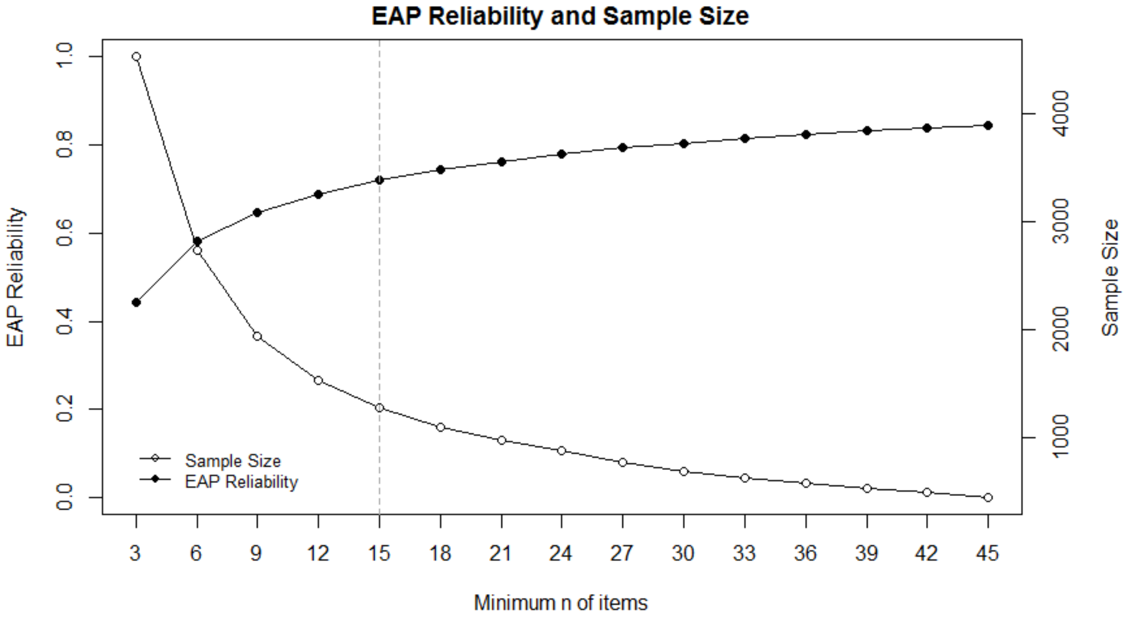

Figure III-1. Mean EAP Reliability and Sample Size Depending on the Number of Items Administered.

Note. Dashed line indicates the cut-off value used for the present analysis.

Figure 1 shows the relation between expected *a posteriori* (EAP) reliability estimates and the respective sample sizes for different cut-off scores. The concept of EAP relates to the principles of Bayesian statistics, that is, person scores and the accompanying reliability estimates rely on a posterior probability distribution given persons' response vector and the model parameters. We excluded eight erroneous

matrix sampling (i.e., sampling not only subjects but also items) and a more generalized form of Balanced Incomplete Blocks designs. In MMCAR data, participants usually only respond to a small, randomly drawn fraction of the existing item pool.

items from the analysis and 84 items with auditory content to avoid presentation mode effects. We also excluded items with poor psychometric quality: a) items that were too easy (observed item mean $> .97$; $n = 243$), b) too difficult (observed item mean $< .28$; $n = 257$), and c) items with a negative discrimination parameter ($n = 134$) in sequential IRT analyses. Based on the final model, we estimated WLE scores for each domain that served as indicators in subsequent analyses.

For all further analysis, we excluded knowledge domains that mainly cover current events knowledge (Beier & Ackerman, 2003; Hambrick, Pink, Meinz, Pettibone, & Oswald, 2008), that is knowledge about current events that is experienced through the media rather than taught in school (i.e., *Celebrities*, *Fashion*, *Film and TV*, *Pop Culture*, *Pop Music*, and *Sports*). Moreover, we removed *Education* as a knowledge domain that showed an exceptionally low EAP reliability below .60, and also *Linguistics* because more than 25% of the items relied on auditory material. In two cases, we calculated joint domain scores for *Mathematics* and *Statistics*, as well as *Technology* and *Computer Science* due to their high conceptual overlap. In total, we used 24 knowledge domain scores in the subsequent analyses, comprising 2,210 items that met all of our inclusion criteria.

Results

In Table 2, we provide descriptive statistics for the knowledge domain scores: The first two columns show the number of items for each knowledge domain after item selection and the number of removed items, respectively. On average, 22 items were excluded per domain (SD = 9.5). Although we had to exclude more than one third of the items for the *Literature* domain, we kept it in the analysis as a cultural and verbal specific domain. The lowest number of items per domain was 69 (*Architecture*), which was still considered sufficient for the subsequent analysis and ample in comparison to the usually found item sample size in the research literature. Besides the proportion correct (on a manifest level), we also report the

EAP reliabilities which vary between .63 and .84 (with a mean of .72, see Table 2). The correlations among the knowledge domains are documented in Table S3 of the online supplement. All but one (*Housekeeping* and *Mathematics*) pairwise correlations were positive with a mean of r = .39 and a range between r = -.10 and r = .66. This positive manifold has often been found between cognitive abilities in general (e.g., Bartholomew, 2004; Van der Maas et al., 2006), and domain-specific knowledge in particular (e.g., Rolfhus & Ackerman, 1999; Schipolowski, 2014).

Table III-2. Descriptive Statistics of the Knowledge Domains

	No. of			Item Difficulty		EAP
Domain	total items	excluded items	cases	Mean	*SD*	Reliabilty
Anthropology	86	24	709	.64	.20	.66
Architecture	69	31	790	.64	.21	.67
Arts	81	19	789	.66	.21	.66
Biology	78	25	751	.70	.21	.80
Chemistry	97	7	789	.64	.20	.80
Ecology	79	23	812	.72	.19	.66
Economics	93	14	582	.66	.20	.75
Finance	78	25	826	.70	.21	.68
Geography	104	13	817	.74	.17	.76
Health	84	17	788	.69	.21	.65
History	92	12	812	.74	.18	.77
Housekeeping	74	30	789	.75	.17	.65
Law	84	23	758	.67	.20	.63
Literature	96	55	644	.67	.21	.72
Mathematics[1]	169	21	914	.58	.18	.84
Medicine	75	31	768	.71	.20	.73
Music	78	22	714	.64	.20	.76
Nutrition	74	30	795	.70	.20	.63
Philosophy	84	17	805	.60	.20	.74
Physics	95	15	738	.67	.19	.79
Politics	85	19	594	.73	.16	.78
Psychology	96	19	827	.59	.20	.77
Religion	84	28	783	.77	.18	.70
Technology[2]	175	15	907	.65	.20	.83
Average	92	22	771	.68	.19	.72

Note. [1] Including also items from Statistics; [2] Including also items from Computer Science.

Hierarchical Structure of Declarative Knowledge

To explore the hierarchical structure of the dimensionality of declarative knowledge, we used a procedure proposed by Goldberg (2006). This method was mostly applied in personality research (e.g., Wright et al., 2012) and involves the estimation of a series of orthogonally rotated principal components analyses (PCA) with an increasing number of components[3]. Based on the number of factors reported in previous research, we modeled one to six components. The pattern matrices for all PCAs are displayed in the online supplement in Table S4 to S9. Factor loadings exceeding .30 were used for the interpretation of the rotated components. To examine relations across the different levels of the hierarchy, we correlated regression-based factor scores. We used orthogonal rotation because unrelated components allow computing unbiased cross-model correlations, which we used as paths between the different levels of the hierarchy. Figure 2 illustrates the hierarchical structure of declarative knowledge. We omitted the six-factor solution, because none of the indicators showed loadings >. 50 on the sixth factor and the explained variance was only 3%.

The one-component solution explained 42% of the variance, whereas the five-components solution explained 67% of the variance. In the one-component solution, all knowledge domains had factor loadings >.40 on a general dimension (mean = .64, min = .42, max =.78) indicating a strong general factor of declarative knowledge. With increasing differentiation, more differentiated dimensions took shape: From the second level on, a *Natural Sciences* factor—containing Mathematics, Physics, and Chemistry—was passed down continuously to the levels with higher dimensions. On the third level, a robust *Life Sciences* factor emerged—containing Medicine, Nutrition, and Housekeeping. On the fourth level, the broad *Social Sciences* factor split up in a *Humanities* factor (subsuming both artistic and cultural knowledge domains) and a *Social Studies* factor. On the last level, a factor with a less apparent

[3] Although we report PCAs we will use the term "factor" when interpreting components to aid readability

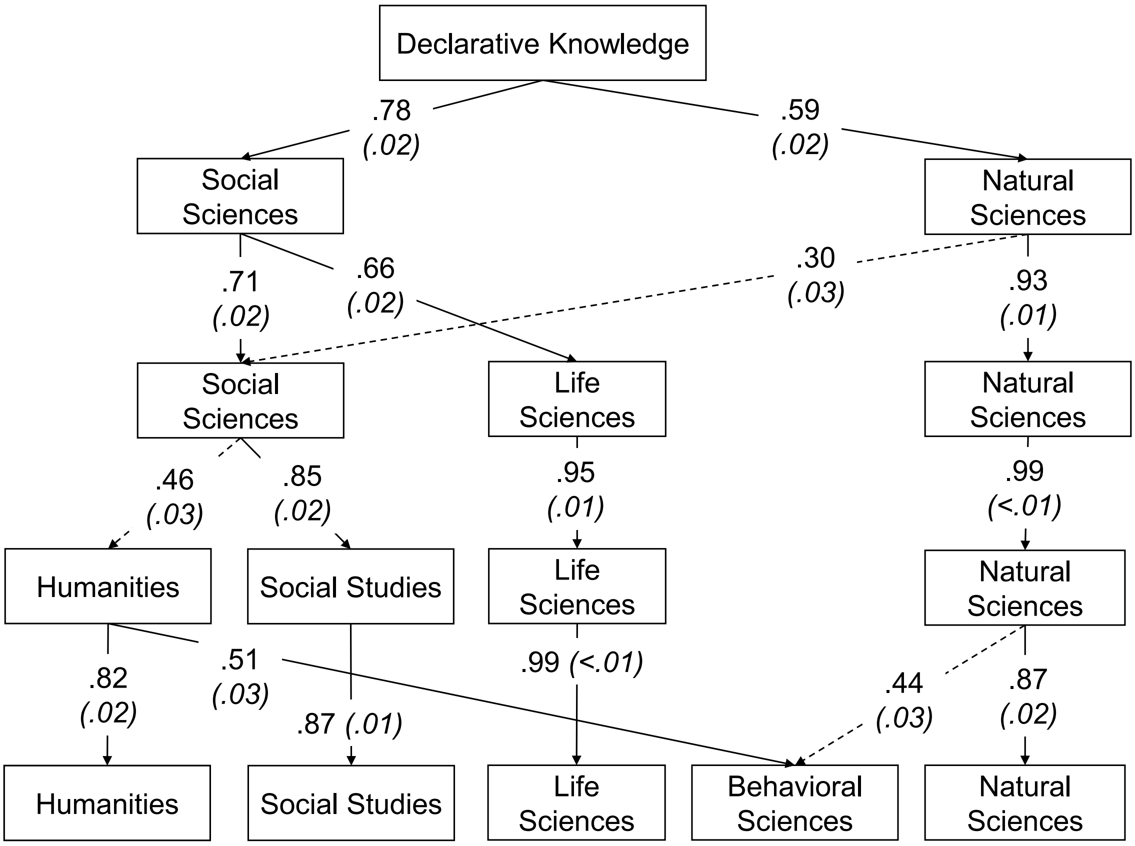

Figure III-2. Correlations between factor scores on the different levels of the dimensional hierarchy of declarative knowledge.

Note. Correlations below .30 were omitted. Correlations with values ≥ .30 are displayed as dashed lines, correlations ≥ .50 as solid lines. Standard errors are given in brackets.

loading pattern is formed from the *Humanities* and *Natural Science* factor. Since its most prominent loading is from the Psychology domain, we label this factor *Behavioral Science*. The findings discussed so far are based on PCA—an information reduction method (Preacher & MacCallum, 2003)—rather than exploratory factor analysis (EFA) which is suited to assess the dimensionality of psychological entities. Thus, to examine the dimensionality of declarative knowledge, we conducted an EFA with maximum likelihood estimation and oblique rotation. The number of factors to be extracted was *a priori* set to 5, as suggested by the hierarchical structure analysis and the parallel analysis (Horn, 1965). In total, the five factors explained 58% of the variance (see Table 3).

Table III-3. Five-Factor Solution with Obliquely Rotated Factor Loadings and Factor Correlations

Domain	Humanities	Social Studies	Life Sciences	Behavioral Sciences	Natural Sciences	h^2	u^2
Arts	**.75**	-.09	.13	.05	-.06	.60	.40
Architecture	**.70**	.14	.07	-.07	-.08	.63	.37
Literature	**.69**	-.07	.10	.11	.00	.56	.44
Religion	**.68**	-.03	.03	.04	.11	.55	.45
Music	**.63**	-.09	.12	.01	.07	.46	.54
History	**.61**	.27	-.06	-.01	.07	.64	.36
Anthropology	**.61**	.17	.00	.00	.08	.58	.42
Geography	**.60**	.15	.03	-.20	.23	.63	.37
Philosophy	**.56**	.06	-.06	**.47**	.00	.69	.31
Politics	**.46**	**.43**	-.04	.02	.10	.68	.32
Economics	.05	**.70**	.08	.19	.08	.71	.29
Finances	.20	**.51**	.28	-.08	-.10	.58	.42
Law	.21	**.49**	.09	.03	.02	.49	.51
Housekeeping	.02	.07	**.71**	-.12	-.12	.52	.48
Health	.02	-.03	**.68**	.07	.16	.56	.44
Medicine	.02	.04	**.66**	.11	.01	.51	.49
Nutrition	.11	.05	**.61**	.01	-.01	.49	.51
Psychology	.06	.14	.19	**.65**	.03	.61	.39
Physics	.08	.07	-.02	-.05	**.79**	.70	.30
Chemistry	.03	-.06	.06	.05	**.74**	.58	.42
Mathematics	.02	.07	-.12	**.49**	**.51**	.68	.32
Technology	.01	**.43**	.04	-.07	**.48**	.53	.47
Biology	.09	-.16	**.32**	.23	**.46**	.53	.47
Ecology	.13	.28	.29	-.01	**.30**	.51	.49
Explained Variance	.21	.10	.11	.06	.11		
Factor correlations							
	Humanities	Social Studies	Life Sciences	Behavioral Sciences			
Social Studies	.60						
Life Science	.55	.28					
Behav. Sci.	.28	.09	.15				
Natural Sci.	.39	.36	.18	.38			

Note. N = 1,117. h^2 = communalities, u^2 = uniqueness. Factor loadings $\geq$ .30 are displayed in bold.

Overall, the factors *Humanities*, *Social Studies*, *Life Sciences*, and *Natural Science* showed plausible loading patterns, which are consistent with classifications reported in previous studies, although no clear simple structure could be established. Some domains showed substantial and plausible cross-loadings, for example, Biology (loading on *Natural Sciences* and *Life Sciences*), or Politics (loading on *Humanities* and *Social Studies*). However, the loadings of Technology on both the *Social Studies* and the *Natural Science* was hard to reconcile with existing theories. The remaining factor, finally labelled *Behavioral Sciences*, was harder to interpret because only three domains showed substantial (i.e., > .30) loadings on the factor.

Given the exploratory nature of the study, we additionally conducted a two-dimensional multi-dimensional scaling and a hierarchical cluster analysis. The results closely mimic the results of the hierarchical PCA—although on different levels of the hierarchy. Whereas in the two-dimensional case, the MDS resembles the third hierarchical level with the factors *Social Sciences*, *Natural Sciences*, and *Life Sciences*, and the cluster analysis mirrors the hierarchical structure found by the hierarchical PCA. Visualizations of both methods are presented in Figure S1 and S2 in the online supplement. Interested readers can conduct further analyses with the data we provide online: https://osf.io/3s492/.

Discussion

Compared to its unquestionable relevance, *gc* can still be labeled the “dark matter” of intelligence research (Ackerman, 2000). Although widely used in different psychological disciplines such as cognitive psychology, educational psychology, *et cetera* there is neither a conclusive conceptual nor a stringent empirical classification of knowledge (Beauducel & Süß, 2011; Rolfhus & Ackerman, 1996). To tackle this question, we administered a knowledge test with 4,050 items covering a large range of domains via smartphone (Harari et al., 2016; Miller, 2012). Hence, this individual differences study comes close to meeting the original definition of *gc* sensu Cattell—“tasks indicating breadth and depth of the knowledge of the dominant culture” (Horn & Noll, 1997, p. 69). Besides the large item pool, electronically distributed measures are well-suited to attract heterogeneous samples, although this largely depends on recruitment strategies and target group orientated advertisement (Buhrmester, Kwang, & Gosling, 2011). Our main goal was to investigate the dimensionality of knowledge with a large item pool and a sample that was diverse in terms of age and educational background. Accordingly, we first discuss our findings and their impact on the search for the dimensionality of knowledge. Second, we discuss study characteristics and their impact on the dimensionality of knowledge and, lastly, we discuss different approaches for modeling crystallized intelligence.

Dimensionality of Declarative Knowledge

Models on the dimensionality of declarative knowledge hitherto reported vary broadly and differ in their labels and taxonomies. Wilhelm and colleagues (2014) reported that—despite their original conceptualization into *Natural Sciences*, *Humanities*, and *Social Studies*—already a parsimonious, unidimensional model described the data sufficiently well, on both an item- and a domain level (Schipolowski, 2014; Schipolowski, Schroeders, & Wilhelm, 2014). In the same vein, our results pointed to a strong general factor with all domains showing substantive loadings on a general *gc*-factor. The two-factor solution resembles the colloquial dichotomy into soft and hard sciences; domains can be distinguished according to their methodological rigor, quantitativity, and objectivity (Smith, Best, Stubbs, Johnston, & Bastiani Archibald, 2000). The same partitioning was reported by Hossiep and Schulte (2008) who assigned eleven content domains to two factors labeled *Social and Society Science* and *Natural and Technical Science*, analogous to the second level in the hierarchical PCA. On the third level, our results deviate from the sometimes used classification, that is, humanities, social studies, and natural sciences (Schipolowski, 2014; Schrank et al., 2014; Schroeders et al., 2015). Instead of a splitting the broad *Social Science* factor into *Humanities* and *Social Studies*, we interpret the third factor that emerges as a *Life Science* factor, because the subtests mostly cover health-related knowledge domains (Beier & Ackerman, 2003). On the fourth level of hierarchy, the *Natural Science* factor and the *Life Science* factor remained stable. Additionally, the broad Social Science factor from previous levels splits into *Humanities* and *Social Studies* which parallels the factors reported by Ackerman and colleagues (2001; but see also Ackerman, 2000; Rolfhus & Ackerman, 1996) who reported four factors labelled *Physical Science* (here: Natural Sciences), *Psychology/Biology* (Life Sciences), *Humanities*, and *Civics* (Social Studies). The findings most likely match because the measures assess knowledge with a comparatively broad item and domain sample. Finally, our results suggest a five-factor structure for the present set of domains: *Humanities*, *Social Studies*, *Life Sciences*, *Natural Sciences*, and *Behavioral Sciences*. While the first four factors were fairly stable with correlations >.80 with their corresponding level four factors, the *Behavioral Sciences* factor is a mix of *Humanities* and *Natural Sciences*. The emergence of *Behavioral Sciences* might be due to the fact that test development is “almost invariably expert selection rather than sampling” (Loevinger, 1965, p. 147; see also Schroeders, Wilhelm, & Olaru, 2016a).

Ultimately, we do not suggest that the five-factor solution is the final word in the debate on the dimensionality of knowledge. Rather the results of the hierarchical PCA point to the reason why models that are commonly discussed in the research literature are so different—because the answer depends on the hierarchical level on which data are gathered. Differently put, the numbers of factors found as well as their interpretation presumably depends on the compilation of domain sampling and item sampling within a given domain as well as person sampling. Thus, it is not surprising that previous studies found fewer factors when analyzing item or domain samples that focused on a more narrowly defined knowledge field. On the other hand, using even broader domain samples than the one used in the current investigation will most likely result in additional factors. Person sampling is another often-neglected determinant of the dimensionality of knowledge. Different factor structures will presumably manifest based on the distribution of age, educational background, and professional orientation. Thus, our results show that the seemingly competing models of declarative knowledge can be integrated in a hierarchical model.

Study Characteristics and Their Impact on the Dimensionality of Knowledge

We argued that the factor structure of knowledge varies as a function of person sampling, item sampling, and study design. In the present study we used an item sampling approach similar to the one in the SAPA project (Condon & Revelle, 2014). This approach implies massively missing data. From a person perspective, the data set includes both participants who completed only a small set of items and participants who responded to all items. It is still an open question, which role self-selection effects play, or—differently put—which person characteristics influence test-taking behavior. Since test-takers received non-monetary incentives for both persistent App usage and good performance, this might affect results because smarter people tend to persist (Mussel, 2013; von Stumm & Ackerman, 2013). Apart from motivational aspects that might influence the dimensionality of the measure, age and ability differentiation are often discussed as potential moderators and best studied together (Lienert & Crott, 1964). A method that is suitable to study such differentiation effects is the estimation of *Local Structural Equation Models* (Hildebrandt, Wilhelm, & Robitzsch, 2009), although recent investigations showed little evidence for age or ability differentiation (Hartung et al., 2018; Molenaar et al., 2010; Schroeders et al., 2015).

Further studies may investigate the impact of occupational specializations on the dimensionality of knowledge. In line with the assumptions of Cattell (1943, 1971) and also Ackerman's PPIK theory (1996), different occupational training should result in idiosyncratic knowledge profiles, leading to different factor structures at lower levels of the hierarchy. For example, one would expect more nuanced factors within the natural science for engineers, as their training mainly comprises technology, mathematics, and physics, whereas physicians arguably develop peaks in their profile within life sciences, as their training mainly comprises medicine, pharmacy, and biology. One possible approach to further investigate the impact of different professions on the dimensionality of knowledge would be to compare the knowledge structure across occupational groups directly.

Furthermore, in the present analysis, our item sample consisted of items that were closer to general knowledge as conceptualized in the *gc* factor of the CHC Theory (McGrew, 2009) rather than the more specialized expert knowledge that could be subsumed under the *gkn* factor. Although we also included domains that are not part of a typical school curriculum—such as psychology, statistics, or architecture—many domains were curriculum-based. Similarly, in the present analysis, we did not include avocational knowledge. Consequently, we might expect different trajectories for avocational and vocational knowledge (Ackerman, 1996), or even current events knowledge (Beier & Ackerman, 2001). Please note that—based both on theoretical and data-driven assumptions—in the present study we analyzed a subset of 2,210 items that met all inclusion criteria.

Lastly, in the present study we used a smartphone-based assessment. Online assessment is often perceived as flexible as it allows administering tests at any time and any geographic region, and it can be conducted without the presence of a research team or a lab. By making the test easily accessible to a larger population via Internet or Smartphone App, it seems possible to reduce the sampling bias relative to traditional lab assessments (Hays, Liu, & Kapteyn, 2015; Henrich et al., 2010). Although online assessments have occasionally been praised as cure against WEIRD samples (Gosling, Sandy, John, & Potter, 2010), we dampen such expectations because highly educated individuals might still be overrepresented due to self-selection. This was also the case in the present sample, despite our genuine attempts to appeal to a heterogeneous sample. Another caveat in smartphone-based assessments is the lack of supervision which results in less control over test-takers' behavior. This is problematic especially for tests of maximal performance (Do, 2009), as participants might use the opportunity to engage in cheating behavior. A recent meta-analysis (Steger, Schroeders, & Gnambs, 2019) found that while differences between a proctored and an unproctored assessment are negligible for tasks that assess fluid abilities, differences exists for tasks that assess crystallized abilities. That is, cheating is easier, more instrumental, and prevalent for knowledge tasks—even in low stakes testing and despite countermeasures. To this end, it might be useful to rely on incidental data such as reaction times or changing the focus of the App to detect cheating behavior *a posteriori* (Couper, 2005; Diedenhofen & Musch, 2017). Future research should seek to determine a set of methods best suited to identify cheaters in online ability assessments.

The Nature of Crystallized Intelligence

These ideas about assessing and modeling crystallized intelligence as declarative knowledge entail questions about the very nature of *gc*. The way *gc* is conceptualized is not as straightforward as one might think; different perspectives on *gc* have been brought forward: In most consensual theories of intelligence *gc* is modelled as a causal entity (i.e., using a reflective model; Borsboom, Mellenbergh, & van Heerden, 2003; Van der Maas, Kan, & Borsboom, 2014). In contrast, the principal components model is a *formative* model, in which component scores are merely composites of the observed data. The role of *gc* in this theoretical framework is controversial: It is often implicitly or explicitly assumed that *gc* is a causal entity that represents a psychological capacity, although not all scholars would follow this interpretation (Kan, Kievit, Dolan, & der Maas, 2011). If we define *gc* as a latent variable in the sense of a causal entity, we assume that there is a causal relationship between the latent variable and the indicators. This means that there is one common cause (*gc*) that influences participants responses on all knowledge items, best modelled using a *reflective* model (i.e., latent variable model; Van der Maas et al., 2014). However, this view was also challenged by some scholars. Van der Maas and colleagues (2014, p. 13) pointed out that using factor models to model intelligence is like "cutting butter with a razor", because factor models are just a very complicated way of deriving results that are in fact nothing more than a weighted sum score. They suggested the use of formative models, arguing that there is no common causal entity that influences all test scores, but a mutualism model in which the positive manifold is explained by interactions between cognitive processes during their development. These relations may be best depicted using network models (van der Maas, Kan, Marsman, & Stevenson, 2017). The authors argue that this approach seems promising especially for the integration of both theories on the development of cognitive abilities and intelligence into one unified framework. But also formative models of intelligence were not spared from criticism (see Bollen & Diamantopoulos, 2017, for further discussion). To this end, Schipolowski and colleagues (2015) interpreted knowledge assessments in the realm of *behavior domain theory* (Markus & Borsboom, 2013), in which item responses are seen as samples from a given behavior domain. In this case, the common factor represents a *behavior domain score* (McDonald, 2003) and not necessarily a causal entity.

Finally, the choice of statistical modeling technique has some impact on the results and their interpretation. Usually, using confirmatory factor analyses relies on simple loading structure which is recommended to unambiguously interpret the solution (Preacher & MacCallum, 2003). In practice, for knowledge assessment, it is difficult (or even impossible) to create measures with broad indicators that closely adhere to simple structure at item and domain levels. This is also evident from the available results: A considerable amount of items and domains show substantial cross-loadings. Ultimately, this questions the presupposition that the majority of items can be unequivocally assigned to single domain. For example, the knowledge domain *Geography* could be assigned to the broader factor *Sciences* and *Humanities*—depending on its compilation: Sub-dimensions such as geology or geodesy might fit best to a *Natural Sciences* factor, while sub-dimensions that cover more cultural or historical aspects of Geography might go best with a *Humanities factor.* Especially, when item samples are very broad and include more aspects of broad knowledge domains, the assignment to only one factor is inconclusive (see also *Behavioral Science*).

The same problem also applies to the item level: Some items might be classifiable to different domains. For example, the question "When did Wolfgang Amadeus Mozart die?" could be equally assigned to *Music* or *History*, which results in fuzziness rather than clear categories. Other statistical procedures could be suitable to model more complex knowledge structures: For example, Q-matrices could be implemented (Rupp & Templin, 2008) to model the relationship between latent variables and individual items, but such matrices are prone to misspecification (Kunina-Habenicht, Rupp, & Wilhelm, 2012) and highly subjective if not backed up by theory. The same holds true for Bayesian approaches to structural equation models (Merkle & Rosseel, 2018), where one needs a solid amount of prior knowledge to specify priors on model parameter estimates. Exploratory structural equation models (Asparouhov & Muthén, 2009), in turn, don't need the theoretical foundation because they are truly explorative in nature, but may leave one with a substantive amount of cross-loadings that render it difficult to postulate a true dimensionality of knowledge. Taken together, different ways of modeling *gc* may yield different results, which might contribute to further understanding the nature of *gc*.

Acknowledgements

We thank Christopher Kannen for developing the IQ App and implementing the Quiz. We would like to thank Florian Schmitz and Christian Montag for their contribution in conceptualizing and promoting the App. Furthermore, we thank all student assistants, colleagues, and experts that helped developing, editing, and proofreading items and the *Bamberg Graduate School of Social Sciences* for their advertising campaign.

References

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227–257. https://doi.org/10.1016/S0160-2896(96)90016-1

Ackerman, P. L. (2000). Domain-specific knowledge as the "dark matter" of adult Intelligence: Gf/gc, personality and interest correlates. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 55*, 69–84. https://doi.org/10.1093/geronb/55.2.P69

Ackerman, P. L., Bowen, K. R., Beier, M., & Kanfer, R. (2001). Determinants of individual differences and gender differences in knowledge. *Journal of Educational Psychology, 93*, 797–825. https://doi.org/10.1037//0022-0663.93.4.797

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R Manual [Manual of the Intelligence Structure Test 2000 R]*. Göttingen: Hogrefe.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 397–438. https://doi.org/10.1080/10705510903008204

Bartholomew, D. J. (2004). *Measuring intelligence: facts and fallacies.* Cambridge, UK; New York: Cambridge University Press.

Beauducel, A., & Süß, H.-M. (2011). Wissensdiagnostik: Allgemeine und spezielle Wissenstests. In L. F. Hornke, M. Amelang, & M. Kersting (Eds.), *Serie II, Psychologische Diagnostik* (Vol. Methodologie und Methoden, pp. 235–273). Göttingen: Hogrefe.

Beier, M. E., & Ackerman, P. L. (2001). Current-events knowledge in adults: An investigation of age, intelligence, and nonability determinants. *Psychology and Aging, 16*, 615–628. https://doi.org/10.1037//0882-7974.16.4.615

Beier, M. E., & Ackerman, P. L. (2003). Determinants of health knowledge: An investigation of age, gender, abilities, personality, and interests. *Journal of Personality and Social Psychology, 84*, 439–448. https://doi.org/10.1037/0022-3514.84.2.439

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods, 22*, 581–596. https://doi.org/10.1037/met0000056

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5. https://doi.org/10.1177/1745691610393980

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing [Abstract]. *Psychological Bulletin, 38*, 592.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*, 153.

Cattell, R. B. (1971). *Abilities: Their Structure, Growth, and Action.* Boston: Houghton Mifflin.

Center for Open Science. (2017). https://cos.io/ [Last access September 29, 2017].

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*, 486–501. https://doi.org/10.1177/0894439305278972

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification" (pp. 1–7). Presented at the

MindTrek'11, Tampere, Finland: ACM Press. https://doi.org/10.1145/2181037.2181040

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

Do, B.-R. (2009). Research on unproctored internet testing. *Industrial and Organizational Psychology, 2*, 49–51. https://doi.org/10.1111/j.1754-9434.2008.01107.x

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., ... Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PLoS ONE, 6*, 1–3. https://doi.org/10.1371/journal.pone.0024974

Engelberg, P. M. (2015). *Ursachen fÃ¼r Geschlechterdifferenzen in Tests des Allgemeinen Wissens [Causes for gender differences in general knowledge tests]* (Doctoral Dissertation). University of Wuppertal, Wuppertal, Germany.

Goldberg, L. R. (2006). Doing it all bass-ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality, 40*, 347–358. https://doi.org/10.1016/j.jrp.2006.01.001

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences, 33*, 94–95. https://doi.org/10.1017/S0140525X10000300

Hamari, J. (2017). Do badges increase user activity? A field experiment on the effects of gamification. *Computers in Human Behavior, 71*, 469–478. https://doi.org/10.1016/j.chb.2015.03.036

Hambrick, D. Z., Pink, J. E., Meinz, E. J., Pettibone, J. C., & Oswald, F. L. (2008). The roles of ability, personality, and interests in acquiring current events knowledge: A longitudinal study. *Intelligence, 36*, 261–278. https://doi.org/10.1016/j.intell.2007.06.004

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*, 838–854. https://doi.org/10.1177/1745691616650285

Hartung, J., Doebler, P., Schroeders, U., & Wilhelm, O. (2018). Dedifferentiation and differentiation of intelligence in adults across age and years of education. *Intelligence, 69*, 37–49. https://doi.org/10.1016/j.intell.2018.04.003

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–83. https://doi.org/10.1017/S0140525X0999152X

Hildebrandt, A., Wilhelm, O., & Robitzsch, A. (2009). Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology, 16*, 87–102.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179–185. https://doi.org/10.1007/BF02289447

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporal intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York: The Guildford Press.

Hossiep, R., & Schulte, M. (2008). *Bochumer Wissenstest. Manual.* Göttingen: Hogrefe.

Irwing, P., Cammock, T., & Lynn, R. (2001). Some evidence for the existence of a general factor of semantic memory and its components. *Personality and Individual Differences, 30*, 857–871.

Kan, K.-J., Kievit, R. A., Dolan, C., & der Maas, H. van. (2011). On the interpretation of the CHC factor Gc. *Intelligence, 39*, 292–302. https://doi.org/10.1016/j.intell.2011.05.003

Kievit, R. A., Fuhrmann, D., Borgeest, G. S., Simpson-Kent, I. L., & Henson, R. N. (2018). The neural determinants of age-related changes in fluid intelligence: a pre-registered, longitudinal analysis in UK Biobank. *Wellcome Open Research, 38*, 1–11.

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education, 87*, 340–356. https://doi.org/10.1016/j.compedu.2015.07.009

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models: Detection of model misspecification in DCMs. *Journal of Educational Measurement, 49*, 59–81. https://doi.org/10.1111/j.1745-3984.2011.00160.x

Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science, 15*, 155–163.

Lienert, G. A., & Crott, H. W. (1964). Studies on the factor structure of intelligence in children, adolescents, and adults. *Human Development, 7*, 147–163. https://doi.org/10.1159/000270108

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review, 72*, 143–155. https://doi.org/10.1037/h0021704

Lord, F. M. (1955). Estimating Test Reliability. *Educational and Psychological Measurement, 15*, 325–336.

Mariani, C., Sacco, L., Spinnler, H., & Venneri, A. (2002). General Knowledge of the World: a standardised assessment. *Neurological Sciences, 23*, 161–175.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity Theory:*

Measurement, Causation, and Meaning. New York: Routledge.

McDonald, R. P. (2003). Behavior domains in theory and practice. *The Alberta Journal of Educational Research, 49*, 212–230.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software, 85.* https://doi.org/10.18637/jss.v085.i04

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7*, 221–237. https://doi.org/10.1177/1745691612441215

Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Maas, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence, 38*, 611–624. https://doi.org/10.1016/j.intell.2010.09.002

Mussel, P. (2013). Intellect: A theoretical framework for personality traits related to intellectual achievements. *Journal of Personality and Social Psychology, 104*, 885–906. https://doi.org/10.1037/a0031918

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 1420–1422. https://doi.org/10.1126/science.aab2374

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*, 13–43. https://doi.org/10.1207/S15328031US0201_02

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2017). Web and phone based data collection using planned missing designs. In N. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE Handbook of Online Research Methods.* Los Angeles, CA: SAGE.

Riediger, M., Wrzus, C., & Wagner, G. G. (2014). Happiness is pleasant, or is it?

Implicit representations of affect valence are associated with contrahedonic motivation and mixed affect in daily life. *Emotion, 14*, 950–961. https://doi.org/10.1037/a0037711

Roberts, R. D., Goff, G. N., Anjoul, F., Kyllonen, P. C., Pallier, G., & Stankov, L. (2000). The Armed Services Vocational Aptitude Battery (ASVAB): Little more than acculturated learning (Gc)!? *Learning and Individual Differences, 12*, 81–103.

Rolfhus, E. L., & Ackerman, P. L. (1996). Self-report knowledge: At the crossroads of ability, interest, and personality. *Journal of Educational Psychology, 88*, 174–188.

Rolfhus, E. L., & Ackerman, P. L. (1999). Assessing individual differences in knowledge: Knowledge, intelligence and related traits. *Journal of Educational Psychology, 91*, 511–526.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research & Perspective, 6*, 219–262. https://doi.org/10.1080/15366360802490866

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior, 69*, 371–380. https://doi.org/10.1016/j.chb.2016.12.033

Schipolowski, S. (2009). *Struktur und Messung kristalliner Intelligenz [Structure and measurement of crystallized intelligence]* (Diplomarbeit). Humboldt-Universität, Berlin.

Schipolowski, S. (2014). *Zur Struktur, Messung und Entwicklung kristalliner Intelligenz [On the structure, measurement and development of crystallized intelligence]* (Doctoral Dissertation). Humboldt-Universität, Berlin.

Schipolowski, S., Schroeders, U., & Wilhelm, O. (2014). Pitfalls and challenges in constructing short forms of cognitive ability measures. *Journal of Individual*

Differences, 35, 190–200. https://doi.org/10.1027/1614-0001/a000134

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2015). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence, 46*, 156–168. https://doi.org/10.1016/j.intell.2014.05.014

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV Tests of Achievement.* Rolling Meadows, IL: Riverside.

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2015). Age-related changes in the mean and covariance structure of fluid and crystallized intelligence in childhood and adolescence. *Intelligence, 48*, 15–29. https://doi.org/10.1016/j.intell.2014.10.006

Schroeders, U., Wilhelm, O., & Olaru, G. (2016a). Meta-Heuristics in Short Scale Construction: Ant Colony Optimization and Genetic Algorithm. *PloS One, 11*, 1–19.

Schroeders, U., Wilhelm, O., & Olaru, G. (2016b). The influence of item sampling on sex differences in knowledge tests. *Intelligence, 58*, 22–32. https://doi.org/10.1016/j.intell.2016.06.003

Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., & Bastiani Archibald, A. (2000). Scientific graphs and the hierarchy of science: A Latourian Survey of Inscription Practices. *Social Studies of Science, 30*, 73–94.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Steger, D., Schroeders, U., & Gnambs, T. (2019). A meta-analysis of test scores in proctored and unproctored ability assessment. *European Journal of Psychological Assessment.* https://doi.org/10.1027/1015-5759/a000494

Van der Maas, H., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*, 842–861. https://doi.org/10.1037/0033-295X.113.4.842

Van der Maas, H., Kan, K.-J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence, 2*, 12–15. https://doi.org/10.3390/jintelligence2010012

Van der Maas, H., Kan, K.-J., Marsman, M., & Stevenson, C. E. (2017). Network models for cognitive development and intelligence. *Journal of Intelligence, 5*, 1–17. https://doi.org/10.3390/jintelligence5020016

von Stumm, S., & Ackerman, P. L. (2013). Investment and intellect: A review and meta-analysis. *Psychological Bulletin, 139*, 841–869. https://doi.org/10.1037/a0030746

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. https://doi.org/10.1007/BF02294627

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). Seattle: Hogrefe & Huber.

Wilhelm, O., Schroeders, U., & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz fÃ¼r die 8. bis 10. Jahrgangsstufe (BEFKI 8-10).* Göttingen: Hogrefe.

Wright, A. G. C., Thomas, K. M., Hopwood, C. J., Markon, K. E., Pincus, A. L., & Krueger, R. F. (2012). The hierarchical structure of DSM-5 pathological personality traits. *Journal of Abnormal Psychology, 121*, 951–957. https://doi.org/10.1037/a0027669

Wrzus, C., Wagner, G. G., & Riediger, M. (2014). Feeling good when sleeping in? Day-to-day associations between sleep duration and affective well-being differ from youth to old age. *Emotion, 14*, 624–628. https://doi.org/10.1037/a0035349

IV. Manuscript 3

Caught in the Act: Predicting Cheating in Unproctored Knowledge Assessment

Diana Steger

Ulm University, Germany

Ulrich Schroeders

University of Kassel, Germany

Oliver Wilhelm

Ulm University, Germany

Status

Steger, D., Schroeders, U., & Wilhelm, O. (in press). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment*.

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

Introduction

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 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, Wilhelm, & Schipolowski, 2010; Sliwinski et al., 2016) 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, Olderbak, Sauter, & Wilhelm, 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, Rosseel, & Lombardi, 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 in proctored versus unproctored ability tests (Steger, Schroeders, & Gnambs, 2019). 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 Multiphasic Personality Inventory (Hathaway & McKinley, 1943), followed by other instruments such as the Sixteen Personality Factor Questionnaire in 1949 (Cattell, Eber, & Tatsuoka, 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.”).

While these lie scales are best applied to detect faking on self-report scales, the Honesty-Humility factor of the HEXACO model (Ashton & Lee, 2007, 2008) has been linked successfully with cheating and other dishonest behavior (Heck, Thielmann, Moshagen, & Hilbig, 2018). In general, self-reports of Honesty-Humility seem to be valid under low-stakes condition (Ashton, Lee, & De Vries, 2014; Zettler, Lang, Hülsheger, & Hilbig, 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 non-existent items (Paulhus, Harms, Bruce, & Lysy, 2003; Phillips & Clancy, 1972). Overclaiming can be assessed by juxtaposing familiarity ratings for a list of items consisting of existing terms (reals) and non-existing 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, Kluemper, Kristl Davison, Taylor, & Novicevic, 2011; but see also Müller & Moshagen, 2019a, 2019b). 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, Hilbig, & Zettler, 2018). In contrast to social desirability, overclaiming does not seem to be confounded with personality or intelligence measures as such (Bensch, Paulhus, Stankov, & Ziegler, 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 easily be 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 to 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, Do, Drasgow, & Fine, 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, Diedenhofen, Verschuere, & Musch, 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, Meijer, & Tendeiro, 2016). However, in a comprehensive simulation study, Karabatsos (2003) evaluated the performance of 36 different person-fit indices and found that cheating—as compared to 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 to 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, Bucholtz, Formazin, & Wilhelm, 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 that 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 for not focusing on the task at hand. In their study, participants worked on an online knowledge 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 it 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, Schroeders, & Gnambs, 2019), 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. Lastly, 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, Moshagen, & Zettler, 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, König, Jung, Sorg, & Ziegler, 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, Oud, & Dam, 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 large, multi-centered 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 that 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 respect to the cheating instruction and the gift card was distributed amongst all participants at random. The time period between online and lab session varied between one days and three 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, Schroeders, & Gnambs, 2019). 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 (see also Table S1 in the online supplement). Questions were sampled from a larger item pool of multiple choice items (Steger, Schroeders, & Wilhelm, 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. Self-report data. First, to assess cheating-related personality traits, we used the German 60-item version of the HEXACO (Moshagen, Hilbig, & Zettler, 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, Ashton, & de Vries, 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 $\alpha = .70$. 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 non-existing (*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 6 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, Wilhelm, & Schipolowski, 2011). As expected, mean familiarity of all *foils* was low, ranging from 1.13 to 2.70 ($M = 1.55$, $SD = 0.43$) compared to 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 ($\alpha = .90$). 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 was 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 that 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 s) in the lab condition and 9.86 seconds (SD = 3.2 s) 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 non-compliance 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* (FIML) 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 ten 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. Lastly, 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, Schroeders, & Wilhelm, 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 latent change score 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 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 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.

Table IV-1. Descriptive Statistics of the Knowledge Tests, *S*-data, *T*-data, and *P*-Data indicators.

	Descriptives							Correlations									
	N	*M*	*SD*	min	max	Skewness	Kurtosis	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Declarative knowledge																	
(1) Total Score (Onl)	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 (Onl)	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 (Onl)	312	.34	.19	0.06	.97	1.45	1.55	.56	.00	.63	-.03	.15					
(7) Difficult Items (Lab)	310	.23	.07	0.06	.47	0.34	0.29	-.02	.00	-.02	-.05	-.02	-.02				
Para Data																	
(8) Flagged RTs (Onl)	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 (Onl)	312	20.70	29.62	0	127	1.63	1.90	.56	-.05	.66	.01	.12	.76	-.02	.59	.06	
(11) Defocusing (Lab)	310	0.09	0.09	0	5	7.61	79.23	-.11	-.05	-.08	.02	-.07	-.02	-.05	.05	.20	-.04

Note. For declarative knowledge scales and difficult items, we report the percentage correct answers; for defocusing events and reaction times, 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 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. 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 inter-individual 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 cut-off scores; for example, see also Hair, Anderson, Tatham, & Black, 1995; Neter, Wassermann, & Kutner, 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).

Table IV-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			.01			.40			.53
ΔR^2						.39			.13
AIC			-478.18			-628.23			-702.70
BIC			-463.30			-609.63			-676.66

Note. * $p < .05$

To complement the analyses, we also computed a latent change score model, which we also extended by several variables to predict the latent change score. 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 non-normality in the data, all models were estimated using a Maximum Likelihood estimator with robust standard errors which is also suited for non-normally distributed indicators (Gao et al., 2019). Next, we estimated the latent change score model (Figure S3 in the online supplement), which fits the data well (N = 315, χ^2 = 69.02, *df* = 42, $p <$ .01, CFI = .97, RMSEA = .05, SRMR = .05). The negative correlation (ρ = -.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 latent change score 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, χ^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.

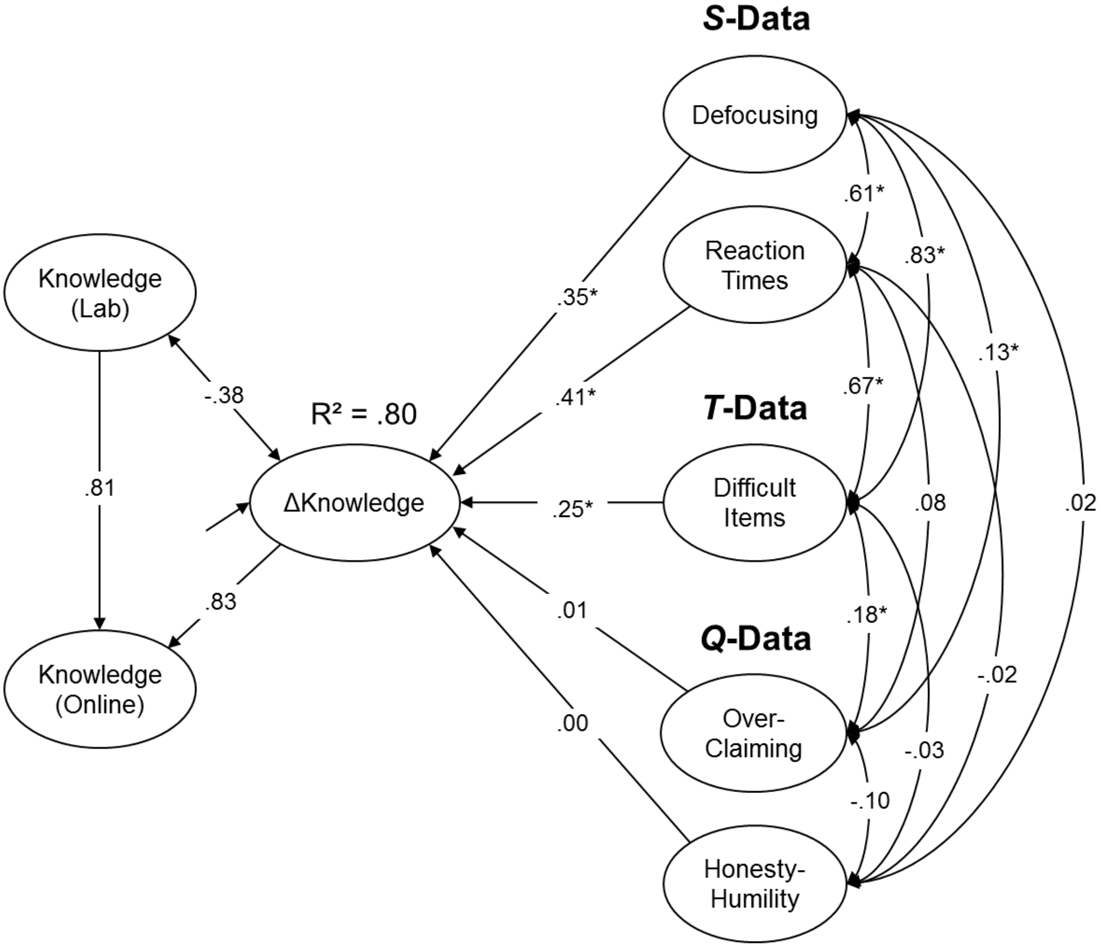

Figure IV-1. Extended latent change model. Indicators, residual correlations, and correlations between predictors and the proctored knowledge score were omitted for readability. A complete overview over correlations between latent factors can be found in Table S5 in the online supplement. * p < .05

In the extended latent change score model, all but one predictors are uncorrelated with the lab knowledge score (Table S5 in the online supplement). The only exception is reaction time, which correlates negatively (ρ = -.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, Stavropoulos, Jaeggi, & Seitz, 2018; Stieger, Lewetz, & Reips, 2018), online panels (Hays, Liu, & Kapteyn, 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., 2016; 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, Marsch, McHugo, & Bartels, 2015 for an overview). However, initial enthusiasm about these new data sources was rapidly followed by critical concerns about data quality (e.g., Aust, Diedenhofen, Ullrich, & Musch, 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 paper, 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 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, Schroeders, & Gnambs, 2019). 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 paper 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, 2019a, 2019b). In the same vein, overclaiming did not contribute substantially to predicting cheating—neither on a manifest nor on a latent level. Considering past effort put into the scale construction of various self-report instruments, these results are sobering. 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, Kuhn, & Holling, 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, Atkinson, Clark, & Lowell, 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, Kincaid, Smith, & Trawick, 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, Weiss, & Wilhelm, 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 latent change score. 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, but can also 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 latent change score model, difficult items, response times, and defocusing events significantly predicted the latent change score, explaining 80% of its variance. Response times have been linked to faking behavior in self-report assessments (Maricuțoiu & 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 responses 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, Kröhne, Goldhammer, Schreiber, & Brefeld, 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 upon 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 gold standard to prevent cheating behavior. However, cheating can also occur and succeed in proctored testing (Drasgow, Nye, Guo, & Tay, 2009). But how can we determine if someone cheated in unproctored settings? Cheating is only directly observable using supervision, sometimes in form of screen monitoring or webcam surveillance (Karim, Kaminsky, & Behrend, 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, Diedenhofen, Verschuere, & Musch, 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, Musch, & Erdfelder, 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 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 non-cheaters 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 participant differ in the anticipated costs and utility for the participants (Thielmann & Hilbig, 2018) and also 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 because it also 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, Schroeders, & Gnambs, 2019) 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 estimate the extent to which the data is 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 does not only apply 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.

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*, 150–166. https://doi.org/10.1177/1088868306294907

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*, 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*, 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*, 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*, 21–29.

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). 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*, 1–14.

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.* Minneapolis, MN: University of Minnesota Press.

Carpenter, R. W., Wycoff, A. M., & Trull, T. J. (2016). Ambulatory assessment: New adventures in characterizing dynamic processes. *Assessment, 23*, 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.* Champaign, IL: Institute for Personality and Ability Testing.

Chatterjee, S., & Price, B. (1991). *Regression diagnostics.* New York: John Wiley.

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

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*, 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*, 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*, 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*, 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*, 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*, 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*, 629–634. https://doi.org/10.1016/j.paid.2011.05.033

Gao, C., Shi, D., & Maydeu-Olivares, A. (2019). 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*. 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*, 1–15. 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*, 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*, 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*, 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*, 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.). New York: 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*, 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.* Minneapolis, MN: 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*, 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*, 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*, 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*, 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*, 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*, 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*, 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*, 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*, 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, 53*, 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*, 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.* New York: Appleton-Century-Crofts.

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

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 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*, 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*, 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. (2019a). True virtue, self-presentation, or both?: A behavioral test of impression management and overclaiming. *Psychological Assessment, 31*, 181–191. https://doi.org/10.1037/pas0000657

Müller, S., & Moshagen, M. (2019b). Controlling for response bias in self-ratings of personality: A comparison of impression management scales and the overclaiming technique. *Journal of Personality Assessment, 101*, 229–236.

https://doi.org/10.1080/00223891.2018.1451870

Murdock, T. B., & Anderman, E. M. (2006). Motivational perspectives on student cheating: Toward an integrated model of academic dishonesty. *Educational Psychologist, 41*, 129–145.

Neter, J., Wassermann, W., & Kutner, M. H. (1989). *Applied Linear Regression Models.* Homewood, IL: 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*, 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., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 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*, 112–120. https://doi.org/10.1080/09639284.2011.590012

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving Surveys with Paradata* (pp. 43–72). 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*, 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.* 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*, 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*, 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*, 2153–2162. https://doi.org/10.1177/0956797613487221

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.1). Retrieved from 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.* https://doi.org/10.1027/1015-5759/a000511

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*, 1–36.

Rovai, A. P. (2000). Online and traditional assessments: What is the difference? *The Internet and Higher Education, 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*, 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). Washington, D.C.: American Psychological Association.

Sliwinski, M. J., Mogle, J. A., Hyun, J., Munoz, E., Smyth, J. M., & Lipton, R. B. (2016). Reliability and validity of ambulatory cognitive assessments. *Assessment, 25*, 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*, 1133–1152.

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

Steger, D., Schroeders, U., & Gnambs, T. (2019). A meta-analysis of test scores in proctored and unproctored ability assessments. European Journal of Psychological Assessment. https://doi.org/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*, 185–209. https://doi.org/10.1002/per.676

Stowell, J. R., & Bennett, D. (2010). Effects of online testing on student exam performance and test anxiety. *Journal of Educational Computing Research, 42*, 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*, 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*, 49–54.

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). Seattle: Hogrefe & Huber.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational*

Behavior and Human Decision Processes, 115, 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*. 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*, 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*, 461–472. https://doi.org/10.1111/jopy.12172

V. Epilogue

Summary

In the following section, I give an overview of the main findings of the three manuscripts included in this thesis. Taken together, the manuscripts explore the central question of how technology-based assessment can be applied to the measurement of cognitive abilities. Manuscript 1 focuses on the impact of unproctored test settings on the assessment of cognitive abilities in general. In manuscript 2, technology-based assessments are applied to measure declarative knowledge—a psychological construct that is generally hard to measure with traditional means—and in manuscript 3, methods to assess data quality in unproctored knowledge assessments are examined.

Manuscript 1: A meta-analysis of test scores in proctored and unproctored ability assessments

The central question of the first manuscript concerned the impact of unproctored test environments on ability test scores. To this end, we conducted a three-level random effects meta-analysis (Cheung, 2014): First, we examined mean score differences between proctored and unproctored assessment to investigate if test scores are on average higher in one of the two settings, using a pool of 109 effect sizes from 49 studies (total N = 100,434). We then examined the impact of potential moderators—that is, a) the perceived consequences of the assessment, b) countermeasures taken against cheating, c) the susceptibility to cheating of the measure itself, and d) the use of different test media—on the test score differences between proctored and unproctored settings. Lastly, because the comparison of mean scores does not warrant conclusions about the equivalence of two measurements (AERA, APA, & NCME, 2014), we also used a smaller pool of five studies that also reported correlations between test scores of proctored and unproctored ability tests and thus allowed us to explore rank order stability. Due to the comparatively small study pool, we did not pursue further moderator analyses.

In the analysis of mean score differences between proctored and unproctored settings, we found a weighted mean effect size of $\Delta = 0.20$, 95% CI [0.10, 0.31], indicating that participants achieved slightly higher mean scores in unproctored tests. Our explorations of potential publication bias (i.e., comparison of effect sizes from published vs. unpublished sources, funnel plot analyses, and a rank correlation test) did not indicate any evidence for potential publication bias. We conducted a mixed-effects regression analysis to quantify the influence of moderators on the pooled effects. Our analysis revealed a significant effect only of the *searchability* of the measure. That is, the pooled mean difference is significantly smaller for measures that are difficult to research on the Internet (e.g., tasks that measure mainly fluid abilities). Lastly, we identified a pooled correlation of $\rho = .58$ ($SE = .19$), 95% CI [.38, .78], indicating substantial rank order changes for proctored versus unproctored assessments.

With the first manuscript, we provided a comprehensive overview of existing findings on the effect of proctored and unproctored test environments. Overall, we found a small but significant effect, which indicated that participants, on average, achieved higher test scores in unproctored test environments. Together with the finding that this effect was most pronounced for measures that are especially easy to look up on the Internet (e.g., knowledge tests), results suggest that unproctored assessments are biased by cheating. Our results suggest further that the conduction of low stakes assessments and the use of countermeasures against cheating are insufficient to prevent fraud, indicating that at least some participants will always cheat if they have the opportunity—regardless of countermeasures or anticipated consequences. Conversely, one could also say that participants will not cheat if they are not given the opportunity. Based on these findings, we recommended designing unproctored assessments carefully (e.g., select tests with low *searchability*) and interpreting findings from unproctored assessments with caution. In case of tests with a high searchability, as for example declarative knowledge tests, *post hoc* strategies can be applied to identify potential cheaters (e.g., Diedenhofen & Musch, 2017) to secure data quality.

Manuscript 2: On the dimensionality of crystallized intelligence: A smartphone-based assessment

With the second manuscript, our aim was to investigate the dimensionality of crystallized intelligence using a mobile quiz app that allowed us to administer a large set of knowledge items to a heterogeneous sample. More specifically, knowledge questions were drawn randomly from a pool of 4050 items from 34 subject domains. For the present manuscript, we analyzed a data set with 1117 participants (58% female; $M_{\text{age}} = 36.6$ years, $SD_{\text{age}} = 15.3$ years) who met the inclusion criterion (i.e., having answered at least 15 items in at least two domains) collected from October 2016 to February 2018. To investigate the dimensionality of knowledge, we used a subset of 25 domains (excluding domains that covered mostly current events knowledge, see also Beier & Ackerman, 2001; Hambrick, Pink, Meinz, Pettibone, & Oswald, 2008; and domains with poor psychometric properties) and computed *Weighted Likelihood Estimates* (Warm, 1989) based on two-parameter logistic item response models separately for each domain. Using the domain scores, we subsequently explored the hierarchical factor structure of declarative knowledge using the *bass-ackwards method* (Goldberg, 2006), in which a series of orthogonally rotated principal component analyses (PCA) are conducted.

The hierarchical principal components analyses yielded well-interpretable results for the extraction of one to five components. At the first level, the one-component solution explained 42% of the variance and all knowledge domains loaded substantially on the general dimension with mean factor loadings of .64 (min = .42, max = .78). Throughout the hierarchical levels, the factor structure evolved mirroring results from previous studies on the dimensionality of declarative knowledge on various levels and resulting in the five components *Humanities*, *Social Studies*, *Life Sciences*, *Behavioral Sciences*, and *Natural Sciences*, which explained 67% of the variance. Because principal components analyses are a mere information reduction method (Preacher & MacCallum, 2003), we additionally conducted an exploratory factor analysis to further examine the dimensionality of declarative knowledge. As suggested by both the hierarchical structure analysis and a parallel analysis (Horn, 1965), we extracted five factors, which explained 58% of the variance. Overall, the factors *Humanities*, *Social Studies*, *Life Sciences*, and *Natural Sciences* showed plausible loading patterns, while the remaining factor of *Behavioral Sciences* was harder to interpret, because only Philopsophy, Psychology, and Mathematics showed substantial loadings (i.e., $\lambda \geq .30$) on this factor. As a last step, we additionally conducted a multi-dimensional scaling and a hierarchical cluster analysis to check the robustness of the results. Despite representing different levels of the hierarchy, the results from the additional analyses correspond closely the results of the hierarchical PCA.

Although previously presented models on the dimensionality of declarative knowledge offered various factor structures, labels, and taxonomies, the present study allows for a conciliatory conclusion: There is not the one true model that depicts the truth about the dimensionality of declarative knowledge. In fact, the results of the hierarchical structure analysis explain why the models commonly reported in the literature are so different: Results depend on the hierarchical level on which the data are gathered (i.e., domain sampling and item sampling within these domains) and sample characteristics (e.g., age or educational background). Consequently, it is not surprising that previous studies found fewer factors when analyzing more narrow item or domain samples. On the other hand, using even broader domain samples than the one used in the current investigation will most likely result in more robust factors (e.g., strengthening the *Behavioral Sciences* factor) or even reveal additional factors. Similarly, person sampling is another often-neglected determinant of the dimensionality of knowledge. Different factor structures will presumably manifest based on the distribution of age (especially when comparing children or adolescent samples to adult samples) or educational background. Taken together, our results demonstrate how the seemingly competing models of declarative knowledge can be integrated into a broad, hierarchical model.

Manuscript 3: Caught in the act: Predicting cheating in unproctored knowledge assessments

In the third manuscript, we addressed the question of how we can secure data quality in unproctored knowledge assessment. To this end, we investigated the potential of different data types—that is, self-report data (from questionnaires), test data (from ability tests), and para data (incidental data from technology-based assessments; Couper, 2005; Kroehne & Goldhammer, 2018)—to identify potential cheaters in unproctored knowledge tests. We analyzed data from 315 participants (71.7% female; $M_{\text{age}} = 25.5$ years, $SD_{\text{age}} = 7.8$ years) who worked in a within-person design on two parallel knowledge tests with 102 items each, once in an unproctored online setting and once in a proctored lab setting. Knowledge tests covered a broad range of declarative knowledge and were sampled from a larger pool of multiple choice items (Steger, Schroeders, & Wilhelm, 2019). To increase the propensity of cheating, participants were told that everyone who answers 80% or more of the questions correctly participates in a lottery. As self-report data, we used the honesty-humility scale of the German version of the HEXACO-60 (Moshagen, Hilbig, & Zettler, 2014) and a newly-developed overclaiming questionnaire. As test data, we analyzed participants' performance on 34 knowledge items that were designed to be virtually impossible to solve and were presented with the actual knowledge test. Lastly, as para data, we recorded response times and the occurrence of browser tab switches using a JavaScript similar to PageFocus (Diedenhofen & Musch, 2017). To investigate the potential of different indicators to predict cheating behavior, we first conducted a hierarchical regression analysis and subsequently an extended *latent change score model* (McArdle, 2009) in which the change score reflected the difference between lab and online assessment.

First, our results suggest that cheating did occur in the unproctored assessment: Overall, participants achieved higher knowledge scores in the unproctored assessment ($M = .66$, $SD = .12$) as compared to the proctored assessment ($M = .57$, $SD = .10$), with 12% of the sample achieving score differences so high that they could be flagged as potential cheaters. In the hierarchical regression analysis, both test and para data significantly predicted cheating, explaining 53% of the variance in score differences between unproctored and proctored knowledge tests. The extended latent change score model replicated these findings: Again, both test and para data predictors significantly predicted cheating, explaining 80% of the variance in the latent change score variable.

Taken together, test and para data appear to be valid predictors of cheating in declarative knowledge tasks, while self-report data was not able to contribute significantly to predicting cheating. However, neither conspicuously high scores on difficult items, nor prolonged response times, nor browser switch times can be directly equated with cheating: We simply do not know what participants are doing when leaving the test page or taking long to produce an answer. Additionally, the difference score we used as an indicator of cheating does not equal cheating directly; other factors such as declining motivation during a longer lab session or reduced test anxiety in unproctored test settings (Stowell & Bennett, 2010) might have influenced the results as well. Furthermore, for different types of ability tests, other indicators might help to identify cheaters correctly. With this manuscript, we provide an approach that can help to assess data quality and to develop a transparent method to identify and exclude potential cheaters from analyses in unproctored knowledge assessment.

Technology-based Assessment

In the following section, I discuss the findings of the three manuscripts in the light of existing literature on technology-based assessment. In more detail, I consider four main issues connected with technology-based assessment in general: a) the effects of the test setting, b) person sampling and sample characteristics, c) item sampling, and d) the use of auxiliary data. Lastly, I address main caveats of technology-based assessment, that is, ethical concerns that arise with these modern assessment techniques, concerns about data quality, and the need for new methodological approaches.

Unproctored Assessments

When conducting unproctored online assessments, cheating was widely discussed to be a problem (Tippins, 2009; Tippins et al., 2006) and it was questioned whether results from psychological instruments administered online can be trusted. Results on the central questions on whether cheating actually biases test scores in unproctored testing were mixed (Do, 2009). To this end, we provided a systematic review on the impact of test environment on the participants' test scores. We found that the main influencing factor for score differences between proctored and unproctored assessments were characteristics of the applied measure itself (namely, whether questions are easy to search on the Internet), regardless of countermeasures taken against cheating, incentive structure, or test medium. Differently put, it is likely that participants will cheat when they are given the opportunity to do so and, conversely, will not cheat if we take this opportunity from them. Accordingly, a straightforward recommendation for ability assessments in unproctored settings is the development of test batteries that are limited to measures of fluid abilities, which go along with tasks that are hard to look up on the Internet.

Contrarily, when using tasks that are easy to look up on the Internet in an unproctored environment, the results clearly call for caution. Therefore, we examined the potential of self-report data, test data, and para data to predict cheating in an unproctored knowledge assessment, finding that it is possible to identify cheaters in an online knowledge task using test and para data predictors. With the application of *post hoc* strategies to control for data quality, it is also possible to conduct unproctored assessments of constructs with a high searchability. However, despite the promising results, it remains unclear to which extent it is also possible to detect cheating "in the wild". Future research must examine whether it is possible to use the same indicators to detect cheaters in smartphone-based assessment or which other indicators could be derived in smartphone-based assessment to be included in models for cheating detection.

Furthermore, this result illustrates the usefulness of para data indicators derived from technology-based assessment, because these indicators are less prone to faking than traditional self-report instruments and offer a more direct measurement of human behavior. Although self-report instruments are widely used in psychological research, these instruments are especially prone to biases due to memory effects, self-deception, or deliberate faking (Schwarz, 2012). On the contrary, para data indicators are obtained unobtrusively, since they come simply as a “by-product” of technology-based assessment (Couper, 2005), making it almost impossible for participants to deliberately fake these measures.

Sample Characteristics

Online- or smartphone-based assessment strategies were widely advertised to overcome biased lab samples by targeting a more heterogeneous audience (Gosling & Mason, 2015; Harari et al., 2016), that is in particular to overcome biases that arise from investigating mainly traditional samples from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich, Heine, & Norenzayan, 2010). In many psychological studies, this bias might be even worse, since they often rely on samples of Psychology students (Gosling, Vazire, Srivastava, & John, 2004; Sears, 1986)—that is, samples of highly educated, young females with above-average cognitive abilities—simply because this group is most easily accessible for the traditional lab studies. To overcome the issue of lacking representativity, we used a mobile quiz app to address a sample that is more heterogeneous with regard to age, gender, and educational background than the traditional lab sample—with mixed success: Generally, our sample was balanced regarding age and gender, but participants with academic qualifications were still overrepresented. Participation in smartphone studies hinges on different factors, such as the access to the technological device and the motivation to search and use this kind of apps. These factors decrease the probability that certain subgroups (e.g., old adults or lower educated classes) participate in these studies, resulting in biased results.

But this might not be the only source of bias in online samples. Generally, it seems true that, using online samples, it is easy to reach large groups of people quickly (e.g., Revelle et al., 2017), but large dropout rates have to be anticipated (Seifert, Hofer, & Allemand, 2018): For example, Figure 1 shows a plot similar to a survival plot (but with the number of answered items rather than time points), illustrating the relation between number of participants and number of knowledge questions answered. A large amount of participants only answered a few items, resulting in a considerably smaller sample of people who actually answered enough items to produce reliable results. Since data from online assessments is potentially more noisy and complex than data from traditional data sources, data processing leaves room for many *researcher's degrees of freedom* (Simmons, Nelson, & Simonsohn, 2011) and due to the lack of standardized data cleaning procedures, analyzing these data is often very exploratory in nature. Accordingly, this calls for the need of transparent reporting and open science practices (e.g., sharing data and code). Sensitivity analyses could also be conducted using specification curve analyses (Simonsohn, Simmons, & Nelson, 2015), which can help to assess the robustness of results.

On an exploratory base, I investigated the impact of participant attrition in the mobile quiz data, that is, to what extend groups of participants that differ in their persistence yield different results. In the following, I present the results from a multi-group confirmatory factor analysis, using the number of items answered in total by a participant as a proxy for persistence. Three groups were created, depending on the number of items participants completed: low persistence ($<$ 1000 items), medium persistence (between 1000 and 2000 items), and high persistence ($>$ 2000 items). Declarative knowledge was modeled using a bifactor model with an overarching factor for general knowledge, *humanities* as a reference factor, and *arts*, *social studies*, *life sciences*, and *physical sciences* as nested factors.

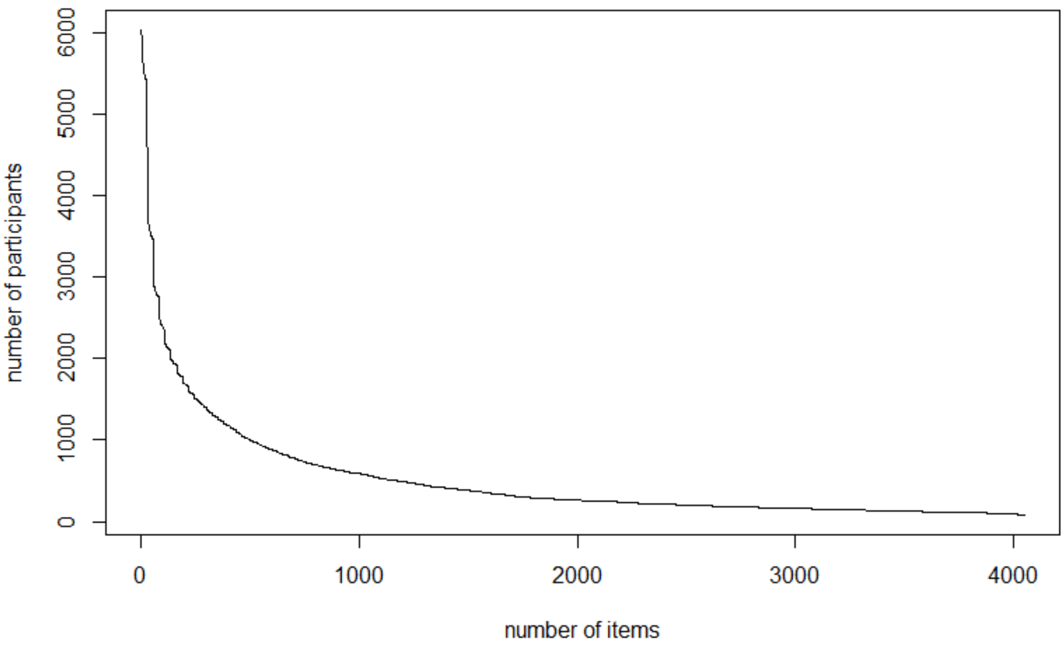

Figure V-1. “Survival plot” illustrating the number of participants depending on the number of items they answered. Total N = 6039.

Results show that scalar measurement invariance holds (CFI = .93, RMSEA = .05) and therefore factor means can be compared across groups. The standardized mean differences for the general knowledge factor are .16 (low vs. medium persistence) and .27 (low vs. high persistence) and the standardized mean differences of the nested factors range between -.10 and .18 (see Table 1). A model with fixed factor means across groups did not obtain a significantly worse fit (CFI = .93, RMSEA = .05), indicating that differences in factor means across groups are negligible. Furthermore, factor saturation was fairly consistent across groups, with a strong general knowledge factor and lower factor saturation of the nested factors. Thus, the dimensionality of declarative knowledge does not seem to be affected by participants’ motivation.

Although these insights into effects of participant attrition seem promising, we still do not know enough about other sources of bias to quantify the risk of biases in the sample. Especially, in the assessment of declarative knowledge, a quiz app might attract people with distinct patterns of investment traits (Mussel, 2013; von Stumm

Table V-1. Factor Means and Factor Saturation across Groups.

	Persistence		
	Low	Medium	High
N	664	202	201
Factor means			
General Knowledge	.00	.16	.27
Arts	.00	-.01	.14
Physical Sciences	.00	-.03	-.10
Social Sciences	.00	.18	-.08
Life Sciences	.00	.09	.15
Factor Saturation			
General Knowledge	.93	.79	.84
Arts	.13	.14	.16
Physical Sciences	.40	.42	.39
Social Sciences	.17	.15	.17
Life Sciences	.40	.40	.56

& Ackerman, 2013). In general, although mobile data collection strategies might be more flexible than traditional lab assessments, we advise caution because data quality must be monitored closely. However, our results also indicate that samples collected with smartphone-based assessments can offer an alternative to the traditional lab samples, as they are more balanced in terms of gender and cover a broader age span.

Item Sampling

Technology-based assessment is said not only to facilitate the collection of data from large samples, but also to make data collection more flexible. Novel data collection strategies allow measuring subjects across multiple time points using intensive longitudinal designs (Bolger & Laurenceau, 2013), enriching traditional test- and self-report data with additional behavioral data (Harari et al., 2016), and using more flexible item sampling such as the *Synthetic Aperture Personality Assessment* (SAPA, Condon & Revelle, 2014; Revelle et al., 2017). This way we are able to gain new insights into well-known constructs: For example, we can collect data that allows us to examine dynamical processes and within-person variability (Zimmermann et al., 2019) and we are also able to test broad and multi-faceted constructs in comparatively short time (Revelle et al., 2017).

When measuring declarative knowledge, testing a broad item and domain sample is of particular interest (Ackerman, 1996). Thus, in the mobile quiz app, we used an approach similar to the SAPA approach, where participants answered sets of items that were randomly drawn from a large pool of knowledge items. Since every participant can answer as many questions as he or she likes, risk of fatigue effects or lacking participant motivation is reduced, but this approach entails data that is *Massively Missing Completely at Random* (Revelle et al., 2017). However, with the use of synthetic covariance matrices as proposed by Revelle and colleagues (2017), it also possible to easily derive results on factor level, thereby allowing the evaluation of broadly measured constructs without the side-effects of hour-long test sessions. Accordingly, the use of modern item sampling strategies makes it possible to validate a broad item pool from various knowledge domains, which subsequently can be used as basis for the compilation of novel instruments. However, with massive missingness (up to 95% missings per item) at item level, the feasibility of some item-level analyses is limited. Similarly, the usability for individual diagnostics is mixed, since measurement precision largely depends on the number of items a person answered.

Collecting Auxiliary Data

The use of technology-based assessment allows the collection of a multitude of auxiliary data including para data (Couper, 2005; Kroehne & Goldhammer, 2018), such as response times or log data, or mobile sensing data (Harari et al., 2016), providing direct information about participants' behavior (e.g., Stachl et al., 2019). These data types provide us with immediate access to human behavior rather than indirectly measuring behavior using self-reports. For example, mobile sensing data such as data on phone calls and text messages can be used to predict social behavior patterns in young adults (Harari et al., 2019). This way, we have the possibility to measure behavior without the danger of response distortions such as memory effects or social desirability and lessen the chance of deliberate faking, as these data are collected incidentally and unobtrusively.

In this thesis, para data was used to predict cheating (a behavior that is otherwise hard to assess)—using response times and browser tab switches as a direct measure of cheating behavior—and ultimately, to control data quality of unproctored assessment *post hoc*. Additionally, in cognitive ability assessment, auxiliary data yield additional diagnostic information about the participants, complementing the information that we get from traditional questionnaire or test data: For example, we can derive information on which participants are more likely to cheat, which participants are more able to cheat, or under which circumstances participants will cheat more frequently. Furthermore, para data can be useful to augment all forms of psychological measures (Kroehne & Goldhammer, 2018), such as large-scale assessments or surveys (Kreuter, 2013), for example by identifying careless responding, non-compliant participants, or bots (Buchanan & Scofield, 2018). Differently put, these data can be used to learn more about the test environment and test-taking behavior of participants when test settings are unstandardized and therefore a potential source of measurement error exists (Stieger & Reips, 2010).

However, when working with auxiliary data, it is vital to transform this data into "psychological meaningful" variables (Seifert et al., 2018, p. 2). The more complex the data gets, the more important sophisticated data transformation techniques become. For example, when using sensory input in smartphone-based assessment, data processing might require the transformation of GPS signals into mobility patterns (e.g., Harari, Gosling, Wang, & Campbell, 2015) or data on smartphone-usage into day-night behavior patterns (Schoedel, 2019). Especially in this field, interdisciplinary projects combining expertise from the field of computer science, data science, and psychology are needed to develop reliable and valid indicators for human behavior that can eventually be merged with other types of data collection (Seifert et al., 2018). If, for example, mouse clicks should be used as an indicator for test-taking behavior, the raw data needs to be aggregated and transformed into a meaningful behavior indicator, reflecting for example careless responding. On such fine-grained level, simply no psychological theories exist on how these complex behaviors translate into psychological constructs or interact with each other. As we simply do not know enough about these complex data patterns, this also shifts the focus from confirmatory research to approaches that are more exploratory in nature.

Caveats of Technology-based Assessment

It becomes clear that with technology-based assessment, we have a tool that facilitates data assessment in many ways: It becomes easier to address large and diverse samples (Gosling, Sandy, John, & Potter, 2010), it becomes easier to collect a vast amount of data without overstraining participants' motivaton (Condon & Revelle, 2014), and it is possible to collect additional information about participants' behavior apart from traditional self-reports or test instruments (e.g., using sensory data, Harari et al., 2016; or para data, Kroehne & Goldhammer, 2018). However, technology-based assessments also have their limitations, which we need to consider when implementing them in psychological research.

First, with the rise of technology-based assessment, concerns were raised about data security and ethical use of data. From a researcher's perspective, the unobtrusiveness with which we can now assess auxiliary data is promising, but participants need to be informed what data is collected and how these data is used. Especially in studies that record many additional information, this could either influence participants' behavior when they are aware of all information that is tracked (e.g., not visiting specific places when they know that their GPS information is tracked), or discourage participants from enrolling in the study altogether. Furthermore, when we use smartphone technology to record a large amount of personal information this results in highly individualized profiles that render anonymization almost impossible (see also Seifert et al., 2018). Accordingly, new data and privacy models have been developed that are tailored to the challenges of technology-based assessment to ensure that ethically sound research is possible (e.g., Beierle et al., 2018; Zook et al., 2017). Adherence to this principles needs to become imperative to maintain participants' privacy.

Second, technology-based assessment allows us to assess broad constructs, to take the intra-individual perspective, and to collect a large amount of auxiliary information. All these possibilities result in large data sets that can clearly be described as *Big Data*. Such data sets are characterized by high dimensionality and a large sample size, and both features entail their own challenges, leading to noise accumulation, additional measurement error, and potential biases (Fan, Han, & Liu, 2014). Additionally, when working with data obtained from smartphone sensors or other external data sources, we must assure that these sensors provide valid signals in the first place (Harari et al., 2015): If these data are noisy, indicators derived from these data become invalid. In addition, with invalid indicators, no valid results can be obtained or conclusions be drawn. Taken together, we must carefully consider all aspects of data quality when working with these types of data sets.

Lastly, when analyzing such large data sets, these analyses oftentimes require high computational power and the use of adequate statistical methods especially suited for this data type (Fan et al., 2014). Many traditional statistical methods that are well-suited for conventional data sets (with moderate sample sizes and moderate number of variables) do not perform well when used on big data. Accordingly, we need to develop new tools that are especially suited to address the issues that arise with these data sets, as for example machine learning approaches (e.g., Bleidorn & Hopwood, 2019), which can be used to identify associations between digital footprints and established psychological constructs. Although this approach has also been criticized as “atheoretical”, it is arguably difficult to actually generate appropriate theories for this amount of high-dimensional and fine-grained data. Ultimately, it has been argued that psychological research should also put an increased focus on predictive research (using for example machine learning), rather than relying on explanatory research as a gold-standard (Yarkoni & Westfall, 2017), shifting from a confirmatory approach to more exploratory research.

Improving the Measurement of Declarative Knowledge

In the following section, I discuss the present findings in the light of existing literature on declarative knowledge and outline ideas for future research. More specifically, I focus on implications of the current findings for traditional assessment of declarative knowledge and pending questions on the measurement of declarative knowledge.

Traditional Assessments of Declarative Knowledge

According to Cattell (1943), crystallized intelligence is a broad construct, enclosing a broad set of skills, knowledge, and language-related abilities, with declarative knowledge as the central indicator. This conceptualization entails the question of how to measure declarative knowledge. A broad knowledge assessment should ideally include “the whole variety of [culturally valued] knowledge that people *can* acquire during their lives” (Wilhelm & Schroeders, 2019, p. 264). Even after an extensive literature search—which resulted in the identification of 34 knowledge domains—the knowledge domains used in our study are likely just a fraction of all knowledge domains one could possibly think of. A thorough assessment of declarative knowledge has previously been described as complicated to conduct: If the test battery was carefully designed and truly tested a broad range of different knowledge domains with a sufficient amount of items to fully depict the realm of declarative knowledge (Ackerman, 1996; Wilhelm & Schroeders, 2019), the extensive test sessions were usually long and tedious. If shorter test batteries were applied, item selection carries the risk of unbalanced or biased item samples, which can lead to biased results. For example, although a male advantage in knowledge test is widely reported in the literature (e.g., Ackerman, Bowen, Beier, & Kanfer, 2001; Lynn, Irwing, & Cammock, 2002), sex differences in knowledge test performance mainly hinge on the content of the specific item sample (Schroeders, Wilhelm, & Olaru, 2016b). On a more general stance, knowledge test batteries that lead to a male advantage are likely biased and neglect knowledge domains in which females usually outperform males (e.g., humanities or health-related subjects).

Selection effects become also prevalent on domain-level when investigating the dimensionality of declarative knowledge. Also in our study, the domains we included in the study vary with regard to their specificity, ranging from very broad and general domains (e.g., biology, or arts) to more specialized domains (e.g., statistics, or fashion). The dimensionality of knowledge largely depends on the selection of knowledge domains, and the more domains we include in our measurement and the more fine-grained the domains are, the more fine-grained the taxonomy will become. Accordingly, the taxonomy we provide in our analysis is not the final solution to the dimensionality of knowledge—rather, it only depicts a section of it. Our results demonstrate that previous work on the dimensionality of declarative knowledge can be integrated into a taxonomy of declarative knowledge—despite the seemingly competing models reported in the literature. Effectively, the different tests measure different sections of declarative knowledge and on different levels of the taxonomy.

Roadmap to an Improved Measurement of Declarative Knowledge

Generally, the use of a mobile quiz app has proven to be a promising approach to make a broad and comprehensive assessment of declarative knowledge possible and to yield the potential to improve knowledge assessment. However, results from comprehensive mobile assessment of declarative knowledge can also contribute to other substantial pending questions on declarative knowledge.

First, to date there are no conclusive results to which extent age might influence the dimensionality of knowledge. Following Cattel's (1971) notion that in adult years, knowledge becomes highly idiosyncratic, it seems very likely that knowledge structures differentiate with age, leading to a differentiation of the factor structure of knowledge over time (e.g., Baltes, Staudinger, & Lindenberger, 1999). However, although widely cited and investigated, results concerning the differentiation hypotheses suggest that reported differentiation effects more likely depict statistical artifacts rather than depicting true effects (Hartung, Doebler, Schroeders, & Wilhelm, 2018; D. Molenaar, Dolan, Wicherts, & van der Maas, 2010). But most research on the differentiation effects investigate mostly fluid abilities, excluding possible differentiation effects of declarative knowledge. A study that investigated age-related differentiation of both fluid and crystallized intelligence for children and adolescents was published by Schroeders, Schipolowski, and Wilhelm (2015). The authors found little evidence of age-related differentiation for the different reasoning facets and knowledge domains, suggesting that the relatively homogeneous scholastic learning environment in secondary education prevents the development of more pronounced ability or knowledge profiles. In the case of declarative knowledge, this might be due to the fact that, according to Cattell (1971), a structural differentiation of crystallized intelligence takes place *after school*. Based on a comprehensive, smartphone-based assessment of declarative knowledge, it would be possible to investigate differentiation over a broad age range, potentially depicting the unfolding of idiosyncratic knowledge profiles that develop after the individuals leave the standardized learning environments.

Second, also not only the selection of knowledge domains might play a role for the dimensionality of knowledge, but also the composition of domains: In our study, a considerable amount of items and domains show substantial cross-loadings. Ultimately, this questions the presupposition that the majority of items can be unequivocally assigned to single domain. For example, the question “When did Wolfgang Amadeus Mozart die?” could be equally assigned to *Music* or *History*, which results in fuzziness rather than clear categories—violating the condition of a simple structure of a factor, which is necessary to unambiguously interpret the solution (Preacher & MacCallum, 2003). Accordingly, future studies should also focus on the impact of item selection on the dimensionality of knowledge. For example, meta-heuristics could be used to derive “pure” domain scales by minimizing item cross-loadings with other domains.

Especially when compiling new knowledge scales, item selection strategies (so-called meta-heuristics; Olaru, Schroeders, Hartung, & Wilhelm, 2019; Schroeders, Wilhelm, & Olaru, 2016a) can be used to derive short forms that adhere to certain pre-specified criteria and to overcome issues caused by expert selection (Loevinger, 1965). These pre-specified criteria could be optimized with regard to model fit, predictive validity, dimensional structure, or measurement invariance across groups (e.g., gender, or educational or cultural background). Our smartphone-based assessment of declarative knowledge provides an optimal basis for this approach, because it helps collecting information on a large item pool that can be used to derive short forms to also aid the assessment of declarative knowledge in more traditional settings.

Lastly, for decades, experimental and correlational research have been the dominating paradigms in psychological research (Borsboom, Kievit, Cervone, & Hood, 2009), mainly focusing on between-person variations rather than within-person variation (Voelkle, Brose, Schmiedek, & Lindenberger, 2014). However, with the advancement of modern assessment technologies, we have the perfect tools at hand to conduct intensive longitudinal designs (Bolger & Laurenceau, 2013) in the framework of ambulatory assessment (Carpenter, Wycoff, & Trull, 2016; Trull & Ebner-Priemer, 2013) and therefore to gain insight into intra-individual dynamics and processes (Wright & Zimmermann, 2019; Zimmermann et al., 2019).

Also for the assessment of cognitive abilities, this intraindividual perspective might be benefical, since, to date, basically all investigations on cognitive abilities have been based on correlational studies that focus on between-person variations rather than within-person performance fluctuations (Schmiedek, Lövdén, von Oertzen, & Lindenberger, 2019). Yet, we cannot simply assume a close correspondence between within-person and between person structures (Molenaar, Huizenga, & Nesselroade, 2003), leading to the necessity to study cognitive abilities also from the within-person perspective. For example, Könen, Dirk, and Schmiedek (2015) investigated sleep patterns and their impact on the performance of cognitive ability tasks. The four-week long investigation of sleep patterns and working memory task performance in 110 elementary school children revealed not only group effects (i.e., the days children reported better sleep quality were also on average the days on which children performed better in cognitive tasks), but also effects at an individual level. That is, individuals differed in the strength of association between sleep quality and cognitive performance, or—differently put—sleep quality is not as important for everyone.

The intraindividual perspective becomes especially important when developing and evaluating interventions (Schmiedek & Neubauer, 2019). Taking the within-person perspective, we do not only learn how effective an intervention is *on average*, but rather which inter- and intraindividual conditions may influence the effectiveness of a specific treatment. In the realm of cognitive assessment, this could potentially be used to more thoroughly evaluate learning curves for specific educational trainings and help students to benefit the most from different learning strategies. In this case, a smartphone application could provide the environment in which students could log their training sessions, monitor their progress, and also record their test results. This way, the application could not only serve as a data collection tool for psychological research, but based on the results, the app could also use the input to provide direct feedback to the student.

Conclusion

The present thesis provided a first insight into the possibilities of modern technology-based assessment for advances in the measurement of cognitive abilities. To this end, we explored the advantages and disadvantages of online ability assessments in general and of smartphone-based assessment of declarative knowledge in particular. Taken together, modern assessment technologies in general offer great opportunities to complement traditional psychological assessment. With its flexibility and omnipresence in people's everyday life, it allows researchers to collect data that would be otherwise hard to obtain: direct measurements of behavior, data on within-person variations and intrapersonal dynamics, and data from large and heterogeneous samples that would be otherwise hard to target.

Despite all advantages, along with the development of new assessment technologies, we also have to rethink and refine how we do research. We need to develop new (or adjusted) theories that also allow us to take the within-person perspective and take into account real-life behavior (see also Seifert et al., 2018), develop advanced statistical methods for data processing and analysis that meet the requirements of the complex data, and reflect carefully when interpreting or evaluating results. Additionally, although smartphone-based assessment now attracts more and more attention from researchers from all scientific fields, the history of technology-based assessment teaches us that we should not stop to look out for new techniques and approaches to data collection, because smartphone-based assessment will certainly not be the final step in the evolution of technology-based assessment. In general, technology-based assessment is a field that offers great potential for further research—undoubtedly, Francis Galton would be excited.

References

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227–257. https://doi.org/10.1016/S0160-2896(96)90016-1

Ackerman, P. L., Bowen, K. R., Beier, M., & Kanfer, R. (2001). Determinants of individual differences and gender differences in knowledge. *Journal of Educational Psychology, 93*, 797–825. https://doi.org/10.1037//0022-0663.93.4.797

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology, 50*, 471–507. https://doi.org/doi.org/10.1146/annurev.psych.50.1.471

Beier, M. E., & Ackerman, P. L. (2001). Current-events knowledge in adults: An investigation of age, intelligence, and nonability determinants. *Psychology and Aging, 16*, 615–628. https://doi.org/10.1037//0882-7974.16.4.615

Beierle, F., Tran, V. T., Allemand, M., Neff, P., Schlee, W., Probst, T., ... Zimmermann, J. (2018). Context data categories and privacy model for mobile data collection apps. *Procedia Computer Science, 134*, 18–25. https://doi.org/10.1016/j.procs.2018.07.139

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Science Review, 23*, 190–203. https://doi.org/10.1177/1088868318772990

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research.* New York, NY: Guilford Press.

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or the disunity of psychology as a working hypothesis. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudhary (Eds.), *Dynamic Process Methodology in the Social and Developmental Sciences* (pp. 67–97). Retrieved from http://link.springer.com/10.1007/978-0-387-95922-1_4

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

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

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*, 153–193. http://dx.doi.org/10.1037/h0059973

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*, 211–229. https://doi.org/10.1037/a0032968

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*, 486–501. https://doi.org/10.1177/0894439305278972

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

Do, B.-R. (2009). Research on unproctored internet testing. *Industrial and Organizational Psychology, 2*, 49–51. https://doi.org/10.1111/j.1754-9434.2008.01107.x

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review, 1*, 293–314. https://doi.org/10.1093/nsr/nwt032

Goldberg, L. R. (2006). Doing it all bass-ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality, 40*, 347–358. https://doi.org/10.1016/j.jrp.2006.01.001

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences, 33*, 94–95. https://doi.org/10.1017/S0140525X10000300

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59*, 93–104. https://doi.org/10.1037/0003-066X.59.2.93

Hambrick, D. Z., Pink, J. E., Meinz, E. J., Pettibone, J. C., & Oswald, F. L. (2008). The roles of ability, personality, and interests in acquiring current events knowledge: A longitudinal study. *Intelligence, 36*, 261–278. https://doi.org/10.1016/j.intell.2007.06.004

Harari, G. M., Gosling, S. D., Wang, R., & Campbell, A. T. (2015). Capturing Situational Information with Smartphones and Mobile Sensing Methods: Capturing situations with smartphone sensing. *European Journal of Personality, 29*, 509–511. https://doi.org/10.1002/per.2032

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*, 838–854. https://doi.org/10.1177/1745691616650285

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., ... Gosling, S. D. (2019). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology.* https://doi.org/10.1037/pspp0000245

Hartung, J., Doebler, P., Schroeders, U., & Wilhelm, O. (2018). Dedifferentiation and differentiation of intelligence in adults across age and years of education. *Intelligence, 69*, 37–49. https://doi.org/10.1016/j.intell.2018.04.003

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–83. https://doi.org/10.1017/S0140525X0999152X

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179–185. https://doi.org/10.1007/BF02289447

Könen, T., Dirk, J., & Schmiedek, F. (2015). Cognitive benefits of last night's sleep: Daily variations in children's sleep behavior are related to working memory fluctuations. *Journal of Child Psychology and Psychiatry, 56*, 171–182. https://doi.org/10.1111/jcpp.12296

Kreuter, F. (2013). Improving surveys with paradata: Introduction. In F. Kreuter (Ed.), *Improving Surveys with Paradata: Analytic Uses of Process Information* (pp. 1–9). Hoboken, New Jersey: Wiley.

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

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review, 72*, 143–155. https://doi.org/10.1037/h0021704

Lynn, R., Irwing, P., & Cammock, T. (2002). Sex differences in general knowledge. *Intelligence, 30*, 27–39. https://doi.org/10.1016/S0160-2896(01)00064-2

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

Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Maas, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence, 38*, 611–624. https://doi.org/10.1016/j.intell.2010.09.002

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The Relationship Between the Structure of Interindividual and Intraindividual Variability: A Theoretical and Empirical Vindication of Developmental Systems Theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding Human Development* (pp. 339–360). https://doi.org/10.1007/978-1-4615-0357-6_15

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*, 86–97. https://doi.org/10.1026/0012-1924/a000112

Mussel, P. (2013). Intellect: A theoretical framework for personality traits related to intellectual achievements. *Journal of Personality and Social Psychology, 104*, 885–906. https://doi.org/10.1037/a0031918

Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (2019). Ant Colony Optimization and Local Weighted Structural Equation Modeling. A tutorial on novel item and person sampling procedures for personality research. *European Journal of Personality.* https://doi.org/10.1002/per.2195

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*, 13–43. https://doi.org/10.1207/S15328031US0201_02

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2017). Web and phone based data collection using planned missing designs. In N. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE Handbook of Online Research Methods*. Los Angeles, CA: SAGE.

Schmiedek, F., Lövdén, M., von Oertzen, T., & Lindenberger, U. (2019). *Within-person structures of daily cognitive performance cannot be inferred from between-person structures of cognitive abilities* [Preprint]. https://doi.org/10.7287/peerj.preprints.27576v1

Schmiedek, F., & Neubauer, A. B. (2019). Experiments in the Wild: Introducing the Within-Person Encouragement Design. *Multivariate Behavioral Research.* https://doi.org/10.1080/00273171.2019.1627660

Schoedel, R. (2019, September). *Day-and-night behavior patterns in the wild.* Presented at the 15th DPPD conference, Dresden.

Schroeders, U., Wilhelm, O., & Olaru, G. (2016a). Meta-Heuristics in Short Scale Construction: Ant Colony Optimization and Genetic Algorithm. *PloS One, 11*, 1–19.

Schroeders, U., Wilhelm, O., & Olaru, G. (2016b). The influence of item sampling on sex differences in knowledge tests. *Intelligence, 58*, 22–32. https://doi.org/10.1016/j.intell.2016.06.003

Schwarz, N. (2012). Why researchers shoudl think “real-time”: A cognitive rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of Research Methods for studying Daily Life.* New York, NY: Guilford Press.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology’s view of human nature. *Journal of Personality and Social Psychology, 51*, 515–530. https://doi.org/10.1037/0022-3514.51.3.515

Seifert, A., Hofer, M., & Allemand, M. (2018). Mobile data collection: Smart, but not (yet) smart enough. *Frontiers in Neuroscience, 12*, 971. https://doi.org/10.3389/fnins.2018.00971

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive Psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). *Specification curve: Descriptive and inferential statistics on all reasonable specifications* [Preprint].

Stachl, C., Au, Q., Schoedel, R., Buschek, D., Völkel, S., Schuwerk, T., ... Bühner, M. (2019). *Behavioral Patterns in Smartphone Usage Predict Big Five Personality Traits* [Preprint]. https://doi.org/10.31234/osf.io/ks4vd

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

Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior, 26*, 1488–1495. https://doi.org/10.1016/j.chb.2010.05.013

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

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology, 2*, 2–10. https://doi.org/10.1111/j.1754-9434.2008.01097.x

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*, 189–225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9*, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research, 49*, 193–213.

https://doi.org/10.1080/00273171.2014.889593

von Stumm, S., & Ackerman, P. L. (2013). Investment and intellect: A review and meta-analysis. *Psychological Bulletin, 139*, 841–869. https://doi.org/10.1037/a0030746

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. https://doi.org/10.1007/BF02294627

Wilhelm, O., & Schroeders, U. (2019). Intelligence. In R. J. Sternberg & J. Funke (Eds.), *The Psychology of Human Thought* (pp. 255–275). Heidelberg: Heidelberg University Publishing.

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

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, 12, 1100–1122. https://doi.org/10.1177/1745691617693393

Zimmermann, J., Ritter, S., Masuhr, O., Jaeger, U., Spitzer, C., Woods, W. C., ... 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

Zook, M., Barocas, S., boyd, danah, Crawford, K., Keller, E., Gangadharan, S. P., ... Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology, 13*, e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Anhang

Erklärung über den Eigenanteil an den veröffentlichten oder zur Veröffentlichung vorgesehenen wissenschaftlichen Schriften meiner Dissertationsschrift

Nummerierte Aufstellung der eingereichten Schriften

1. Manuskript 1: Steger, D., Schroeders, U., & Gnambs, T. (2019). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment.* Advance online publication. https://doi.org//10.1027/1015-5759/a000494

2. Manuskript 2: 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

3. Manuskript 3: Steger, D., Schroeders, U., & Wilhelm, O. (2019). Caught in the act: Predicting cheating in unproctored knowledge assessment. Manuscript submitted for publication.

Darlegung des eigenen Anteils an diesen Schriften

Manuskript 1 Ich bin Erstautorin des Textes. Die Datenerhebung und Datenauswertung wurden vollständig von mir durchgefürt. Die Ergebnissdiskussion, Erstellung des Manuskripts und die Literaturrecherche wurden überwiegend von mir durchgeführt und in Teilen von Ulrich Schroeders und Timo Gnambs. Die Studienkonzeption wurde in gleichen Teilen von mir, Ulrich Schroeders und Timo Gnambs durchgeführt.

Manuskript 2 Ich bin Erstautorin des Textes. Die Datenaufbereitung und -auswertung wurde vollständig von mir durchgeführt. Die Ergebnisdiskussion, Erstellung des Manuskripts und die Literaturrecherche wurde überwiegend von mir durchgeführt und in Teilen von Ulrich Schroeders und Oliver Wilhelm. Die Studienkonzeption wurde in gleichen Teilen von mir, Ulrich Schroeders und Oliver Wilhelm durchgeführt.

Manuskript 3 Ich bin Erstautorin des Textes. Die Datenaufbereitung und -auswertung wurde vollständig von mir durchgeführt. Die Ergebnisdiskussion, Erstellung des Manuskripts und die Literaturrecherche wurde überwiegend von mir durchgeführt und in Teilen von Ulrich Schroeders und Oliver Wilhelm. Die Studienkonzeption wurde in gleichen Teilen von mir, Ulrich Schroeders und Oliver Wilhelm durchgeführt.

Anschriften der jeweiligen Mitautoren

Timo Gnambs	timo.gnambs@lifbi.de
Ulrich Schroeders	schroeders@psychologie.uni-kassel.de
Oliver Wilhelm	oliver.wilhelm@uni-ulm.de

Bestätigung der Richtigkeit der oben abgegebenen Erklärung

Ich bestätige die von Frau Diana Steger abgegebene Erklärung

Timo Gnambs ________________________

Ulrich Schroeders ________________________

Oliver Wilhelm ________________________